Projektgruppe InMemory Planung mit SAP HANA

Projektdokumentation

Abteilung Wirtschaftsinformatik 1:
Very Large Business Applications

Betreuer: Prof. Dr.-Ing. habil. Jorge Marx Gómez
Dipl.-Inform. Nils Giesen
Dr.-Ing. Dirk Peters
Dipl.-Math. Jens Siewert
M.Eng&Tech. Viktor Dmitriyev

Danksagung
Die Mitglieder der Projektgruppe OliMP bedanken sich bei ihren Betreuern Prof. Dr.-Ing. habil. Jorge Marx Gómez, Dipl.-Inform. Nils Giesen, Dr.-Ing. Dirk Peters, Dipl.-Math. Jens Siewert sowie M.Eng&Tech. Viktor Dmitriyev und Hans-Hermann Redenius für die Koordination und die Unterstützung während der gesamten Projektlaufzeit. Hierbei gilt unser Dank auch für die Zeit der wöchentlichen Projektgruppentreffen und der Teilnahme an außerordentlichen Terminen. Über den gesamten Zeitraum waren sie Ansprechpartner für inhaltliche sowie organisatorische Fragen und haben mit interessierten Diskussionen zur Verbesserung unseres Projektergebnisses beigetragen. Darüber hinaus gilt unser Dank ebenso für die Initiierung und Organisation der Ausstellung auf der Messe CeBIT in Hannover, da es uns die Möglichkeit gegeben hat, Teile unserer Projektarbeit einem interessierten Publikum vorzustellen und weiteres Feedback zu erhalten.

Ebenso geht unser Dank an Herrn Dr. Joachim Kurzhöfer von der AS Inpro GmbH für die hilfreiche Unterstützung der Projektgruppe sowie Martin Donauer und Deyan Stoyanov von der eXin AG für die Betreuung der Seminararbeiten sowie die SAP-Schulungen. Ein weiterer Dank geht an das Hasso Plattner Institut in Potsdam für die Bereitstellung der notwendigen Infrastruktur.

Projektgruppe OliMP im März 2015

Inhaltsverzeichnis

Glossar **15**

1 Einleitung **17**

2 Problemstellung und Zielsetzung **18**
2.1 Problemstellung . . . 18
2.2 Visionsfindung . . . 18
2.3 Vision . . . 25
2.4 Geplante Vorgehensweise . . . 28

3 Hypothesenbildung **31**
3.1 Einflussfaktoren . . . 31
3.2 Formulierung der Hypothesen . . . 32

4 Projektmanagement **36**
4.1 Rahmenbedingungen . . . 36
4.2 Projektorganisation . . . 38
4.3 Projektplanung . . . 44
4.3.1 Grobe Planung . . . 44
4.3.2 Feinplanung . . . 47

5 Auswahl der Software **51**

6 Anforderungsanalyse **56**

7 Beschreibung der Datenbasis **59**
7.1 Astro-Daten . . . 59
7.2 Strompreise . . . 59
7.3 Stromverbrauchsdaten . . . 59
7.4 Temperatur-Daten . . . 60
7.5 Zeitdimension . . . 60

8 Datenstruktur **62**
8.1 EEX-Daten . . . 62
8.1.1 EEX-Ex-Ante-Daten . . . 62
8.1.2 EEX-Ex-Post-Daten . . . 66
8.2 Verbrauchsdaten von Entso-E . . . 69
8.3 Wetterdaten . . . 69
8.4 Astro-Daten . . . 71
8.5 Stompreisdaten für Haushalt und Industrie . . . 73

9 Architektur-Entwurf **75**
9.1 Definitionen . . . 75
9.2 Typisierung des Softwaresystems . . . 76
9.3 Bausteinsicht . . . 76
9.3.1 Allgemeine Schichtenarchitektur . . . 77
9.3.2 EL-Prozess Design . . . 78

9.3.3 Data Warehouse Design 79
9.4 Verteilungssicht 81
9.5 Laufzeitsicht 81

10 Implementierung **83**
10.1 Code Conventions 83
10.1.1 Java 83
10.1.2 SAP HANA 83
10.2 Javadoc des Programmcodes 84
10.3 ELTA-Prozess 85

11 Auswahl der Algorithmen **89**
11.1 Zeitreihenanalyse 89
11.1.1 Linear Regression with damped Trend and seasonal Adjust 89
11.1.2 ARIMA 90
11.1.3 Forecast Smoothing 91
11.1.4 Single Exponential Smoothing 92
11.1.5 Double Exponential Smoothing 92
11.1.6 Triple Exponential Smoothing 93
11.2 Regressionsanalyse 94
11.2.1 Bi-variate natürliche logarithmische Regression 94
11.2.2 Bi-variate geometrische Regression 95
11.2.3 Multilineare Regression 97
11.2.4 Polynomiale Regression 98
11.2.5 Exponentielle Regression 99
11.3 Support Vector Machine (SVM) 100
11.4 Messkriterien 106
11.4.1 R-Squared 106
11.4.2 R-Squared adjusted 107
11.4.3 MAE (Mean Absolute Error) 107
11.4.4 RMSE (Root Mean Squared Error) 107
11.4.5 Variationskoeffizienten CV(MAE) und CV(RMSE) 108
11.5 Korrelation zwischen den Faktoren 108

12 Test und Evaluation **110**
12.1 Konzept 110
12.2 Testergebnisse des Datenimports 113
12.3 Ergebnisse der Vorhersagen 116
12.3.1 Datenbasis 116
12.3.2 Zweiter Versuch der Datenbasis 130
12.4 1. Hypothese 145
12.5 2. Hypothese 163
12.6 3. Hypothese 187
12.7 Evaluation der Ergebnisse 209
12.8 Evaluation der Hypothesen 209
12.9 Betriebswirtschaftlicher Mehrwert 215
12.10Ausblick 216
12.11Evaluation der Anforderungen 218

13 SAP HANA Verbesserungen **220**
13.1 Fehlerhafte Modelle AFM . . . 220
13.2 Dokumentation PAL vertiefen . . . 220
13.3 Vereinfachung Anwendung Algorithmen . . . 221
13.4 Limitierung der Resultsets . . . 221
13.5 Fehlerdatei CSV-Import ungenau . . . 221
13.6 Assistent Control Files . . . 221
13.7 Abbrechen komplexer Operationen . . . 221
13.8 Einfrieren bei komplexen Operationen . . . 222

14 Fazit **223**

15 Veranstaltungen **227**
15.1 Design Thinking Workshop . . . 227
15.2 Boule Turnier . . . 228
15.3 Schulung in SAP PA und SAP BPC mit eXin AG . . . 228
15.4 HPI Cloud Symposium und Future SOC Lab Day . . . 229

16 CeBIT **232**
16.1 Kick-Off Ausstellertreffen am 26.11.2014 . . . 232
16.2 Bericht des Messeauftritts aus Sicht des OliMP Projekts . . . 234

Literaturverzeichnis **236**

Seminararbeiten **239**

Abbildungsverzeichnis

1 Grober Architekturentwurf der Vision . . . 28
2 Projektsverlaufübersicht . . . 44
3 Meilensteinplan ab Oktober . . . 50
4 SAP HANA Architektur [SAP14a, S. 14] . . . 52
5 Application Function Modeler (AFM) . . . 53
6 Nutzungsvarianten von SAP PA[Gra13] . . . 54
7 Der Konfigurationsassistent zur Zeittabelle in SAP HANA . . . 61
8 Ausschnitt aus der SAP HANA Zeittabelle . . . 61
9 Architektur Schichten . . . 78
10 Architektur Pakete EL-Programm . . . 79
11 Snowflake-Schema . . . 80
12 Star-Schema . . . 80
13 Architektur Verteilungssicht . . . 81
14 Sequenzdiagramm für einen beispielhaften Extraktions- und Ladeprozess . . 82
15 Pseudocode des Reports auf Rohdaten- und Vorverarbeitungsebene . . . 85
16 Pseudocode der CTL-Datei . . . 86
17 Pseudocode des Import-Befehls . . . 87
18 Pseudocode der Fehlerdatei . . . 87
19 Separierende eindimensionale Hyperebene [JWHT13, Vgl. S. 340] . . . 102
20 maximum margin hyperplane [JWHT13, Vgl. S. 342] . . . 103
21 Testkonzept . . . 111
22 Protokoll zur Prognose . . . 112
23 Visualisierter Verlauf . . . 113
24 View `VIEW_CONSUMPTION_ALL_DIRTY` . . . 117
25 Bereinigung der Entsoe-Stromdaten . . . 118
26 View für die Trainingsdaten der Stromverbräuche . . . 119
27 View für die Testdaten der Stromverbräuche . . . 119
28 Grafische Darstellung der Vorhersagen . . . 120
29 Grafische Darstellung des zweiten Durchlaufes . . . 121
30 Grafische Darstellung des dritten Durchlaufes . . . 122
31 Grafische Darstellung des vierten Durchlaufes . . . 123
32 SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung . . . 126
33 SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung für Wochenarbeitstage . . . 126
34 SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung für Wochenenden . . . 127
35 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Jan2014“ . . . 127
36 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Zusammen“ . . . 128
37 Grafische Darstellung der Vorhersage mit der polynomialen Regression . . . 129
38 View für die Trainingsdaten der Stromverbräuche mit zusätzlichen Zeitangaben . . . 131
39 View für die Testdaten der Stromverbräuche mit zusätzlichen Zeitangaben . 132
40 SQL-Statement für den gesamten Trainingsdatensatz (2009-2013) . . . 133

41 SQL-Statement für das Jahr 2013 als Trainingsdatensatz 134
42 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2009_2013“ 135
43 Vergleich von tatsächlichen Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2013“ 135
44 SQL-Statement für den gesamten Trainingsdatensatz 137
45 SQL-Statement für Juni bis Dezember 2013 als Trainingsdatensatz 137
46 SQL-Statement für das Jahr 2013 als Trainingsdatensatz 137
47 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2009-2013“ 139
48 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Jun2013“ 139
49 Diagramm Vergleich Ist und Forecast 140
50 SQL-Statement für den gesamten Trainingsdatensatz bei der SVM 141
51 Support Vector Maschine mit $\gamma = 1$ und $C = 1$ 142
52 Support Vector Maschine mit $\gamma = 0.01$ und $C = 100$ 143
53 Support Vector Maschine mit $\gamma = 0.01$ und $C = 1000$ 143
54 SQL-Code zum Übertragen der Feiertagsinformationen 145
55 View für die Trainingsdaten der Stromverbräuche mit zusätzlichen Zeitangaben sowie Feiertage und Wochenende 146
56 View für die Testdaten der Stromverbräuche mit zusätzlichen Zeitangaben sowie Feiertage und Wochenende 147
57 SQL-Statement für den gesamten Trainingsdatensatz mit Feiertagen 148
58 SQL-Statement für den gesamten Trainingsdatensatz mit Wochenenden . . 148
59 SQL-Statement für den gesamten Trainingsdatensatz mit Wochenenden und Feiertagen 148
60 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe“ 150
61 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Wo“ 150
62 Vergleich von tatsächlichen Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe_Wo“ 150
63 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe“ 153
64 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Wo“ 154
65 Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe_Wo“ 154
66 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe“ 155
67 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Wo“ 155
68 Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe_Wo“ 156
69 SQL-Statement für den gesamten Trainingsdatensatz bei SVM 157
70 SQL-Statement für den Trainingsdatensatz bezogen auf das Wochenende bei SVM 157

71 SQL-Statement für den Trainingsdatensatz bezogen auf die Arbeitswoche und Feiertage bei SVM . . . 158
72 Support Vector Machine mit $\gamma = 0.01$ und $C = 1000$ . . . 159
73 Support Vector Machine, WE/WD mit $\gamma = 0.01$ und $C = 1000$ . . . 159
74 Support Vector Machine, WE/WD mit $\gamma = 0.001$ und $C = 1000$ . . . 160
75 Support Vector Machine, WD mit $\gamma = 0.01$ und $C = 100$ . . . 160
76 Support Vector Machine, WD mit $\gamma = 0.01$ und $C = 1000$ . . . 161
77 Support Vector Machine, WD mit $\gamma = 0.001$ und $C = 1000$ . . . 161
78 View `VIEW_AIRTEMP_ALL_DIRTY` . . . 164
79 Prozedur zum Bereinigen der Temperaturdaten . . . 165
80 Erweiterte View `CONSUMPTION_TRAINING` . . . 166
81 Erweiterte View `CONSUMPTION_FORECAST` . . . 167
82 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur . . . 167
83 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen . . . 168
84 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen in der Woche . . . 168
85 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen an Wochenenden . . . 168
86 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Temp“ . . . 170
87 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Temp_Fe_Wo“ . . . 170
88 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Spl_Wo“ . . . 171
89 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Spl_Woe“ . . . 172
90 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Zusammenfassung der Durchläufe „Spl_Wo“ und „Spl_Woe“. . . . 173
91 SQL-Code der Trainingsdaten für das Jahr 2013 . . . 173
92 SQL-Code der Trainingsdaten für Dezember 2013 . . . 174
93 SQL-Code der Trainingsdaten für Januar 2013 . . . 174
94 SQL-Code der Trainingsdaten für Januar 2009-2013 . . . 174
95 Erstellung der Sequenz zum korrekten durchnummerieren der Zeilen . . . 174
96 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „2013_komplett“ . . . 175
97 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Dez_2013“ . . . 175
98 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Jan_2013“ . . . 176
99 Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Jan_2009_2013“ . . . 176
100 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur . . . 178
101 SQL-Code für den gesamten Trainingsdatensatz mit Temperatur-Feiertag-Wochentag . . . 178
102 Diagramm Vergleich Ist und Forecast. Durchlauf: „Temp“ . . . 179
103 Diagramm Vergleich Ist und Forecast. Durchlauf: „Temp_Wo_W“ . . . 180
104 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_W“ . . . 180

105 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_Wo“ 181
106 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_zusamm“ 181
107 SQL-Code für den gesamten Trainingsdatensatz bei SVM 182
108 SQL-Code für den Trainingsdatensatz bezogen auf die Temperatur bei SVM 183
109 Support Vector Machine, T mit $\gamma = 0,001$ und $C = 1000$ 184
110 Support Vector Machine, T mit $\gamma = 0,01$ und $C = 100$ 184
111 Support Vector Machine, T+WD+WE mit $\gamma = 0,001$ und $C = 1000$ 185
112 Support Vector Machine, T+WD+WE mit $\gamma = 0,01$ und $C = 100$ 185
113 Erweiterte View `CONSUMPTION_TRAINING` 188
114 Erweiterte View `CONSUMPTION_FORECAST` 189
115 SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten in Deutschland 191
116 SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten der Industrie in Deutschland 191
117 SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland 191
118 SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland 191
119 SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland 191
120 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „H-Halt“ 193
121 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Ind“ 194
122 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „H-Halt-Ind“ 194
123 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Spl-Woch“ 194
124 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Spl-WEnde“ 195
125 Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Split-zusam“ 195
126 SQL-Statement für den gesamten Trainingsdatensatz mit Gesamtpreis . . . 197
127 SQL-Statement für den gesamten Trainingsdatensatz mit Haushaltspreis . . 197
128 SQL-Statement für den gesamten Trainingsdatensatz mit Industriepreis . . 197
129 SQL-Statement für den gesamten Trainingsdatensatz mit Woche ohne Wochenende 198
130 SQL-Statement für den gesamten Trainingsdatensatz mit Wochenende . . . 198
131 Diagramm Vergleich Ist und Forecast. Durchlauf: „Gesamt “ 199
132 Diagramm Vergleich Ist und Forecast. Durchlauf : „H_halt “ 200
133 Diagramm Vergleich Ist und Forecast. Durchlauf: „Indus “ 200
134 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_W “ 201
135 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_Wo “ 201
136 Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_zusamm “ 202
137 SQL-Statement für den gesamten Trainingsdatensatz bei SVM 202
138 SQL-Statement für den Trainingsdatensatz bezogen auf den Haushalt bei SVM 203

139 SQL-Statement für den Trainingsdatensatz bezogen auf die Industrie bei SVM . . . 203
140 Support Vector Maschine mit $\gamma = 0,01$ und $C = 1000$ . . . 204
141 Support Vector Maschine, (IN) mit $\gamma = 0,01$ und $C = 1000$ . . . 205
142 Support Vector Maschine,(IN) mit $\gamma = 0,001$ und $C = 1000$ . . . 206
143 Support Vector Maschine, (IN/HH) mit $\gamma = 0,01$ und $C = 1000$ . . . 206
144 Support Vector Maschine, (IN/HH) mit $\gamma = 0,01$ und $C = 1000$ . . . 207
145 Support Vector Maschine,(IN/HH) mit $\gamma = 0,001$ und $C = 1000$ . . . 207
146 Vergleich von tatsächlichen Stromverbrauch und vorhergesagten Verbrauch. Versuch: „LSDNA-Zusammen“ . . . 212
147 Vergleich zwischen tatsächlichen und prognostizierten Stromverbrauch mit der lineare Regression mit gedämpften Trend und saisonaler Anpassung . . . 217
148 Projektseite des OliMP Projekts in der CeBIT Broschüre des Landes Niedersachsen . . . 235

Tabellenverzeichnis

1 Namen und Rollen der Projektteilnehmer . . . 39
2 Tooleinsatz zur Abwicklung des Projektes . . . 42
3 Zuordnung und Thema der Seminararbeiten . . . 46
4 Meilensteinplanung ab Oktober Teil 1 . . . 48
5 Meilensteinplanung ab Oktober Teil 2 . . . 49
6 Musskriterien . . . 56
7 Wunschkriterien . . . 56
8 Abgrenzungskriterien . . . 57
9 Qualitätsanforderungen . . . 57
10 Kennzahlensteckbrief erwartete Solarenergie . . . 62
11 EEX_Ex_ante_Planned_Generation_Solar . . . 62
12 Kennzahlensteckbrief erwartete Windenergie . . . 63
13 EEX_Ex_ante_Planned_Generation_Wind . . . 63
14 Kennzahlensteckbrief erwartete Nichtverfügbarkeitskapazität . . . 63
15 EEX_Ex_ante_Non_usability_Generation . . . 64
16 Kennzahlensteckbrief geplante Kapazität aller Einheiten . . . 64
17 EEX_Ex_ante_Available_Capacity . . . 64
18 Kennzahlensteckbrief geplante maximale Energieerzeugung . . . 64
19 EEX_Ex_ante_Planned_Energy . . . 65
20 Kennzahlensteckbrief geplante Energieerzeugung bezogen aufs Land . . . 65
21 EEX_Ex_ante_Planned_Generation . . . 65
22 Kennzahlensteckbrief Summe installierter Kapazitäten über 100 MW . . . 65
23 EEX_Ex_ante_Sum_installed_capacity . . . 66
24 Kennzahlensteckbrief tatsächliche Energieerzeugung . . . 66
25 EEX_Ex_post_Actual_Generation . . . 66
26 Kennzahlensteckbrief tatsächliche Windenergieerzeugung . . . 67
27 EEX_Ex_post_Generation_Wind . . . 67
28 Kennzahlensteckbrief tatsächliche Solarenergieerzeugung . . . 67
29 EEX_Ex_post_Generation_Solar . . . 67
30 Kennzahlensteckbrief tatsächliche Nichtverfügbarkeitskapazität . . . 68

31 EEX_Ex_post_Non_usability_Generation . . . 68
32 Kennzahlensteckbrief tatsächliche Energieerzeugung des Vortages . . . 68
33 EEX_Ex_post_Previous_Day_Generation . . . 68
34 Kennzahlensteckbrief Stromverbrauch . . . 69
35 Entsoe_Power_Consumption . . . 69
36 Kennzahlensteckbrief Lufttemperatur . . . 69
37 DWD_Weather_AirTemp . . . 70
38 Kennzahlensteckbrief Bewölkungsgrad . . . 70
39 DWD_Weather_Cloudiness . . . 70
40 Kennzahlensteckbrief Luftdruck . . . 70
41 DWD_Weather_Pressure . . . 70
42 Kennzahlensteckbrief Bodentemperatur . . . 71
43 DWD_Weather_SoilTemp . . . 71
44 Kennzahlensteckbrief Winddaten . . . 71
45 DWD_Weather_Wind . . . 71
46 Kennzahlensteckbrief Sonnenaufgang/-untergang . . . 72
47 Kennzahlensteckbrief Mondphasen . . . 72
48 Kennzahlensteckbrief Arbeitzeitfaktor . . . 72
49 Kennzahlensteckbrief Feiertage . . . 73
50 DWD_Astro Referenzspalten Zeit . . . 73
51 DWD_Astro Referenzspalten Sonnenaufgang/-untergang . . . 73
52 DWD_Astro Referenzspalten Mondphasen . . . 73
53 DWD_Astro Referenzspalten Arbeitszeitfaktor . . . 73
54 DWD_Astro Referenzspalten Feiertage . . . 73
55 Kennzahlensteckbrief Stompreise . . . 74
56 Strom-Preis_Deutschland_Haushalt . . . 74
57 Strom-Preis_Deutschland_Industrie . . . 74
58 Entwicklungsrichtlinien für Java . . . 84
59 Entwicklungsrichtlinien für SAP HANA . . . 84
60 Tabelle `OLIMP.ENTSOE_POWER_CONSUMPTION` . . . 116
61 Erforderliche Datenfelder für die SAP-PAL-Algorithmen . . . 117
62 View „CONSUMPTION“ . . . 119
63 Modellbildung des ersten Durchlaufes. Algorithmus: Arima . . . 120
64 Modellbildung des zweiten Durchlaufes. Algorithmus: Arima . . . 121
65 Modellbildung des dritten Durchlaufes. Algorithmus: Arima . . . 122
66 Modellbildung des vierten Durchlaufes. Algorithmus: Arima . . . 122
67 Nicht durchführbare Arima-Konfigurationen . . . 123
68 Parametereinstellungen der linearen Regression mit gedämpftem Trend . . . 126
69 Fehlerkennzahlen der Anwendung des Modells . . . 127
70 View `PAL.CONSUMPTION.TRAINING` . . . 132
71 View `PAL.CONSUMPTION.FORECAST` . . . 133
72 Relevante Dateien für die Multiple Lineare Regression . . . 133
73 Parametereinstellungen der multiplen linearen Regression . . . 134
74 Fehlerkennzahlen der Multiplen Linearen Regression . . . 135
75 Relevaten Dateien für die Exponentielle Regression . . . 137
76 Parametereinstellungen der exponentiellen Regression (für alle Durchläufe) 138
77 Fehlerkennzahlen der Prognose mit der exponentiellen Regression . . . 138
78 Relevante Dateien für die Support Vector Machine . . . 141

79 Parametereinstellungen der Support Vector Machine mit radialem Kernel . 142
80 Tabelle OLIMP.DWD_ASTRO 145
81 Relevante Dateien für die multiple lineare Regression 148
82 Parametereinstellungen der multiplen linearen Regression (für alle Durchläufe) 149
83 Ergebnisse der Hypothese 149
84 Relevante Dateien für die exponentiale Regression 152
85 Parametereinstellungen der exponentialen Regression (für alle Durchläufe) . 152
86 Fehlerkennzahlen der Prognose des ersten Teilversuchs (2009-2013) 153
87 Fehlerkennzahlen der Prognose des zweiten Teilversuchs (2013) 153
88 Relevante Dateien für die Support Vector Machine 157
89 Durchläufe SVM für die 1. Hypothese 158
90 Tabelle OLIMP.DWD_WEATHER_AIRTEMP 163
91 Tabelle PRE.AIRTEMP_ALL_CLEAN 164
92 Relevante Dateien für die lineare Regression 167
93 Parametereinstellungen der multiplen linearen Regression 169
94 Ergebnisse der 2. Hypothese 169
95 Ergebnisse der 2. Hypothese 174
96 Relevaten Dateien für die Exponentiale Regression 177
97 Fehlerkennzahlen der Anwendung des Modells auf die Testdaten 179
98 Relevante Dateien für die Support Vector Machine 182
99 Durchläufe SVM für die 2. Hypothese 183
100 Tabelle PRE.PRICE_HOUSEHOLD_ALL_CLEAN 187
101 Tabelle PRE.PRICE_INDUSTRY_ALL_CLEAN 187
102 Relevante Dateien für die Multiple Lineare Regression 190
103 Parametereinstellungen der Multiple Lineare Regression (für alle Durchläufe) 192
104 Ergebnisse der Hypothese 193
105 Relevaten Dateien für die Exponentiale Regression 197
106 Fehlerkennzahlen der Anwendung des Modells auf die Testdaten 199
107 Relevante Dateien für die Support Vector Machine 203
108 Durchläufe SVM für die 3. Hypothese 204
109 Promethee-Ranking zu den Durchläufen 211
110 Fehlerkennzahlen der besten Prognosen der multiplen lineare Regression für alle Hypothesen 213
111 Fehlerkennzahlen der besten Prognosen der exponentiellen Regression für alle Hypothesen 214
112 Fehlerkennzahlen der besten Prognosen der Support Vector Machine für alle Hypothesen 214
113 Legende 218
114 Umsetzung der Musskriterien 219
115 Umsetzung der Wunschkriterien 219
116 Umsetzung der Qualitätsanforderungen 219

Abkürzungsverzeichnis

AFM Application Function Modeller . . . 47
AFL Application Function Library . . . 53
BFL Business Function Library . . . 51
BPC Business Planning and Consolidation . . . 20
BWIP Business Planning Integrated Planning . . . 20
CRM Customer-Relationship-Management . . . 37
CMS Content Management System . . . 39
CSV Comma Seperated Values . . . 54
CTL Control File . . . 86
DB Datenbank . . . 22
DBMS Datenbank-Management-System . . . 51
DOD Definition of Done . . . 43
DWD Deutscher Wetterdienst . . . 25
DWH Data Warehouse . . . 21
EEN Enterprise Europe Network . . . 232
EEX European Energy Exchange . . . 25
ERP Enterprise Ressource Planning . . . 23
ETL Extract Transform Load . . . 20
HPI Hasso-Plattner-Institut . . . 37
JDBC Java Database Connectivity . . . 51
KDD Knowledge Discovery in Data . . . 21
MAE Mean Absolute Error . . . 107
MDX Multidimensional Expressions . . . 23
ODBC Open Database Connectivity . . . 23
OLAP Online Analytical Processing . . . 76
OTB Oldenburger Turnerbund . . . 228
UML Unified Modeling Language . . . 75
PAL Predictive Analysis Library . . . 28
PMML Predictive Model Markup Language . . . 95
PROMETHEE Preference Ranking Organization Method for Enrichment Evaluation 209
RMSE Root Mean Squared Error . . . 107
ROI Return on Investment . . . 22
SFTP Secure File Transfer Protocol . . . 85
SQL Structured Query Language . . . 29
SVM Support Vector Machine . . . 100
VLBA Very Large Business Applications . . . 36
XML Extensible Markup Language . . . 26
XLS Excel-Datei . . . 85

Glossar

ABC-Analyse Ein universelles Verfahren zum Priorisieren von Aufgaben, Produkten und Problemen.

Application Function Library Ermöglicht datenintensive Berechnungen unmittelbar in der Datenbank auszuführen, um einen überflüssigen Transport von Daten in die Anwendungsschicht zu vermeiden.

Business Function Library Eine Anwendungsbibliothek, die vorgefertigte parametergesteuerte Funktionen aus dem ERP-Bereich enthält.

Charting Visualization Object Models Bietet eine neue Standard Visualisierungslösung für SAP BI-Client Tools.

Confluence Ist eine Kollaborationssoftware für die Gruppenarbeit.

Content-Management-System Eine Software, die die Erstellung, Bearbeitung und Organisation von Inhalten in Webseiten sowie in anderen Medienformen ermöglicht.

CSV Datei Beschreibt die Struktur einer Textdatei zur Speicherung und zum Austausch von Informationen.

Eclipse Ist ein Java-basiertes Programmierwerkzeug zur Entwicklung von Software.

Enterprise-Resource-Planning Bezeichnet die unternehmerische Aufgabe, interne Ressourcen, wie z.B. Kapital, Personal, Betriebsmittel und Material, rechtzeitig zu planen.

European Energy Exchange Ist eine Börse für Energie und energienahe Produkte.

European Network of Transmission System Operators for Electricity Ist der Verband Europäischer Übertragungsnetzbetreiber.

Eurostat Ist das statistische Amt der Europäischen Union in Luxemburg.

Extensible Markup Language Eine erweiterbare Auszeichnungssprache, die Darstellung hierarchisch strukturierter Textdateien ermöglicht.

Extraktion Transformation Laden Bei einem ETL-Prozes werden die Daten aus verschiedenen Datenquellen mit unterschiedlichen Strukturen in einer Zieldatenbank zusammengeführt.

Hasso-Plattner-Institut Ein universitäres Institut in Potsdam, dass sich vor allem durch den Studiengang IT-Systems Engineering auszeichnet.

Java Database Connectivity Eine einheitliche Schnittstelle, die die Verbindung von Java Plattform und Datenbanken verschiedener Hersteller ermöglicht.

Jira Projektmanagementtool für die Gruppenarbeit, in der Softwareprojekte verwaltet werden.

Knowledge-Discovery-in-Database Ein umfassender Datenanalyseprozess, in dessen Kern Verfahren des Data Mining zur Anwendung kommen.

Multidimensional Expressions Eine Datenbanksprache für multidimensionale Datenbanken.

Online Analytical Processing Zählt zu den Methoden der analytischen Informationssysteme. Dabei werden die wachsenden Datenbestände sichtbar in mehreren Dimensionen dargestellt.

Open Database Connectivity Eine Datenbankschnittstelle, die SQL als Datenbanksprache verwendet.

OpenVPN Ein Programm zum Aufbau eines Virtuellen Privaten Netzwerkes.

R Eine Programmiersprache, die statistische Datenanalyse und Anfertigung von statistischen Grafiken ermöglicht.

SAP AG Ist ein führender Anbieter von Unternehmenssoftware.

SAP Business Objects Ist eine Business Intelligence Lösung der SAP AG.

SAP Business Warehouse Bezeichnet eine Data-Warehouse-Anwendung der SAP AG.

SAP Business Warehouse - Integrated Planning Ist ein Tool der SAP AG zur Durchführung einer integrierten Unternehmensplanung in Unternehmen.

SAP Predictive Analytics Ist ein Tool der SAP AG für die Durchführung von Data Mining und statistischer Analyse.

SAP Sybase IQ Ist eine von der SAP AG für Analyse und Business Intelligence Zwecke optimierte Datenbank.

Secure File Transfer Protocol Eine für die Secure Shell (SSH) entworfene Alternative zum File Transfer Protocol (FTP), die die verschlüsselte Übertragung von Daten ermöglicht.

Strategic Enterprise Management - Business Planning and Simulation Ist ein Tool der SAP AG zur Unterstützung der Unternehmensplanung.

Unified Modeling Language Ist eine Modellierungssprache zur Spezifikation, Konstruktion sowie Visualisierung von Modellen für Softwaresysteme.

View Kann in SQL als virtuelle Tabelle definiert werden, die als Objekt gespeichert wird. Diese enthalten keine konkreten Daten, sondern lediglich Verweise auf die entsprechenenden Spalten der zugrundeliegenden Basistabellen.

1 Einleitung

Das Masterstudium der Informatik und Wirtschaftsinformatik an der Carl von Ossietzky Universität Oldenburg sieht die Durchführung einer einjährigen Projektgruppe vor, in der ein Team von 6 - 12 Studierenden anhand eines gegebenen Problems die vollständige Entwicklung von der Problemanalyse bis hin zur Realisierung des Systems durchführt. Neben Methoden und Inhalten des Studienfachs erlernen die Studierenden dabei berufstypische Arbeitsweisen wie das Arbeiten im Team, die Arbeitsteilung und die Übernahme von Verantwortung. Zugleich werden persönliche Fähigkeiten wie die Aufbereitung von Inhalten, zielorientiertes Argumentieren sowie die Präsentations- und Urteilsfähigkeit gefördert. Dies geschieht unter anderem im Rahmen von Seminararbeiten, die in ihrer Thematik mit dem Oberthema oder der Projektgruppe im Allgemeinen in Verbindung stehen. Die nähere Erläuterung zu den Seminararbeiten befindet sich im Kapitel 4.3.1.

Die Projektgruppe *In Memory Planung mit SAP HANA* steht unter der Leitung von Prof. Dr. Jorge Marx Gómez und wird in der Abteilung *Very Large Business Applications (VLBA)* durchgeführt. Die Frage- und Aufgabenstellung der Projektgruppe lautet wie folgt: Wie kann durch den Einsatz von In-Memory-Planungs- und Prognosewerkzeugen eine bessere Simulation zukünftiger Auswirkungen heute zu treffender Entscheidungen unterstützt werden? Wie kann der Unternehmenslenker unterstützt werden, transparente Entscheidungen statt aus dem „Bauch heraus“ mit Hilfe von In Memory Planung zu treffen? Weiterhin soll evaluiert werden, ob die technologische Weiterentwicklung betriebswirtschaftliche Mehrwerte erzeugen kann. Zielsetzung ist also die Erzeugung von Transparenz sowie die Optimierung des Planungsprozesses. Näheres dazu ist im Kapitel 2 zu finden.

Die Projektgruppe hat sich den Eigennamen *OliMP* gegeben. Dies kann einerseits als Kurzform von *Ol*denburger *I*n*M*emory *P*lanung gesehen werden, andererseits wird mit diesem Wortspiel und der grafischen Präsentation des Kurznamens in Form des Logos direkt der Bezug zur Anwendung von SAP Software sichtbar, welche den Menschen auf dem Gipfel (dem Olymp) als marketingwirksames Mittel nutzt.

2 Problemstellung und Zielsetzung

Dieses Kapitel erläutert ausführlich die Problemstellung, mit der sich die Projektgruppe über den Verlauf des Projektes beschäftigt. Hierzu wird zunächst die allgemeine, übergeordnete Problemstellung erläutert. Anschließend wird im Abschnitt 2.2 und 2.3 ausführlich auf die konkrete Problemstellung eingegangen. Das Kapitel schließt mit der Vorgehensweise zur Problemlösung ab.

2.1 Problemstellung

Im Rahmen der Projektgruppe soll evaluiert werden, ob der Einsatz von In-Memory Planungs- und Prognosewerkzeugen am Beispiel von SAP HANA eine bessere Simulation zukünftiger Auswirkungen unterstützt. Außerdem ist zu bewerten, ob die Entscheidungsträger dadurch zu transparenteren Entscheidungen gelangen anstatt diese aus dem „Bauch-heraus“ treffen zu müssen und ob sich in diesem Kontext weitere betriebswirtschaftliche Vorteile ergeben. Insgesamt soll hierdurch der gesamte Planungsprozess optimiert und transparent gestaltet werden. Diese Zielsetzung wird in den folgenden beiden Abschnitten anhand einer praxisnahen Problemstellung konkretisiert und verfeinert.

2.2 Visionsfindung

Da die Zusammenarbeit mit dem externen Partner der Projektgruppe aufgrund nicht näher spezifizierter Gründe am 13.08.2014 endete (siehe Kapitel 4) und auch vor diesem Zeitpunkt keine konkrete Aufgabenstellung an die Projektgruppe erfolgte, wurden in einem internen Brainstorming mögliche praxisnahe Aufgabenstellungen und Szenarien besprochen. Insbesondere sind in diesem Brainstorming auch die Erkenntnisse aus den Seminararbeiten (Siehe Kapitel 16.2) eingeflossen. Dieser Abschnitt dokumentiert die Ergebnisse des Brainstormings. Im darauf folgenden Abschnitt wird die von den Projektgruppenmitgliedern auf Basis des Brainstormings erstellte Vision vorgestellt.

Im Brainstorming erzählt jedes Projektgruppenmitglied die wichtigsten Erkenntnisse aus der jeweiligen Seminararbeit. Die relevanten Punkte werden dabei näher diskutiert.

Seminarbereich Planung

Seminararbeit „Planungsprozesse in Unternehmen aus fachlicher und organisatorischer Sicht und Ihre Ziele“

- Systemation
- Trends erkennen und reagieren
- Unternehmensplanung

- 5 Schritteplan: Informationsbeschaffung, Analyse, Ziele, Strategie, Maßnahmen
  Problem: Informationsaufbereitung und Beschaffung: Zu viele Daten vorhanden, Aufbereitung notwendig.
- Die richtigen Informationen müssen in der richtigen Form zum richtigen Zeitpunkt bereitgestellt werden.
- Planungsziele allgemein:
  - kurzfristig: den Gewinn erhöhen (Liquidität erhöhen, z. B. durch zusätzliche Verkaufsmaßnahmen).
  - langfristig: Sicherung der Unternehmensexistenz (Marktanteil, Wettbewerbsvorteil, Expansion).

Seminararbeit „Chancen und Herausforderungen in der klassischen Unternehmensplanung“

- Chancen (Probleme):
  - Prognose: viele Stellschrauben, die Prognosen beeinflussen können.
  - Simulation: eine Stellschraube beeinflusst alle anderen Attribute
    d.h. es können mehrere Umweltsituationen betrachtet werden.
  - Zukunftsplanung: Große Datenmengen, Problem: die richtigen Daten auswählen.
  - Übersicht notwendig, welche Daten relevant sind.
- Herausforderungen
  - Dynamik der Umwelt: Diverse unbekannte Einflüsse, z.B. Konkurrenz, Dynamik des Kunden.
  - Interessenkonflikt: Unterschiede der einzelnen Interessengruppen.
  - Fragwürdige Gesetzmäßigkeiten
  - verschiedene Zusammenhänge der Daten
- Planungshorizonte (als Beispiel)
  - Generell: Unternehmenstätigkeit, z. B. Wir bauen qualitativ hochwertige Autos.
  - Strategisch: Betrachtung von Geschäftsfeldern, Regionen,... Horizont: 5 Jahre.
  - Mittelfristig: Betrachtung von Maßnahmen, Projekte... Horizont: 2-3 Jahre.
  - Operativ: Betrachtung von Prozessen, Ressourcen... Horizont: 1 Jahre.

Seminarbereich Planungstools- und Werkzeuge

Seminararbeit „Planungs- und Prognosewerkzeuge mit ihren Stärken und Schwächen am Beispiel der Systeme SEM-BPS, BW-IP und BPC der SAP AG“

Warum sind Planungssysteme erforderlich: Unterstützung für Entscheider bei (komplexen) Planungsaufgaben, da die Komplexität solcher Planungsaufgaben so hoch ist, dass diese durch entsprechende Planungssysteme übernommen werden müssen. Was bietet zentrale Planung?

- Berücksichtigung verschiedener Zeitperioden.
- dezentrale Planung
  - Abteilungsgebunden
  - Ortsgebunden
  - Sparten
- zentrale Planung
  - Gesamtplanung
- Konsolidierung
- Planungs-/ Forecast-Funktionen (Algorithmen integriert)
- Organisatorische, betriebliche, Daten- Prozesse in Modellen abbildbar.
- Modellierung der Daten
- Datenimport aus verschiedenen Quellen
- Grafische Darstellung der Ergebnisse
- Business Planning Integrated Planning (BWIP) = aktuelles Programm der SAP für integrierte Unternehmensplanung
- Excel-Integration vorhanden

Probleme/Anforderungen an Planungstools:

- zentrale Datenbasis erforderlich
- Zuverlässige, bereinigte Daten
- Extract Transform Load (ETL) Prozess erforderlich, ggf. externes Programm
- keine Redundanzen in den Daten

Planungstools der SAP AG:

- Business Planning and Consolidation (BPC):
  - dezentrale Planung (z. B. einzelne Fachbereiche)

  - flexibel einsetzbar
  - Anwenderfreundlich
  - Office Integration
  - Konsolidierung möglich
- BWIP:
  - zentrale Planung
  - Data Warehouse (DWH) integriert
  - Planungsprozesse vordefiniert (durch zentrale Planung)
  - Konsolidierung nicht möglich ohne zusätzliches Tool

Ausblick: SAP möchte beide Tools zu einem Tool verschmelzen. SAP HANA kann für beide Systeme eingesetzt werden.

Seminararbeit „Statistische Verfahren zur Fortschreibung historischer Daten“

- Zeitreihenanalyse
  - M5-Modellbaumverfahren
  - Elman-Netze

M5-Modellbäume und Elman-Netze berücksichtigen den Faktor Zeit implizit bei ihren Prognoserechnungen. Beide Verfahren sind in R verfügbar. Kombination beider Verfahren möglich, was die Prädiktionsgenauigkeit möglicherweise erhöhen kann.

Der Knowledge Discovery in Data (KDD)-Prozess:

1. Datenauswahl (Mensch)
2. Vorverarbeitung (Mensch, Maschine)
3. Transformation (Maschine)
4. Data-Mining - Anwendung von Algorithmen (Maschine)
5. Interpretation (Mensch)
6. Visualisierung (Mensch, Maschine)

Der Data-Mining-Prozessschritt nimmt nur ca. 10% der Arbeitszeit in Anspruch. Datenauswahl, Vorverarbeitung und Interpretation nehmen den Großteil der Zeit in Anspruch.

Seminararbeit „Bewertung des Einsatzes von prädiktiven Methoden und Werkzeugen im SAP-Umfeld (SAP Predictive Analysis)“
Vorhersagen allgemein:

- Erstellung von Prognosemodellen

- effektive Prognostizierung
- effektive Verteilung von knappen Ressourcen

Typische Anwendung für Prognosemodelle:

- Responsemodelle (welcher Kunde kauft was)
- Cross-Selling (Was könnte einen Kunden sonst noch interessieren)
- Up-Selling (Erhöhung des Verkaufes)
- Abwanderungs- und Reaktivierungsmodelle (welche Kunden wandern am ehesten ab)
- Betrugsmodelle (welche Trans- oder Interaktionen sind betrügerisch oder müssen nachverfolgt werden)

Probleme/Anforderungen:

- Identifizierung der Möglichkeit, Einblicke in umsetzbare Geschäftsentscheidungen zu unterstützen.
- Überforderte Analysten (zu viele Daten, welche Daten sind relevant?).
- Ziele und gewünschte Ergebnisse aus den Analysen müssen definiert werden.
- umsetzbare Ergebnisse mit messbaren Return on Investment (ROI) erwünscht.
- Die verwendeten Datensätze sind entscheidend: Sie müssen sorgfältig konstruiert sein; das Team muss die Daten „verstehen“.
- Die Daten müssen von Anomalien bereinigt sein.
- Kooperation zwischen Abteilungen erforderlich.
- Idealfall: DWH mit allen Daten aus den verschiedenen Abteilungen vorhanden, mit bereits bereinigten Daten. (Auch hier: Bereinigung der Daten ist der aufwändigste Schritt).
- Auswahl eines geeigneten Modellierungswerkzeuges.
- Zentrale Fragen:
  - Wie werden Daten in das System gespeist?
  - Direkter Zugriff auf Datenbank (DB) oder zugriff über Dateien.
  - Schreiben von Ergebnissen zurück in die Datenbank.

SAP Predictive Analysis:

- Austausch von Informationen und Implementierung neuer Informationen.
- kann mit/ohne SAP HANA online/offline genutzt werden.

- Vermeidung von Medienbrüchen, wenn das Programm direkt mit SAP HANA genutzt wird.
- Nutzung der in der PAL-Library implementierten Algorithmen:
  - Datenaufbereitungsalgorithmen, Normalisierung
  - Clustering
  - Entscheidungsbäume
  - Zeitreihenanalyse (Glättung, Prognose, Trends)
  - Social-Network-Analysis
  - ABC-Analysis

Seminararbeit „Analytical capabilities of SAP HANA: Integration with R & Excel“

Excel kann über Multidimensional Expressions (MDX)-Schnittstelle mit SAP HANA verbunden werden. Dadurch ist ein direkter Zugriff auf die Daten in SAP HANA möglich. Alle Funktionen von Excel sind auf die importierten Daten nutzbar.

- Vorteile
  - Schnell
  - bekannte Excel-Umgebung
  - Direkte Verbindung zu Excel und SAP HANA sehr einfach (MDX, Open Database Connectivity (ODBC)).
- Nachteile
  - Sicherheitsproblematik
  - Excel = Flaschenhals?

Wenn Algorithmen angewendet werden sollen kann hierfür SAP Predictive Analysis verwendet werden, für simple Berechnung und Reporting kann Excel verwendet werden.

Aus diesen gesammelten Informationen haben die Projektgruppenmitglieder die erste (grobe) Idee für die Vision formuliert:

Die mittelfristige Planung für Energieunternehmen könnte mit Hilfe von SAP HANA optimiert werden. Dabei werden relevante Unternehmensdaten in eine zentrale Datenbasis importiert und verarbeitet, so dass anschließend Zusammenhänge zwischen diesen Daten mit Hilfe von SAP Predictive Analysis und Excel erkannt werden können. Mögliche Quellen können dabei sein: Webservices (Stromverbrauch, Temperatur, etc.), Enterprise Ressource Planning (ERP)-Systeme, Daten von Strombörsen, Stromlieferanten (Atomkraftwerke),

Flat-Files etc. Diese Datenquellen werden in SAP HANA über eine wohldefinierte Schnittstelle bereinigt und importiert. Auf diesen importierten Datenquellen werden Operationen (z.B. mit Hilfe SAP Predictive Analysis, PAL-Bibliothek, R, oder Excel ausgeführt. Dadurch können folgende Mehrwerte für Unternehmen entstehen:

- kontinuierliche Unternehmensplanung mit großen Datenmengen.
- dadurch könnten genauere Prädiktionen und Planungen (Qualität) resultieren.
- Schnelle Beantwortung von Fragen (Quantität durch mehrfache Simulation).
- Verstellung verschiedener Stellschrauben möglich.
- Förderung der von Abteilungsübergreifsender Zusammenarbeit.

Diese erste, grobe Vision wurde in mehreren Iterationen verfeinert. Unter anderem wurde in diesen Iterationen mehrfache Recherchearbeit von den Projektgruppenmitliedern betrieben und die Vision mit den Ergebnissen entsprechend verfeinert. Die Recherche zu den aktuellen Herausforderungen im Energiebereich werden im folgenden Stichpunktartig zusammengefasst.

Chancen und Herausforderungen in der Energiewirtschaft

- Privatpersonen werden ihren Stromverbrauch vermehrt selbst produzieren.
- Der Anteil von unregelmäßig und nicht planbar produzierten Strom wird zunehmen (z.B. durch Solar-, Bio-, und Windkraftwerke).
- Intelligente Stromnetze werden dadurch unabdingbar.
- Die Nachfrage wird - bezogen auf Bilanzkreise - erstellt.
- Die Bilanzkreisverantwortlichen erstellen hier im Voraus - basierend auf historischen Daten und Erfahrungswerten - einen Fahrplan der prognostizierten Last in ihrem jeweiligen Bilanzkreis.
- Entsprechend dem Fahrplan erwerben sie die Erzeugungsleistung entweder direkt von Energieversorgungsunternehmen, an der Strombörse oder erzeugen den Strom in eigenen Erzeugungsanlagen.
- Weicht später in der Realität die Last von der ursprünglichen Prognose ab, muss kurzfristig weitere Erzeugungsleistung gekauft beziehungsweise generiert werden oder es wird Ausgleichsenergie beim jeweiligen verantwortlichen Übertragungsnetzbetreiber bezogen.
- Im deutschen Regelzonenverbund als Gesamtbilanzebene werden durch den Einsatz von Regelenergie Prognoseabweichungen auf der Erzeugungs- und Nachfrageseite sowie Kraftwerksausfälle ausgeglichen.
- Da durch das Bilanzkreismanagement nicht geregelt wird, dass Erzeugung und Nachfrage auch geografisch im Gleichgewicht stehen, ergeben sich je nach Zeitpunkt in

einigen Regionen Leistungsüberschüsse und in anderen Regionen Leistungsdefizite. Die Stromnetze stellen durch die Übertragung und Verteilung des Stroms das geografische Gleichgewicht her.

- Einflüsse auf den Stromverbrauch:
  - Einfluss der Temperatur : gering.
  - Einfluss der kalendarischen Variablen : sehr stark.
  - Einfluss des Prognosehorizonts : Prognosefehler wachsen mit der Länge des Prognosehorizonts.

Für diese Erkenntnisse wurden die Ergebnisse aus [Ban03], [NRW15], [We13], [HF14], [Ste04] und [BFG04] verwendet.

2.3 Vision

Die von den Projektgruppenmitgliedern erstellte Vision fokussiert die Optimierung der mittelfristigen Unternehmensplanung für Energieunternehmen mit Hilfe der In-Memory-Datenbank SAP HANA: Die integrierte Unternehmensplanung von Energieunternehmen hinsichtlich des Strompreises und des Stromverbrauches basiert auf der Analyse, Simulation und Prognoseberechnung historischer Daten, mit deren Hilfe der kurz- bis mittelfristige Stromverbrauch für Deutschland möglichst optimal prognostiziert werden soll. In einer weiteren Recherchearbeit wurden hierfür potentielle Datenlieferanten ermittelt:

Daten der European Energy Exchange (EEX) Stromgeneration durch Solar- und Windkraftwerke, verfügbare gesamte Stromkapazität, aktuelle Stromgenerationsdaten sowie nicht verfügbare Energiekapazitäten.
Link zur Datenquelle: Die Daten der EEX wurden vom Offis - Institut für Informatik in Oldenburg bezogen.

Meteorologische Daten des Deutscher Wetterdienst (DWD) Temperatur-, Niederschlags- und Winddaten, Sonnenaufgangs- und -untergangsdaten sowie Feiertagskalender.
Link zur Datenquelle: `ftp://ftp-cdc.dwd.de/pub/CDC/`

ENTSO-E historische Stromverbrauchsdaten - bezogen auf Deutschland.
Link zur Datenquelle:
`https://www.entsoe.eu/db-query/consumption/mhlv-a-specific-country-for-a-specific-month`

EUROSTAT Strompreisdaten für Privathaushalte und Industrie.
Links zur Datenquelle: `http://ec.europa.eu/eurostat/data/database`

Die mittelfristige Unternehmensplanung für Energieunternehmen bezieht sich dabei auf den geografischen Bereich Deutschland. Das bedeutet, es werden nur Daten verwendet, die eine Analyse, Simulation und Prognoseberechnung für den Bereich Deutschland ermöglichen.

Auf dieser bisherigen Wissensbasis wird die folgende vorläufige Haupthypothese formuliert:

> Je mehr historische Trainingsdaten für die Prognose und Simulation zur Verfügung stehen, desto höher die Prognosegenauigkeit.

In der Projektarbeit soll dabei verifiziert werden, welche Kombination welcher Daten die höchste Prognosegenauigkeit für die mittelfristige Planung in Deutschland bietet. Dazu soll wie folgt vorgegangen werden:

- Historische Stromverbrauchsdaten aus einer definierten Periode bilden die Grundlage für die Prognose einer darauf folgenden Prognoseperiode.
- Diese Prognose wird mit der Realität verglichen, um die Prognosegenauigkeit zu ermitteln.
- Für die gleiche Prognoseperiode werden anschließend sukzessiv zusätzliche Datenquellen an die Datenbasis angebunden, um damit die Prognosegenauigkeit zu verändern.

Zusätzliche Datenquellen sind:

- Meteorologische Daten (z.B. Temperatur-, Niederschlags-, Winddaten, Sonnenaufgangs und -untergangsdaten).
- Stromverbrauchsdaten
- Feiertagskalender
- Rohstoffpreise für Stromhalberzeugnisse

Insgesamt werden mehrere Prognosedurchläufe geplant, um diejenige Kombination von Datenquellen zu finden, die die höchste Prognosegenauigkeit bietet. Die Datenquellen sind dabei heterogen, d.h. es können Daten aus Webservices, Flat-Files, Extensible Markup Language (XML)-Files, anderen Datenbanken (z.B. MySQL) oder Daten aus ERP-Systemen vorliegen. Die Daten sollen mit Hilfe eines Präprozesses (Datenaufbereitungsalgorithmus) automatisch vorverarbeitet und anschließend in SAP HANA geladen werden. Damit soll die Qualität der Daten sichergestellt werden. Die bereinigten Daten stehen anschließend in SAP HANA für Analysen und Prognosen durch SAP Predictive Analysis (siehe 5), der PAL-Bibliothek oder Microsoft Excel zur Verfügung. An dieser Stelle soll der Geschwindigkeitsvorteil von SAP HANA verifiziert werden: Insbesondere soll geklärt werden, ob

die unternehmerische Planung durch die schnellere Verarbeitungsgeschwindigkeit mit einer gleichzeitig höheren Anzahl von Daten durch die In-Memory-Technologie unterstützt werden kann. Insbesondere soll auch evaluiert werden, ob der In-Memory-Ansatz zu einer Optimierung der Prognosen führt.

Daraus können folgende Mehrwerte aus technischer und wirtschaftlicher Sicht für Unternehmen entstehen:

- Die Daten kommen zwar immer noch aus heterogenen Datenquellen, aber durch den automatisierten ETL-Prozess stehen die Daten in *einer* Datenbank (SAP HANA) zur Verfügung. Die Heterogenität der Daten ist für den Benutzer nicht direkt sichtbar, da er nur auf die Daten in SAP HANA zugreift.
- Informationsüberfluss: Welche Daten sind relevant? Der automatische ETL-Prozess könnte so realisiert werden, dass nicht alle Datenfelder der heterogenen Datenquellen importiert werden, sondern nur die Datenfelder, die von dem Anwender definiert, beziehungsweise für die Prognosen benötigt werden.
- Qualität der Daten: Durch den automatischen ETL-Prozess soll die Qualität der Daten gewährleistet werden.
- Der Planungsprozess wird optimiert, indem nach der idealen Kombination derjenigen Datenquellen gesucht wird, welche zur höchsten Prognosegenauigkeit führt.
- Der Zeitvorteil für die Erstellung von Berichten, Prognosen und Simulationen kann evaluiert werden.
- Förderung der Zusammenarbeit zwischen den Abteilungen (durch Zugriff auf eine zentrale Datenbasis).
- Datenzugriff: Die Zuteilung von Benutzerrechten erfolgt auf Datenbankebene.
- Verarbeitung von großen Datenmengen, was ohne SAP HANA ökonomisch nicht zielführend wäre.

Die Abbildung 1 zeigt den ersten groben Architekturansatz des zu erstellenden Systems.

In diesem vorläufigen Entwurf liegen zunächst viele heterogene Datenquellen und -formate vor (rechts im Bild). Diese Heterogenität wird mit Hilfe des automatischen ETL-Prozesses aufgehoben, so dass die Daten anschließend in einem bereinigten und SAP HANA konformen Format vorliegen. Die bereinigten Daten werden anschließend ebenfalls automatisiert in die SAP HANA Instanz geladen, so dass die Daten nach Abschluss des automatisierten ETL-Prozesses in Form von Tabellen in der SAP HANA Instanz vorliegen. Innerhalb der Datenbank müssen anschließend die notwendigen Verknüpfungen zwischen den Daten erstellt werden, damit anschließend Analysen, Simulationen und Prognosen mit diesen Daten erstellt werden können. Hierzu existieren die folgenden Möglichkeiten:

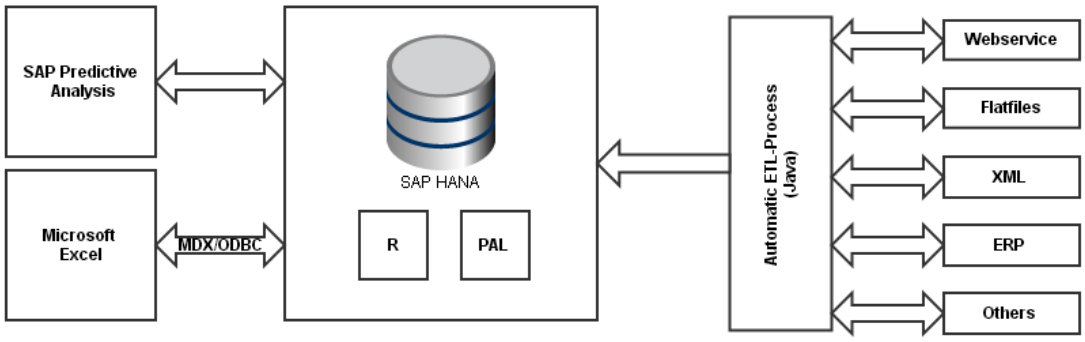


Abbildung 1: Grober Architekturentwurf der Vision

- SAP Predictive Analysis: Das Programm ist eine separate Softwarelösung der SAP AG. Das Programm kann mit SAP HANA Datenbanken verbunden werden. Die Daten werden entweder aus der Datenbank lokal auf den Rechner oder direkt, online in SAP HANA verarbeitet. Dabei ermöglicht das Programm anschließend Analysen, Simulationen, Prognosen und Berichte auf Basis der in der SAP HANA Instanz vorliegenden Daten.
- Microsoft Excel: Analog zu SAP Predictive Analysis besteht mit Microsoft Excel die Möglichkeit der Datenanbindung an SAP HANA. Die Daten werden aus der Datenbank extrahiert und anschließend lokal in Microsoft Excel geladen. Für die Analyse stehen anschließend die gewohnten Microsoft Excel Funktionen und -Diagramme zur Verfügung.
- SAP-Predictive Analysis Library (PAL) und R: Die SAP-PAL ermöglicht eine direkte Einbindung und Ausführung von R-Code (z.B. Prädiktionsalgorithmen) in SAP HANA. Damit entfällt das zusätzliche Laden von Daten aus der SAP HANA Instanz auf lokalen Rechnern. Die Integration der R-Algorithmen in SAP HANA ermöglicht die Verarbeitung der Algorithmen direkt in der Datenbank, was zu einer extrem hohen Performance - auch in Verbindung mit extrem großen Datenmengen - führt [SAP14b].

Die Beschreibung der von den Projektgruppenmitgliedern formulierten Vision ist hiermit abgeschlossen. Es folgt die Erläuterung der zum Zeitpunkt der Visionsformulierung geplanten Vorgehensweise um die in diesem Abschnitt definierten Zielsetzungen zu realisieren.

2.4 Geplante Vorgehensweise

Die folgende Aufzählung zeigt schemenhaft die notwendigen Schritte, um die Inhalte der Vision aus den vorigen Abschnitt zu realisieren.

1. Zunächst müssen die notwendigen Datenquellen beschafft und für das Laden in die SAP HANA Instanz vorbereitet werden. Hierzu ist es erforderlich, die Daten in ein spezielles Format zu transformieren.
2. Sobald die Daten in SAP HANA vorliegen, müssen diese auf Anomalien geprüft und bereinigt werden. Ebenfalls muss geprüft werden, ob für jeden Zeitpunkt des Betrachtungszeitraumes entsprechende Daten vorliegen. Gegebenfalls müssen die bereinigten Daten für die Anwendung spezifischer Algorithmen weiter angepasst werden.
3. Die verschiedenen Datenquellen müssen anschließend mit entsprechenden Structured Query Language (SQL)-Statements in Verbindung zueinander gesetzt werden. Dies kann beispielsweise über eine View realisiert werden.
4. Sobald die Daten in einem passenden Format für die jeweiligen Algorithmen vorgehen, kann die Analyse und Prognose der Daten beginnen. Hierzu soll folgendermaßen vorgegangen werden:
   a) Zunächst soll der Stromverbrauch auf lediglich Basis des historischen Stromverbrauches vorhergesagt werden (Dies bildet in allen weiteren Ausführungen die *Datenbasis*. Der Zeitraum der historischen Daten bildet dabei der Zeitraum Januar 2009 bis Dezember 2013. Der zu prognostizierende Zeithorizont ist der Stromverbrauch für den Monat Januar 2014.
   b) In weiteren Schritten werden sukzessive weitere Datenquellen (Features) für den gleichen Zeitraum hinzugefügt. Solche Features sind zum Beispiel: Die Lufttemperatur, Strompreise oder Angaben zu Feiertagen und Wochenenden.
   c) Dabei wird jede Kombination von Algorithmus und der verwendeten Features dokumentiert und anhand von definierten Fehlerkennzahlen evaluiert. Um hierzu ein systemisches Vorgehen zu ermöglichen, werden Hypothesen formuliert. Diese Hypothesen werden im Kapitel 3 näher erläutert.
5. Sobald die Ergebnisse der verschiedenen Hypothesen unter der Verwendung der verschiedenen Algorithmen vorliegen, kann evaluiert werden, welcher Algorithmus unter Verwendung welcher Features die höchste Prädiktionsgenauigkeit liefert.
6. Mit Hilfe dieser Ergebnisse kann im Anschluss evaluiert werden, ob betriebswirtschaftliche Vorteile durch die Verwendung von SAP HANA im Energiesektor entstehen und ob dies zu einer Optimierung des Planungsprozesses führt.

Die Beschreibung der geplanten Vorgehensweise ist hiermit abgeschlossen. Hierzu ist anzumerken, dass diese Beschreibung lediglich eine grobe, schemenhafte Orientierung der Vorgehensweise für die Projektgruppenmitglieder darstellt. Unter Umständen werden im Projektverlauf weitere Schritte hinzugefügt, ausgelassen oder wesentlich umfangreicher behandelt. Im nächsten Kapitel werden die von den Projektgruppenmitgliedern formulierten

Hypothesen erläutert, die ein systemisches Vorgehen zur Durchführung der Projektgruppenarbeit ermöglichen.

3 Hypothesenbildung

In diesem Kapitel werden die von den Projektgruppenmitgliedern formulierten Hypothesen näher beschrieben. Zunächst erfolgt eine Beschreibung der für den Stromverbrauch relevanten Einflussfaktoren. Anschließend werden die konkreten Hypothesen formuliert.

3.1 Einflussfaktoren

In diesem Abschnitt werden die möglichen Einflussfaktoren beschrieben, die sich auf den Stromverbrauch auswirken. Dazu zählen kalendarische Variablen, Temperatur und Strompreise.

Kalendarische Variablen Sie spielen eine wichtige Rolle für die Bestimmung und Optimierung der Prognosenqualität [Ban03]. Ihre Größe und die Art ihrer Berechnung werden in diesem Abschnitt erläutert.

Wochentagstypen Der Stromkonsum unterscheidet sich grundsätzlich an Wochenenden, Feiertagen von Werktagen. Aufgrund des Einflusses dieser Tage auf benachbarte Tage werden folgende Tagestypen unterschieden: Wochentage ab Montag bis Freitag; Wochenende: Samstag und Sonntag sowie gesetzliche Feiertage (zum Beispiel Weihnachten oder Ostern).

Sommerferienindex Betriebsferien und urlaubsbedingte Abwesenheit führen in vielen Fällen zu einem Absinken der Last in den Sommerferien im Vergleich zu den umliegenden Wochen. Dies kann zum Beispiel durch den Sommerferienindex berücksichtigt werden. Je nach Gebiet und Jahr kann der Einfluss der Sommerferien auf die Last undeutlich/erratisch oder mehr oder weniger deutlich und regelmäßig sein. Deshalb ist die Berücksichtigung des Sommerferienindex optional.

Weihnachtsindex Der Weihnachtsindex verfolgt für die Weihnachtsferien eine ähnliche Idee wie der Sommerferienindex für die Sommerferien. Aufgrund der jährlich wechselnden Konstellation von Feiertagen, Wochentagen sowie Schulferienbeginn und -ende kann aber hier der Ansatz einer Mitteilung der historischen Last über mehrere Jahre und der Übertragung auf andere Jahre (vom Kalibrierungs- auf den Prognosezeitraum) anhand eines zeitlichen Bezugspunktes wie der Ferienmitte nicht verwendet werden. Stattdessen wird in [Ban03] eine Kernphase und Übergangsphasen der Weihnachtsferien vorgeschlagen:

- Die Kernphase dauert stets vom 24. Dezember 0 Uhr bis zum 1. Januar 24 Uhr.
- Die Übergangsphase am Anfang der Weihnachtsferien dauert vom ersten Ferientag 0 Uhr bis zum 23. Dezember 24 Uhr; die Übergangsphase am Ende der Weihnachtsferien dauert vom 2. Januar 0 Uhr bis zum letzten Ferientag 24 Uhr.

Kalendarische Variablen haben generell einen sehr starken Einfluss auf die Stromprognosen [Ste04].

Temperatur Niedrige Temperaturen sind mit zusätzlicher Inbetriebnahme elektrischer Raumheizungen verbunden, was seinerseits zu erhöhtem Stromkonsum führt. Die Berücksichtigung der Temperatur ist optional. In der Literatur wird der Einfluss der Temperatur unterschiedlich, ausgelegt. Aufgabe der Projektgruppe ist es, den Einfluss der Temperatur auf die Prognosen zu evaluieren [Ban03] und [BFG04].

Strompreise Die Strompreise können auch einen Einfluss auf den Stromverbrauch haben. Generell wird zwischen Strompreisen für private Haushalte und Industrie unterschieden. Die Daten hierzu stammen von der Internetpräsenz der „EUROSTAT“. Die EUROSTAT ist der führende Anbieter für hochwertiger Statistiken über ganz Europa. Ihre Aufgabe ist es, die Europäische Union mit hochwertigen Statistiken zu unterstützen.

Die Vorstellung der Einflussfaktoren auf den Stromverbrauch ist hiermit abgeschlossen. Im folgenden Abschnitt werden die daraus generierten Hypothesen vorgestellt.

3.2 Formulierung der Hypothesen

Ziel der Projektgruppe ist die Optimierung der mittelfristigen Unternehmensplanung für Energieunternehmen mit Hilfe von SAP HANA. In diesem Zusammenhang wurde die Hypothese aufgestellt, dass die Prognosegenauigkeit umso höher ausfällt, je mehr Daten- und insbesondere Datenfeatures für die Prognose und Simulation zur Verfügung stehen.

Für die Überprüfung dieser Leithypothese wurden Unterhypothesen gebildet, die jeweils von einem steigenden Einflussfaktor auf den Stromverbrauch ausgehen. Im Folgenden werden die hierzu formulierten Hypothesen - die es im Anschluss zu evaluieren gilt - vorgestellt.

Haupthypothese Die Haupthypothese lautet: Je mehr Datenfeatures und damit Datenmengen vorliegen, desto höher fällt die Prognosegenauigkeit für den Stromverbrauch aus. Mit Datenfeatures sind Einflussgrößen auf den Stromverbrauch gemeint, die sich auf die Prognose des Stromverbrauches auswirken können. Dies sind zum Beispiel Angaben zu Werk- und Sonntagen oder Angaben zur Temperatur. Mit Datenmengen ist der Zeitraum der historischen Daten gemeint: Ist der betrachtete historische Zeitintervall höher, so führt dies in der Regel zu besseren Prädiktionsergebnissen [Ste04]. Diese Hypothese soll mit Hilfe der folgenden Unterhypothesen evaluiert werden.

Basishypothese Die Basishypothese stellt die Grundlage für die Verifizierung der weiteren Hypothesen dar. Mit Hilfe der Basishypothese soll evaluiert werden, ob der Stromverbrauch lediglich anhand des historischen Stromverbrauches prognostiziert werden kann.

Die Basishypothese bezieht demnach lediglich einen Faktor für die Prädiktion und Simulation in die Berechnungen mit ein. Dies ist der historische Stromverbrauch selbst.

Hypothese 1 Die erste Hypothese bezieht den historischen Energieverbrauch sowie Angaben zu Werk-, Sonn- und Feiertagen mit in die Berechnungen des Modells und die Prognose ein. In dieser Hypothese werden demnach drei Faktoren für die Berechnung der Prädiktion einbezogen. Dies ist zunächst der historische Stromverbrauch, anschließend werden die Angaben zu Werk-, Sonn- und Feiertagen hinzugefügt. Dabei soll so vorgegangen werden, dass in mehreren Durchläufen evaluiert wird, welcher dieser Faktoren den höchsten Einfluss auf eine korrekte Vorhersage hat. Hierzu sind die folgenden Durchläufe definiert:

- 1. Durchlauf: historischer Stromverbrauch mit Angaben zu Werk- und Sonntagen.
- 2. Durchlauf: historischer Stromverbrauch mit Angaben zu Feiertagen.
- 3. Durchlauf: historischer Stromverbrauch mit Angaben zu Werk- und Sonntagen sowie Feiertage.

Mit den ersten beiden Durchläufen soll dabei gezeigt werden, wie sich der Einfluss der jeweiligen Variablen isoliert auf die Prognose des zukünftigen Stromverbrauches auswirkt. Im dritten Durchlauf werden dann die Angaben zu Werk- und Sonntagen sowie die Angaben zu Feiertagen gemeinsam in die Prognoseberechnung einbezogen. Hiermit soll untersucht werden, ob die Einbeziehung beider Variablen bessere Prädiktionsergebnisse liefert, als die isolierte Betrachtung der Einflussgrößen. Damit wird auch die Evaluation der Haupthypothese unterstützt: Wenn dieser Durchlauf bessere Ergebnisse liefert als die isolierte Betrachtung der Variablen, kann dies als Zeichen dafür interpretiert werden, dass eine höhere Anzahl von Features - und damit eine höhere Datenmenge - zu besseren Vorhersagen führt. Wenn das Ergebnis des dritten Durchlaufes ungenauere oder gleichwertige Ergebnisse als der erste und zweite Durchlauf liefert, ist dies ein Zeichen dafür, dass aufgrund der Menge an Daten und Features nicht unbedingt bessere Prädiktionsergebnisse produziert werden.

Hypothese 2 Die zweite Hypothese bezieht den Energieverbrauch, die Angaben zu Werk-, Sonn- und Feiertagen sowie die durchschnittliche Lufttemperatur Deutschlands mit in die Berechnungen des Modells und der Prognose ein. In dieser Hypothese werden demnach vier Faktoren für die Berechnung der Prognosen einbezogen. Dies sind die Angaben zu Werk- und Sonntagen, die Angaben zu Feiertagen sowie die Angaben zur durchschnittlichen Lufttemperatur in Deutschland. Um zu verifizieren, ob die Temperatur einen stärkeren Einfluss auf den Stromverbrauch hat als die Angaben zu Werk-, Sonn- und Feiertage, wird diese Hypothese mit Hilfe von zwei Durchläufen evaluiert:

- 1. Durchlauf: historischer Energieverbrauch mit Angaben zur Temperatur.

- 2. Durchlauf: historischer Energieverbrauch mit Angaben zu Werk-, Sonn-, Feiertage und durchschnittlicher Lufttemperatur.

Dabei soll mit dem ersten Durchlauf der Einfluss des Faktors der durchschnittlichen Lufttemperatur auf die Prädiktionsergebnisse evaluiert werden. Im zweiten Durchlauf werden die Features der ersten Hypothese mit in die Berechnung einbezogen, um hiermit auch wieder die Haupthypothese zu evaluieren: Wenn dieser Durchlauf bessere Ergebnisse liefert als die isolierte Betrachtung der Variable, ist dies ein Zeichen dafür, dass eine höhere Anzahl von Features zu besseren Prädiktionsergebnissen führt. Wenn das Ergebnis des zweiten Durchlaufes ungenauere oder gleichwertige Ergebnisse als der erste Durchlauf liefert, ist dies ein Zeichen dafür, dass aufgrund der Menge an Daten und Features nicht unbedingt bessere Prädiktionsergebnisse produziert werden.

Hypothese 3 Die dritte Hypothese bezieht den Energieverbrauch, die Angaben zu Werk-, Sonn- und Feiertagen, die durchschnittliche Lufttemperatur Deutschlands, sowie die durchschnittlichen Strompreise für Haushalte und der Industrie in Deutschland mit in die Berechnungen ein. In dieser Hypothese werden demnach 6 Faktoren für die Berechnung der Prädiktion einbezogen. Dies sind die Angaben zu Werk- und Sonntagen, die Angaben zu Feiertagen, die Angaben zur Lufttemperatur in Deutschland sowie die Angaben zu Strompreise für Haushalt und Industrie in Deutschland. Um zu verifizieren, ob die Strompreise - bezogen auf Haushalte und Industrie einen stärkeren Einfluss auf den Stromverbrauch haben, wird auch diese Hypothese in mehreren Durchläufen evaluiert:

- 1. Durchlauf: historischer Energieverbrauch mit Angaben zu Werk-, Sonn-, Feiertage, durchschnittlicher Lufttemperatur und Haushaltsstrompreise.
- 2. Durchlauf: historischer Energieverbrauch mit Angaben zu Werk-, Sonn-, Feiertage, durchschnittlicher Lufttemperatur und Industriestrompreise.
- 3. Durchlauf: historischer Energieverbrauch mit Angaben zu Werk-, Sonn-, Feiertage, durchschnittlicher Lufttemperatur sowie Haushalts- und Industriestrompreise.

Mit den ersten beiden Durchläufen soll zunächst die Auswirkung des Haushaltsstrompreises sowie des Industriestrompreises isoliert betrachtet werden. Im dritten Durchlauf werden anschließend beide Features zusammen betrachtet. Wenn dieser Durchlauf bessere Ergebnisse liefert als die isolierte Betrachtung der Variablen, ist dies ein Zeichen dafür, dass eine höhere Anzahl von Features zu besseren Prädiktionsergebnissen führt. Wenn das Ergebnis des dritten Durchlaufes ungenauere oder gleichwertige Ergebnisse als der erste und zweite Durchlauf liefert, ist dies ein Zeichen dafür, dass aufgrund der Menge an Daten und Features nicht unbedingt bessere Prädiktionsergebnisse produziert werden.

Die Vorstellung der Hypothesen ist hiermit abgeschlossen. Zu beachten gilt hier, dass dies die vorläufigen Hypothesen sind, die zunächst von der Projektgruppe abgearbeitet

werden. Die jeweiligen Features der Hypothesen wurden anhand der Literaturrecherche ausgewählt. Insbesondere sind hier diejenigen Features aufgenommen worden, die laut der Literatur den höchsten Einfluss auf den Stromverbrauch haben[1] [BFG04]. Sollte der zeitliche Rahmen Raum für weitere Hypothesen zulassen, können weitere Hypothesen - und damit weitere Features - getestet und evaluiert werden. Im folgenden Kapitel wird auf das Projektmanagement eingegangen.

[1] Für das Feature „durchschnittliche Lufttemperatur“ differieren die Literaturangaben deutlich.

4 Projektmanagement

In diesem Kapitel wird das Projektmanagement der Projektgruppe näher beschrieben. Hierzu werden im folgenden Abschnitt die Rahmenbedingungen der Projektgruppe erläutert. Es erfolgt die Benennung der internen und externen Projektgruppenmitglieder. Im darauf folgenden Abschnitt werden die einzelnen Aufgaben und Rollen der Projektgruppenmitglieder beschrieben und definiert. Danach wird in Abschnitt 4.2 auf die verwendeten Tools zur Projektabwicklung und das von den Projektgruppenmitgliedern fokussierte Vorgehensmodell zur Realisierung der Problemlösung eingegangen.

4.1 Rahmenbedingungen

Der Projektstart ist am 01. April 2014, dass Projektende ist auf den 31. März 2015 datiert. Insgesamt besteht die Projektgruppe aus elf Studierenden der Informatik sowie Wirtschaftsinformatik, die im folgenden Projektmitglieder aufgeführt sind:

- Rima Adhikari (K.C.)
- Farhad Gavan El-Yazdin
- Abdulmasih Hadaya
- Benjamin Hemken
- Ivaylo Ivanov
- Mariska Janz
- Igor Perelman
- Eduard Rajski
- Steffen Scheer
- Jonas Schlemminger
- Daniel Stratmann

Die Studierenden werden von Angehörigen der Abteilung Very Large Business Applications (VLBA) betreut. Namentlich sind das:

- Prof. Dr.-Ing. Jorge Marx Gómez
- Dipl.-Inform. Nils Giesen
- Dr.-Ing. Dirk Peters (bis 17.12.2014)
- Dipl.-Math. Jens Siewert
- M. Eng & Tech. Viktor Dmitriyev

Die Betreuung schlägt sich insbesondere darin nieder, dass es pro Woche ein Treffen mit den Projektgruppenmitgliedern sowie Betreuern gibt, in dem Fragen und Probleme betreffend der erfolgreichen Durchführung der Projektgruppe beantwortet werden. Zusätzlich nehmen die Betreuer der Abteilung VLBA eine Vermittlerrolle mit den externen Kooperationspartnern der Projektgruppe ein. Ebenso nimmt der externe Kooperationspartner AS Inpro GmbH an den wöchentlichen Treffen teil. Die Kooperationspartner werden in den folgenden Absätzen benannt.

Kooperation mit dem Hasso-Plattner-Institut (HPI) Das Hasso-Plattner-Institut für Softwaresystemtechnik hat seinen Sitz in Potsdam. Das HPI besteht zum einem aus einer Universität und zum anderen aus einer Stätte der Forschung im Bereich IT-Systems-Engineering. Das Future SOC Lab ist eine Forschungseinrichtung, die interessierten Wissenschaftlern eine Infrastruktur von neuester Hard- und Software kostenfrei für Forschungszwecke zur Verfügung gestellt. Dazu zählen teilweise noch nicht am Markt verfügbare Technologien. Diese Möglichkeiten zur Zusammenarbeit richtet sich insbesondere an Wissenschaftler in den Gebieten Informatik und Wirtschaftsinformatik. Einige der Schwerpunkte sind Cloud-Computing, Parallelisierung und In Memory Technologien. Die Aufgabe der Projektgruppe OliMP liegt in der Auswertung des Einsatzes der In Memory Technologie SAP HANA bei der Unternehmensplanung. Aus diesem Grund wurde zu Beginn der Projektarbeit eine Bewerbung mit der Bitte um eine Kooperation mit dem Future SOC Lab versendet und im Juli 2014 erfolgreich bestätigt. Seitens des HPI nehmen keine Vertreter an die wöchentlichen Treffen der Projektgruppe teil.

Kooperation mit der eXin AG Die eXin AG ist eine auf Prozess- und Technologielösungen im Bereich Analytics und Customer-Relationship-Management (CRM) spezialisierte Unternehmensberatung. Der Schwerpunkt der Tätigkeit liegt in der Unterstützung der Unternehmen beim effektiven Auf- und Ausbau von Business-Intelligence- und CRM-Konzepten zur gezielten Kontrolle und Steuerung der Geschäftsaktivitäten. Ansprechpartner der eXin AG sind Martin Donauer und Deyan Stoyanov. Mit Hilfe des externen Partners soll eine möglichst praxisnahe Aufgabenstellung mit realen Fragestellungen ermöglicht werden. Ein weiterer Vorteil, der sich aus der Zusammenarbeit mit dem externen Partner ergibt, ist der Kontakt zu weiteren Projektpartnern sowie Einblicke in die Arbeitswelt der IT- und Business-Beratung bereits während des Studiums.

Aufgrund nicht weiter spezifizierter Gründe endet die Zusammenarbeit mit der eXin AG am 13.08.2014 ohne die Stellung einer praktischen Aufgabe an die Projektgruppe. Ungeachtet dessen übernahmen Martin Donauer und Deyan Stoyanov jedoch auch weiterhin die Betreuung der Seminararbeiten (siehe Abschnitt 16.2). Zusätzlich erhielten die Projektgruppenmitglieder an zwei Tagen Schulungen von der eXin AG zu den Themen *SAP*

Predictive Analysis und *SAP-BPC*. Die Schwerpunkte dieser Schulungen werden im Kapitel 15 erläutert.

Kooperation mit der AS Inpro GmbH Auf der Suche nach einen neuen externen Partner auf Initiative von Herrn Prof. Dr. Jorge Marx Gómez wurde Herr Dr. Joachim Kurzhöfer zu einem Kick-Off-Meeting eingeladen. Herr Dr. Joachim Kurzhöfer ist Geschäftsführer der AS Inpro GmbH, einer Tochtergesellschaft der Lufthansa Systems AG. Lufthansa Systems AG verfügt über ein umfangreiches Portfolio von maßgeschneiderten Lösungen für unterschiedliche Branchen und besitzt eines der leistungsfähigsten Rechenzentren Europas. Das Leistungsspektrum erstreckt sich über die gesamte Breite an IT-Dienstleistungen, von IT-Consulting über die Entwicklung und Implementierung von Branchenlösungen bis zum Betrieb von Rechenzentren.

In diesem Meeting am 21.08.2014 wurde Herr Dr. Kurzhoefer über die Ziele und Herausforderungen der Projektgruppe und über die Erwartungen der Projektgruppenmitglieder an den potentiellen externen Partner informiert. Seit dem 17.09.2014 ist die AS Inpro GmbH offizieller Partner der Projektgruppe. Die Mitarbeit der AS Inpro GmbH schlägt sich darin nieder, dass das Unternehmen für die Projektgruppenmitglieder eine beratene Position hinsichtlich der fachlichen Fragestellung einnimmt.

Die Beschreibung der internen und externen Projektgruppenmitgliedern ist hiermit abgeschlossen. Im folgenden Abschnitt wird der organisatorische Rahmen der Projektgruppe erläutert.

4.2 Projektorganisation

Zunächst werden die Rollen und Aufgaben der internen Projektgruppenmitglieder dokumentiert. Diese Rollen wurden teilweise zu Beginn der Projektgruppe festgelegt. Wenn nicht anders angegeben, sind die zugeteilten Rollen über den gesamten Verlauf der Projektgruppe gültig.

Rollen und Aufgaben In der Phase des Projektstartes hatten die Projektmitglieder die Möglichkeit, sich für die Ausführung bestimmter Rollen und Aufgaben zu bewerben. Diese vorgenommen Einteilung ist in Tabelle 1 zu finden.

Im Verlaufe des Projektes wurden für die verschiedenen Rollen klare Aufgaben und Zuständigkeiten definiert. Der Hintergrund hierzu war, dass zwischenzeitlich Überlappungen und Unstimmigkeiten dazu geführt haben, dass nicht eindeutig ersichtlich war, welche Rolle für welche Aufgabe zuständig war. Mit der folgenden Definition soll gewährleistet werden, dass den jeweiligen Rolleninhaber die Aufgaben und Verantwortlichkeiten seiner

Name	Rolle
Mariska Janz	Webseitenverantwortliche
Jonas Schlemminger	System- und Serveradministrator
Benjamin Hemken	Testbeauftragter
Igor Perelmann	Kommunikationsbeauftragter
Daniel Stratmann	Dokumentationsbeauftragter
Ivaylo Ivanov	Social Events und Finanzen
Steffen Scheer	Projektmanager vom 01.04.2014 bis 06.08.2014
Abdulmasih Hadaya	Projektmanager vom 07.08.2014 bis 30.09.2014
Eduard Rajski	Projektmanager vom 01.10.2014 bis 30.11.2014
Farhad Gavan El-Yazdin	Projektmanager vom 01.12.2014 bis 31.01.2015
Rima Adhikari (K.C.)	Projektmanagerin vom 01.02.2015 bis 31.03.2015

Tabelle 1: Namen und Rollen der Projektteilnehmer

Rolle bewusst sind. In den folgenden Absätzen sind die hierzu entstandenen Aufzeichnungen dokumentiert:

Websitenverantwortliche Die Websitenverantwortliche hat die Aufgabe der Erstellung und Wartung userer Website (`http://www.ol-imp.de`) über den gesamten Verlauf der Projektgruppe. Zusätzlich wird die Websitenverantwortliche relevante Blogeinträge veröffentlichen. Hierzu erhält die Inhaberin dieser Rolle inhaltliche Unterstützung von der gesamten Projektgruppe. Ebenso gehört zu dieser Rolle die Wartung der Internetseite, zum Beispiel das Aufspielen von Updates des Content Management System (CMS).

System- und Serveradministrator Der System- und Serveradministrator ist verantwortlich für die Erstellung, Konfiguration und Wartung der verwendeten Hard- und Softwaresysteme. Zusätzlich ist er Ansprechpartner für alle Projektgruppenmitglieder bei technischen Problemen der verwendeten Systeme.

Testbeauftragter Der Testbeauftragte ist für die Erstellung eines Testkonzeptes der erstellen Software und für die Durchführung der Tests anhand dieses Konzeptes verantwortlich. Dabei obliegt das Testen anhand dieses Konzeptes allen Projektgruppenmitgliedern.

Kommunikationsbeauftragter Der Kommunikationsbeauftragte ist für die Kommunikation mit externen Stakeholdern verantwortlich. Weiterhin sind die Ergebnisse dieser Kommunikation an die restlichen Projektgruppenmitglieder heranzutragen und entsprechend zu bearbeiten.

Dokumentationsbeauftragter Der Dokumentationsbeauftragte ist für die Erstellung der Vorlagen für die Seminararbeiten und die Dokumentation verantwortlich. Zusätzlich ist der

Inhaber dieser Rolle der Ansprechpartner für alle Projektgruppenmitglieder hinsichtlich Fragen zur Erstellung der Seminararbeiten und Dokumentation mittels LaTeX. Er kann Dokumentationsaufgaben auch delegieren.

Social Events und Finanzen Diese Rolle verwaltet die gesamten Finanzen der Projektgruppe. Zusätzlich ist die Rolle für die Planung von Social Events (z.B. ein gemeinsames Frühstück) verantwortlich.

Projektmanagement In den Aufgabenbereich des Projektmanagements fallen:

- Überwachung, dass die zugeteilten Aufgaben von den Projektgruppenmitgliedern fristgemäß erfüllt werden.
- Eingreifen, wenn Aufgaben nicht fristgemäß erfüllt werden.
- Allgemeine Termin- und Aufgabenüberwachung.
- Erstellung und ggf. Anpassung des Meilensteinplanes.
- Zuteilung von Aufgaben, wenn sich hierfür keine Freiwilligen finden.
- Koordinations- und Kontrollfunktion.

Zu Beginn der Projektphase wurden einige grundlegende Vorgaben bezüglich des Projektmanagements definiert. Es wurde von den Betreuern vorgegeben, dass fünf Teilnehmer die Aufgabe des Projektmanagements im Wechsel übernehmen müssen. Diese teilen sich die Aufgabe über die gesamte Projektzeit ein, sodass für alle Beteiligten ein identischer Zeitaufwand entsteht.

Entwicklungsaufgabe Diese Aufgabe wird von jedem internen Projektgruppenmitglied eingenommen. Sie beinhaltet die Verpflichtung zur Gestaltung und Implementierung des Systems.

Protokollaufgabe Jedes Projektgruppenmitglied ist dazu verpflichtet, die Moderation und Protokollierung der wöchentlichen Treffen mit den Betreuern an definierten Zeitpunkten zu übernehmen. Die Aufgabe des Protokollierens der Sitzungen rotiert wöchentlich. Der jeweilige Protokollant hat die Aufgabe, die externen Sitzungen so ausführlich wie möglich zu protokollieren. Das Protokoll muss spätestens am Freitag nach einer Sitzung bis 18:00 Uhr im Projekt-Wiki Confluence verfügbar sein. Zusätzlich hat der Protokollant die Aufgabe, die aus den jeweiligen Sitzungen resultierenden Aufgaben mit einer entsprechenden Aufgabe im Projektmanagement-Tool Jira zu verknüpfen und einem Projektgruppenmitglied zuzuordnen.

Moderation Die Aufgabe der Moderation der Sitzungen rotiert wöchentlich. Der Moderator moderiert sowohl die externe als auch interne Sitzung der jeweiligen Woche. Der Moderator bereitet die Sitzung anhand von Tagespunkten vor. Er lädt die restlichen Projektmitglieder in Form einer E-Mail zu der jeweiligen Sitzung ein. In dieser E-Mail wird der vorläufige Ablauf der Sitzung anhand von Tagespunkten benannt. In den Sitzungen hat der Moderator folgende Aufgaben:

- Leitung der gesamten Sitzung.
- Sicherstellen, dass immer nur eine Person zur Zeit redet.
- Eingreifen, wenn ein Gespräch „aus dem Ruder“ läuft.
- Zielführende und problemorientierte Leitung des Treffens.
- Diskussionspunkte zu einer Entscheidung führen.

Die Benennung und Definition der Rollen und Aufgaben der Projektgruppenmitglieder ist hiermit abgeschlossen. Im folgenden Abschnitt werden die Tools, mit dessen Hilfe das Projekt unterstützt wird beschrieben.

Tools zur Projektunterstützung Für eine erfolgreiche Projektbearbeitung sind einige Tools erforderlich, die dafür sorgen, dass

- Termine und Aufgaben koordiniert werden
- eine Meilensteinplanung erstellt werden kann
- das Projektziel über den gesamten Verlauf verfolgt werden kann
- Jederzeit der aktuelle Stand des Projektes nachvollziehbar ist
- eine gemeinsame Wissensbasis geschaffen werden kann
- die Bearbeitung des Projektes nach einem bestimmten Vorgehensmodell erfolgt
- die effektive und effiziente Kommunikation der Projektgruppenmitglieder ermöglicht wird
- Projektrelevante Daten unter allen Mitgliedern schnell ausgetauscht werden können
- in einer gemeinsamen Entwicklungsumgebung gearbeitet werden kann.

Um diese Aufgaben und Problemstellungen möglichst effektiv zu lösen, wurde der Einsatz der in der Tabelle 2 genannten Werkzeuge zur Unterstützung des Projektes beschlossen.

Die Vorstellung der Tools, welche die Abwicklung des Projektes unterstützen ist hiermit abgeschlossen. An dieser Stellt sei auf Kapitel 5 hingewiesen. Hier werden die Programme, welche direkt zur Durchführung der Aufgabenstellung verwendet werden detailliert erläutert.

Tool	Einsatzbereich
Confluence + Jira	Erstellung und Zuweisung von Projektaufgaben, Teilen von projektrelevanten Dokumenten (z. B. Protokolle), Terminplanung, Meilensteinplanung
SVN	Bearbeiten von Dokumenten und Programmcode im Team, Workspace für Implementierungen, Verfügbarkeit von allen relevanten Daten für das Projekt
Pidgin (XMPP)	Programm für eine schnellere Kommunikation
Latex	Schreiben der Projektdokumentation
SAP HANA Studio	Entwicklungsumgebung für HANA Applikationen

Tabelle 2: Tooleinsatz zur Abwicklung des Projektes

Scrum In der Projektstartphase wurde entschieden, dass das Projekt nach dem Vorgehensmodell *Scrum* durchgeführt wird. Der folgende Abschnitt gibt einen kurzen Überblick über das Thema Scrum. Detaillierte Ausführungen befinden sich in der Seminararbeit im Anhang.

Scrum ist eine agile Produktentwicklungsmethode, die bei der Auslieferung der wichtigsten Geschäftsanforderungen innerhalb kürzester Zeit hilft. Mittels Scrum arbeitet das Produktentwicklungsteam eigenverantwortlich und interdisziplinär. Zentrale Bedeutung für die Zusammenarbeit ist, dass jedes Teammitglied auf seine Aufgaben fokussiert ist. Das bedeutet, dass das Produkt in Serien/Abschnitten von Sprints erstellt wird. Das Besondere an diesen Sprints ist, dass die in einem Sprint festgelegten Teilaufgaben komplett fertiggestellt werden. Das heißt, alle Teilaufgaben eines Sprints haben eine Anforderungs- und Entwurfsanalyse durchlaufen und wurden anschließend implementiert und getestet. Die Teilaufgaben eines Sprints sind anschließend fertiggestellt und können dem Auftraggeber als fertiges Programm im Sinne der Sprintdefinition vorgestellt werden. Scrum-Teams beschäftigen sich also innerhalb kurzer Zeit mit der Anforderungsdefinition, der Entwurfsanalyse, der Implementierung und dem Testen. Diese Phasen wiederholen sich je nach Projektgröße und Anzahl der Sprints mehrfach. Im Gegensatz dazu wird in vielen anderen Vorgehensmodellen jede Phase nur einmalig durchlaufen [Glo13].

Zu Beginn der Entwicklung hat sich die Projektgruppe für eine angepasste Scrum-Variante entschieden. Um den Koordinationsaufwand zu optimieren wurde anfangs zwei Teams mit jeweils fünf und sechs Personen gegründet. Innerhalb der Teams wurden Scrum Master und Product Owner bestimmt. Ein teamübergreifender Product Owner sollte dabei darauf achten, dass ein Gesamtzusammenhang hergestellt werden kann. Dabei war ein Team für die Durchführung der Datenextraktion, der Transformation und des Ladens der Daten in SAP HANA verantwortlich. Ein anderes Team war für die Hypothesenbildung, statistische Aufarbeitung sowie die Analyse verantwortlich. Die Kommunikation zwischen den Teams fand dabei durch regelmäßige gemeinsame Meetings statt. Teammitglieder konnten dazu jederzeit andere Mitglieder teamübergreifend kontaktieren. Jeder Sprint war auf ei-

ne Dauer von eine Woche ausgelegt. Außerdem wurde für jede Sprint-Aufgabe ein Ticket im Ticketsystem (Jira) erstellt. Diese mussten dann bis zum Ende des jeweiligen Sprints abgeschlossen sein. Beginn und Anfang eines jeden Sprints war das wöchentliche Projekt-Meeting. Diese Entscheidung wurde bis zum Ende des Projekts nicht revidiert.

Nach den ersten Sprints wurde jedoch deutlich, dass vereinzelte Tasks ohne adäquate Dokumentation als Abgeschlossen markiert wurden - und damit teilweise in Vergessenheit gerieten. Deshalb wurde gemeinsam eine Definition of Done (DOD) vereinbart, die besagt, dass Tasks erst als abgeschlossen markiert werden dürfen, sobald sie implementiert und dokumentiert wurden. Ab Oktober wurden Kernarbeitszeiten eingeführt, zu denen jedes Projektgruppenmitglied nach Möglichkeit anwesend sein soll. Diese sind Dienstags und Mittwochs von 10:00 bis 17:00 Uhr. Diese Zeiten dienen der Verbesserung der Kommunikation und Zusammenarbeit innerhalb der Gruppe.

Im November 2014 zeigte sich, welche Nebeneffekte die Einteilung in zwei Teams mit sich brachte. Ein Vorteil der zwei kleineren Teams war zwar der verringerte Koordinationsaufwand innerhalb dieser, was zu einer hohen Produktivität bezüglich der Umsetzung der Aufgaben führte. Jedoch haben sich die Teams untereinander nicht wie geplant abgesprochen, die Koordination von teamübergreifenden Aufgabenstellungen funktionierte nicht wie gewünscht. Um die gesamte Projektgruppe neu auf das Ziel zu fokussieren wurden daher die ursprünglichen Teams wieder aufgelöst. Fortan wurde nur noch innerhalb eines großen Teams gearbeitet. Um die alten Strukturen auch informal aufzulösen wurde festgelegt, dass jedes Gruppenmitglied bei jedem neuen Sprint möglichst mit anderen Gruppenmitgliedern (rotierend) an einer Aufgabe arbeitet. Diese Vorgehensweise führte dazu, dass sich die heterogen verteilten Fachkompetenzen (beispielsweise in der Mathematik, Softwareentwicklung und Organisation und Planung) optimal in der Gruppe verteilten.

4.3 Projektplanung

Dieses Kapitel dokumentiert die vorgenommene Projektplanung zur Startphase des Projektes. Generell gilt: Die Projektgruppe dauert 12 Monate und umfasst auch die Beschäftigung in der vorlesungsfreien Zeit. Jedem Teilnehmer werde bei Bestehen der Projektgruppe 24 Kreditpunkte beziehungsweise ECTS vergeben. Daraus folgt ein zeitlicher Aufwand von 24 * 30 Stunden = 720 Stunden. Bei 45 Arbeitswochen entspricht dies einer wöchentlichen Arbeitszeit von 16 Stunden.

4.3.1 Grobe Planung

Die auf Abbildung 2 zu sehende Übersicht vom ersten Projektmanager wurde zu Beginn des Projekts erstellt, nachdem eine frühere Übersicht verfeinert werden sollte.

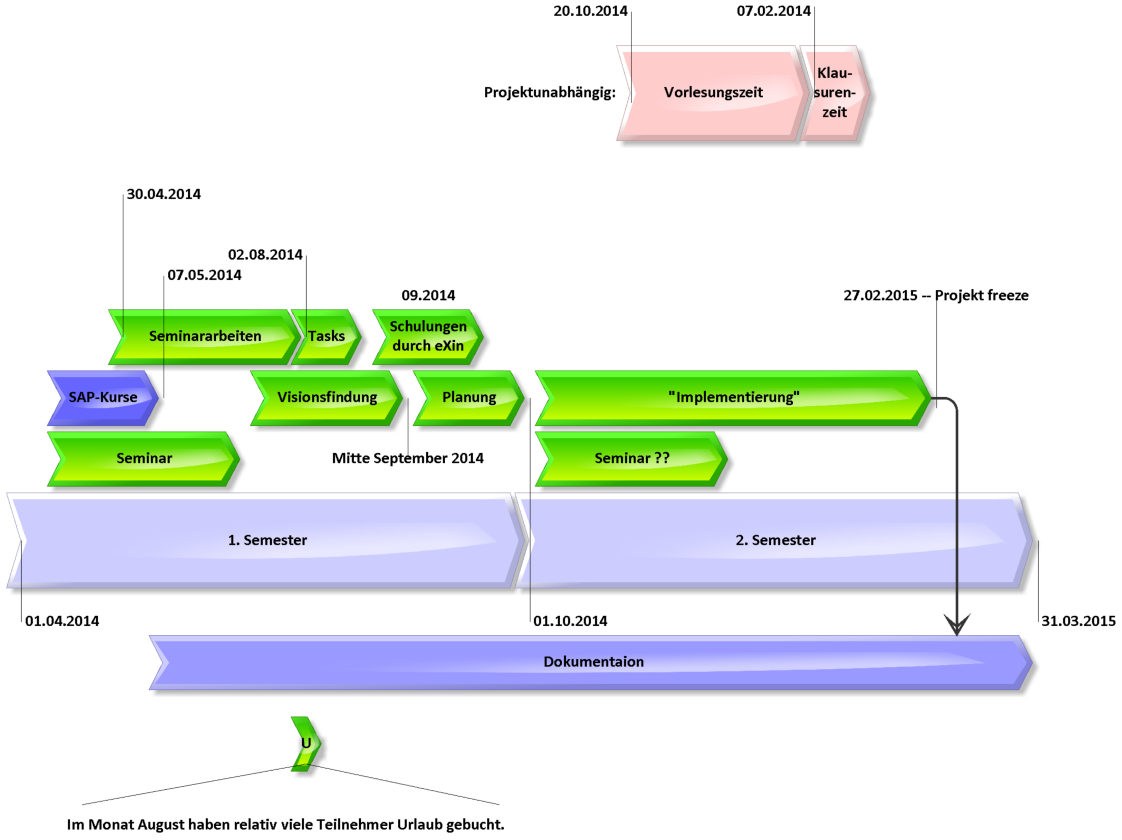


Abbildung 2: Projektsverlaufübersicht

Zu der grafischen Planung wurden folgende erste Meilensteine definiert:

Meilenstein A: Bekanntgabe der Themen der Seminararbeiten (30.04.2014) Damit sich die Projektgruppenmitglieder mit dem umfassenden Thema der In Memory Technik

sowie den Planungs- und Prognosewerkzeugen näher beschäftigen konnten, definieren die Betreuer geeignete Seminararbeitsthemen.

Meilenstein B: openSAP-Kurse abgeschlossen (07.05.2014) Im Rahmen der Projektgruppe wurde von allen internen Mitgliedern 2 SAP Online-Kurse absolviert:

1. An Introduction to SAP HANA by Dr. Vishal Sikka
   (Link: https://open.sap.com/course/hanaintro1)
2. Introduction to Software Development on SAP HANA by Thomas Jung
   (Link: https://open.sap.com/course/hana1)

Die Inhalte der Kurse vermitteln die Grundlagen der In Memory Datenbank SAP HANA und die Entwicklung mit SAP HANA Studio.

Meilenstein C: Ende der Seminar-Präsentationen (Mitte Juni) Ab dem 08.05.2014 wurden die Seminararbeitsthemen von den Projektmitgliedern erarbeitet. Die Tabelle 3 zeigt die entsprechende Zuordnung von Projektgruppenmitglied und Seminarthema. Hierzu wurde vom jeweiligen Projektgruppenmitglied eine Ausarbeitung mit einem Umfang von 10 bis 15 Seiten erstellt[2]; zudem wurde das jeweilige Thema in einer 30 minütigen Präsentation den anderen Projektgruppenmitgliedern vorgestellt[3].

Meilenstein D: Seminarphase abgeschlossen (02.08.2014) Die Seminararbeiten wurden von allen Projektmitgliedern abgegeben und im SVN hochgeladen.

Die Vorstellung der Grobplanung zum Start der Projektgruppe ist hiermit abgeschlossen. Im folgenden Abschnitt wird auf die Feinplanung eingegangen, welche nach Abschluss der Visionsfindung - und damit der Aufgabenstellung an die Projektgruppe - vorgenommen wurde.

[2] Die Ausarbeitungen können im Anhang dieses Dokumentes gefunden werden.

[3] Aufgrund des ähnlichen Themenbereiches von „BW, BPC, SEM - Was leisten heutige Planungssysteme der SAP AG?“ und „Planungs- und Prognosewerkzeuge und ihre Stärken und Schwächen“ wurde von den beiden Projektgruppenmitgliedern eine Gemeinschaftsarbeit und -präsentation erstellt.

Name	**Rolle**
Mariska Janz	BW, BPC, SEM - Was leisten heutige Planungssysteme der SAP AG?
Jonas Schlemminger	Bewertung des Einsatzes von prediktiven Methoden und Werkzeugen im SAP-Umfeld (SAP Predictive Analysis)
Benjamin Hemken	Klassische vs. agile Softwareentwicklung: Einsatz von SCRUM in der Projektgruppe
Igor Perelman	Planungsprozesse im Unternehmen aus fachlicher und organisatorischer Sicht und deren Ziele
Daniel Stratmann	Statistische Verfahren zur Fortschreibung historischer Daten
Steffen Scheer	Chancen und Herausforderungen in der klassischen Unternehmensplanung
Ivaylo Ivanov	Design Thinking
Eduard Rajski	Analytical Capabilities of SAP HANA: integration of R and Excel?
Rima Adhikari (K.C.)	Vor- und Nachteile von inMemory-Computing
Abdulmasih Hadaya	Planungs- und Prognosewerkzeuge und ihre Stärken und Schwächen
Farhad Gavan El-Yazdin	Integrierte Unternehmensplanung

Tabelle 3: Zuordnung und Thema der Seminararbeiten

4.3.2 Feinplanung

Im Zuge der nächsten Phase der Projektgruppe, in der eine praktische Aufgabenstellung umgesetzt und realisiert werden soll, war eine aktualisierte und verfeinerte Version der Meilensteinplanung erforderlich. Dieser Meilensteinplan wurde zunächst jedoch noch in Hinblick auf den vorherigen Praxispartner, der eXin AG erstellt. Da die Zusammenarbeit mit der eXin AG am 13.08.2014 unerwartet endete, waren alle bisherigen Planungen hinfällig. Vor allem hatte die Projektgruppe ab diesem Zeitpunkt die Aufgabe, in Eigenarbeit eine Vision - und damit ein praxisnahe Aufgabenstellung - zu finden, da seitens des ehemaligen Projektpartners keine konkrete Aufgabe gestellt wurde. Parallel hierzu lief die Suche nach einem neuen Praxispartner, was am 17.09.2014 mit der Zusage der AS Impro GmbH erfolgreich abgeschossen werden konnte. Aus diesen Gründen war eine Anpassung der Meilensteinplanung unerlässlich. Dazu wurde den Mitgliedern der Projektgruppe seitens der Betreuung und auch vom neuen Praxispartner empfohlen, die Ergebnisse des bisherigen Meilensteinplanes in der aktualisierten Planung ersichtlich zu machen. Dies soll darstellen, ab wann der Meilenstein als erfüllt gewertet werden kann. Nach mehreren Revisionen, die im Laufe des Oktobers angefertigt wurden, hat sich die Projektgruppe auf eine vorerst finale Version des Meilensteinplans (siehe Tabellen 4 und 5) geeinigt. Die grafische Repräsentation des Meilensteinplanes ist in Abbildung 3 zu finden.

Aufgrund von Komplikationen beziehungsweise Updates der SAP HANA Instanz konnte diese für insgesamt ca. 3 Wochen nicht genutzt werden. Ebenso haben irreparable Probleme der ursprünglich genutzten SAP HANA Instanz einen Wechsel auf eine neue Instanz erfordert, was mit einem Zeitaufwand von einer weiteren Woche verbunden war (Migration der Daten). Weiterhin gab es Schwierigkeiten bei der Nutzung der grafischen Modellierung der Prognosealgorithmen im Application Function Modeller (AFM) in SAP HANA[4]. Aus diesem Grund wurden nur noch selbst erstellte SQL-Skripte für die Berechnung der Modelle und Prognosen verwendet. Diese Punkte führten dazu, dass die beiden Meilensteine „Pilothypothese vorbereitet, überprüft und evaluiert“ und „Vereinfachtes Data Warehouse in SAP HANA fertiggestellt“ nicht zeitgemäß beendet wurden. Stattdessen wurden durch die immer wiederkehrenden Schwierigkeiten neue Lösungsmöglichkeiten realisiert, um dennoch angemessene Ergebnisse zu erzielen.

Die Beschreibung des Projektmanagements ist hiermit abgeschlossen. Im nächsten Kapitel werden die Programme vorgestellt, die für die aktive Durchführung des Projektes und der Aufgabenstellung benötigt werden.

[4]Siehe hierzu Kapitel 13.

Datum (Fertig)	**Meilenstein**	**Ergebnis**
19.10.2014	Vision erstellt	Dokumentation • Vision mitsamt Hypothesen
08.12.2014	Pilothypothese vorbereitet, überprüft und evaluiert	Dokumentation • Ziel der Prognose • Vorgehensweise (Datenselektion, Algorithmen, sonstige Parameter) • Ergebnisse der Analyse • Evaluation der Ergebnisse
19.12.2014	Vereinfachtes Data Warehouse in HANA fertiggestellt	Java-Code • Packages • Klassen • Interfaces • Test-Klassen (wenn nötig) Dokumentation • Diagramme (soweit nötig) • Schemas • Klassendiagramme • Tests • Klasse/Testumgebung • Test-Parameter • Daten in HANA • Views in HANA

Tabelle 4: Meilensteinplanung ab Oktober Teil 1

Datum (Fertig)	**Meilenstein**	**Ergebnis**
02.03.2015	Alle Hypothesen vorbereitet, überprüft und evaluiert	Dokumentation • Ziel der Prognose • Vorgehensweise (Datenselektion, Algorithmen, sonstige Parameter) • Ergebnisse der Analysen • Evaluation der Ergebnisse in Bezug auf Standard-Prognose
09.03.2015	Planung und Vorbereitung für Cebit abgeschlossen	Dokument im Confluence • Mögliche Fragen und Antworten zum Projekt • Wer ist wann am Stand • sonstige Aufgaben
23.03.2015	Dokumentation abgeschlossen	Dokumentation • Vorstellung (Organisationsstruktur u.ä.) • Vision/Idee/Hypothesen • Herangehensweise • Schemata, Diagramme zu programmierten Artefakten • Tests • Analysen (Parameter, Vorgehensweise, Ergebnisse, Evaluation) • Gesamtergebnis • Evaluation des Gesamtergebnisses • Seminararbeiten • Protokolle • Meilensteinplanung • Berichte • Literatur
31.03.2015	Projekt abgeschlossen	Dokumentation JAR-Execution-File Präsentationsfolien Handout (wenn nötig)

Tabelle 5: Meilensteinplanung ab Oktober Teil 2

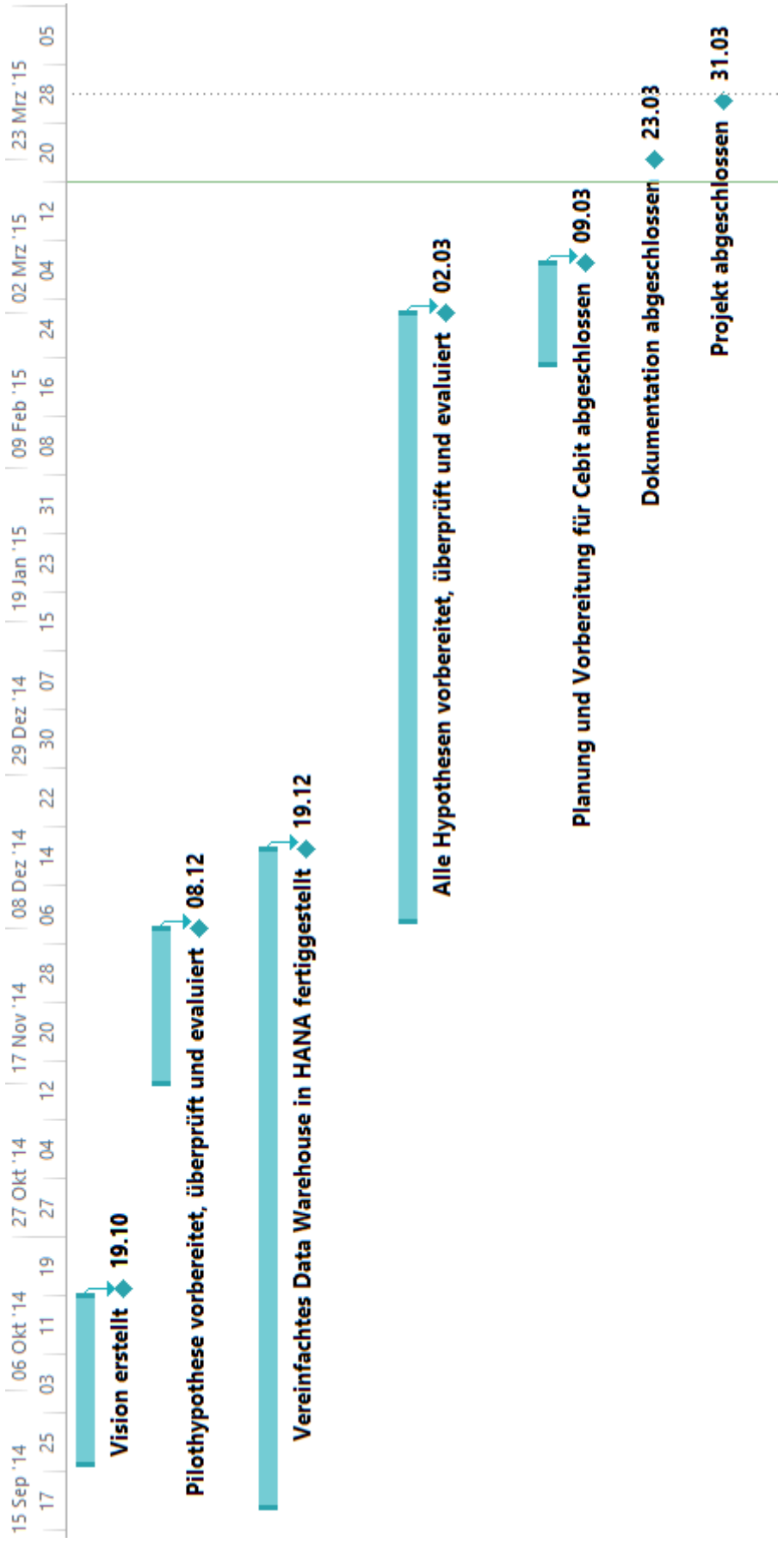


Abbildung 3: Meilensteinplan ab Oktober

5 Auswahl der Software

In diesem Kapitel sollen die für die von der Projektgruppe genutzten Applikationen aufgelistet werden, die im Prognoseprozess verwendet werden.

SAP HANA SAP HANA ist eine von der SAP AG entwickelte Datenbank, die im Jahre 2010 erstmals vorgestellt wurde und deren Basis die In Memory Technologie bildet. Das Speichern der zur Verfügung gestellten Daten erfolgt dabei im Arbeitsspeicher. Daraus entsteht im Vergleich zur Speicherung auf magnetischen Festplatten ein bedeutsamer Geschwindigkeitsvorteil [SAP14a].

SAP HANA gehört zu den relationalen In Memory Datenbanken, die auf SQL basieren. Die SAP AG stellt dem Kunden bei der Nutzung von SAP HANA verschiedene Optionen zur Verfügung; so kann die Lösung entweder als Kombination von Hardware und Software (Appliance) oder in der Cloud verwendet werden. Als Betriebssystem für die Datenbank wird SUSE-Linux verwendet. Eine der wichtigsten Funktionen von SAP HANA ist die Möglichkeit, die Analysen von großen Datenmengen schnell in Echtzeit durchzuführen. Die Funktionsweise basiert auf der Kombination von spaltenorientierten und zeilenorientierten Datenbanktechnologien. Durch das parallele Ausführen von Prozessen durch Verwendung mehrkerniger CPU-Architekturen kann das volle Potenzial des In Memory Ansatzes ausgeschöpft werden. Für die Entwicklungsumgebung und Administration der Datenbank wird das auf Eclipse basierende *SAP HANA Studio* verwendet. Die Datenquellen können hierbei heterogen sein, es können zum Beispiel Daten aus SAP ERP bzw. SAP BW nach SAP HANA extrahiert werden [SAP14a].

SAP HANA verwendet zur Kommunikation mit dem Datenbank-Management-System (DBMS) die Schnittstellen ODBC und Java Database Connectivity (JDBC). In SAP HANA stellt der *Index Server* die Hauptkomponente dar. An dieser Stelle erfolgt das Speichern der Daten. Ebenfalls befinden sich hier die Engines für die Datenverarbeitung. Um die Erstellung von Prozeduren zu ermöglichen, wird *SAP HANA SQLScript* als eigene Scriptsprache verwendet. Außerdem besteht die Möglichkeit, die Programmiersprache *R* für statistische Analysen zu nutzen [SAP14a].

Das Entwicklungs- und Administrationstool für die Datenbank SAP HANA heißt *SAP HANA Studio* und basiert auf Eclipse. SAP HANA Studio wird für die Erstellung der Entwicklungs-Objekte, wie Datenmodelle oder *server-side code files*, verwendet, die mit anderen Entwicklern durch Nutzung des *SAP HANA Repository* geteilt werden. Die Funktionalität des Repository ermöglicht es, dass die Arbeit der Entwicklungsteams gleichzeitig auf den gleichen Entwicklungsobjekten vorgenommen werden kann [SAP14a].

Predictive Analytics Library – PAL SAP stellt Funktionen für HANA zur Verfügung, die in die Bibliotheken PAL und Business Function Library (BFL) unterteilt sind.

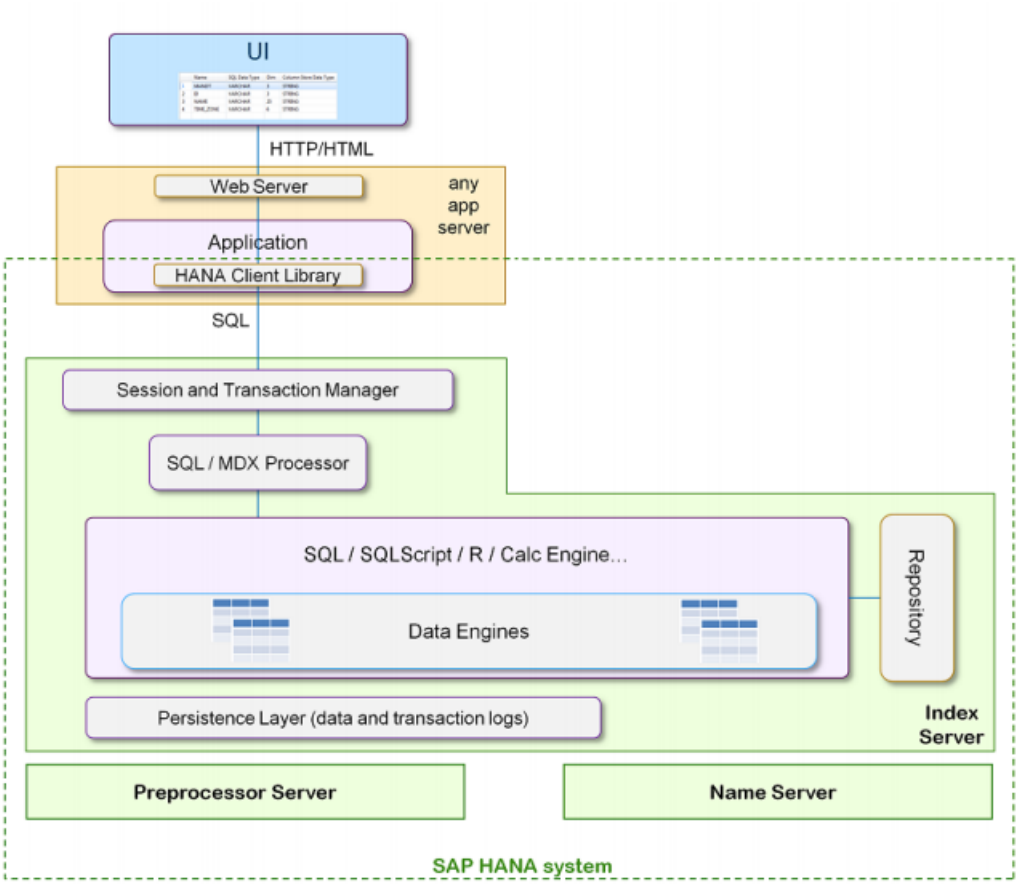


Abbildung 4: SAP HANA Architektur [SAP14a, S. 14]

Predictive Analysis Library (PAL) enthält Funktionen, die per SQLScript aus HANA heraus aufgerufen werden können. Die Algorithmen können für die Durchführung verschiedener prädiktiver Analysen angewendet werden, wie z.B. die lineare Beziehung zwischen zwei oder mehr Variablen, um die zukünftige Entwicklung einer bestimmten Variable zu ermitteln. Die Algorithmen können in neun Kategorien klassifiziert werden [SAP15]:

- Clustering
- Classification
- Regression
- Association
- Time series
- Preprocessing
- Statistics
- Social Network Analysis

- Miscellaneous

Die Algorithmen in den Kategorien *Regression* und Zeitreihen (*Time series*) sind insbesondere für den Zweck der Vorhersage des Stromverbrauchs anhand von Vergangenheitsdaten oder anderen Faktoren relevant.

Application Function Modeler – AFM Der AFM ist ein grafischer Editor für SAP HANA Studio (siehe Abb. 5), durch den eine PAL- oder BFL-Funktion zur Application Function Library (AFL) Model-Datei hinzugefügt werden kann. Eine Anpassung der Parameter und der Ein-/Ausgabe-Tabellentypen ist möglich (ohne SQLScript). Der Benutzer kann eine Prozedur aufrufen und erhält ein Ergebnis. Parallel wird automatisch ein SQLScript-Code generiert, der gespeichert und zu einem späteren Zeitpunkt wieder verwendet werden kann.

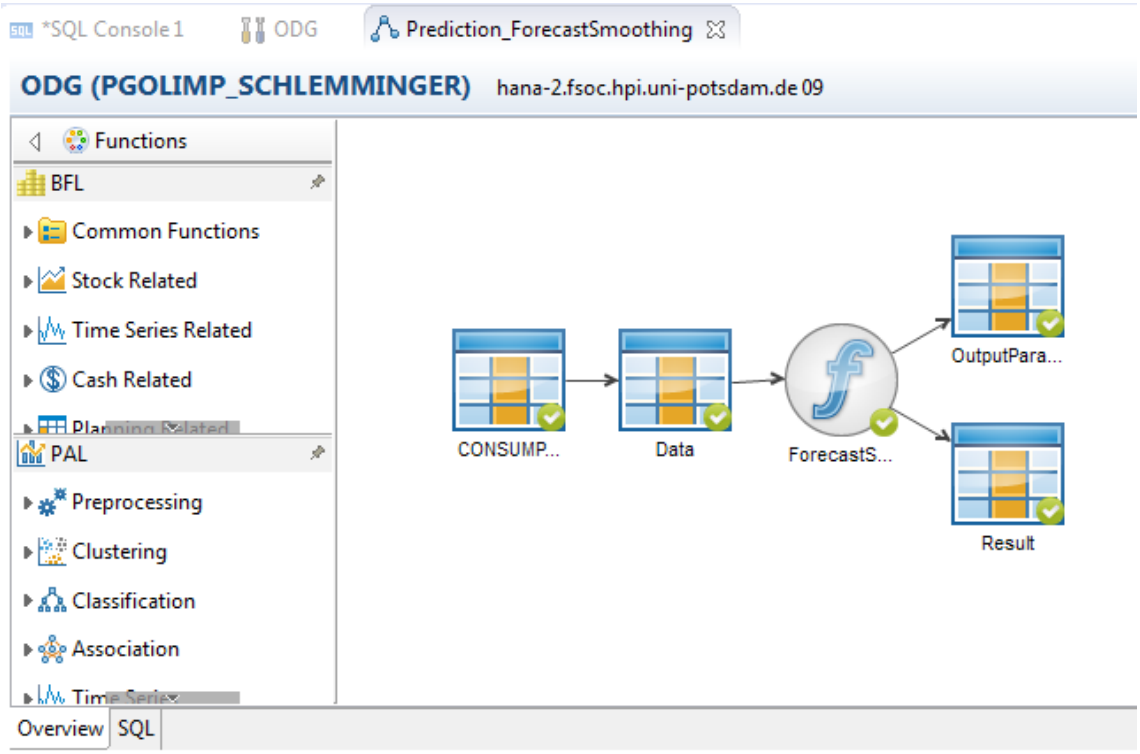


Abbildung 5: Application Function Modeler (AFM)

SAP Predictive Analysis – PA SAP Predictive Analysis ist eine Desktop-Anwendung aus dem Hause SAP, die im Gegensatz zu klassischen, vergangenheitsorientierten BI-Tools die Erstellung statistischer Vorhersagemodelle ermöglicht [Gra13]. Die Software arbeitet dabei in einer Eclipse-basierten visuellen Modellierungsoberfläche, in der Nutzer vordefinierte Statistikalgorithmen auf eigene Datensätze anwenden können [Gra13]. So sind Rückschlüsse auf künftige Entwicklungen möglich [Gra13]. Dabei lassen sich geschäftliche Daten mit öffentlichen Quellen in Bezug setzen [Gra13]. Mögliche Anwendungsszenarien

sind beispielsweise Cross-Selling-Analysen, Kundenschwundanalysen und Vertriebsprognosen [Gra13].

SAP Predictive Analysis kann prinzipiell stand-alone genutzt werden (siehe Abbildung 6) [Gra13]. In diesem Fall laden Anwender ausschließlich lokal vorliegende Datensätze aus Flat Files, beispielsweise Excel-Tabellen oder strukturierten Comma Seperated Values (CSV)-Dateien [Gra13].

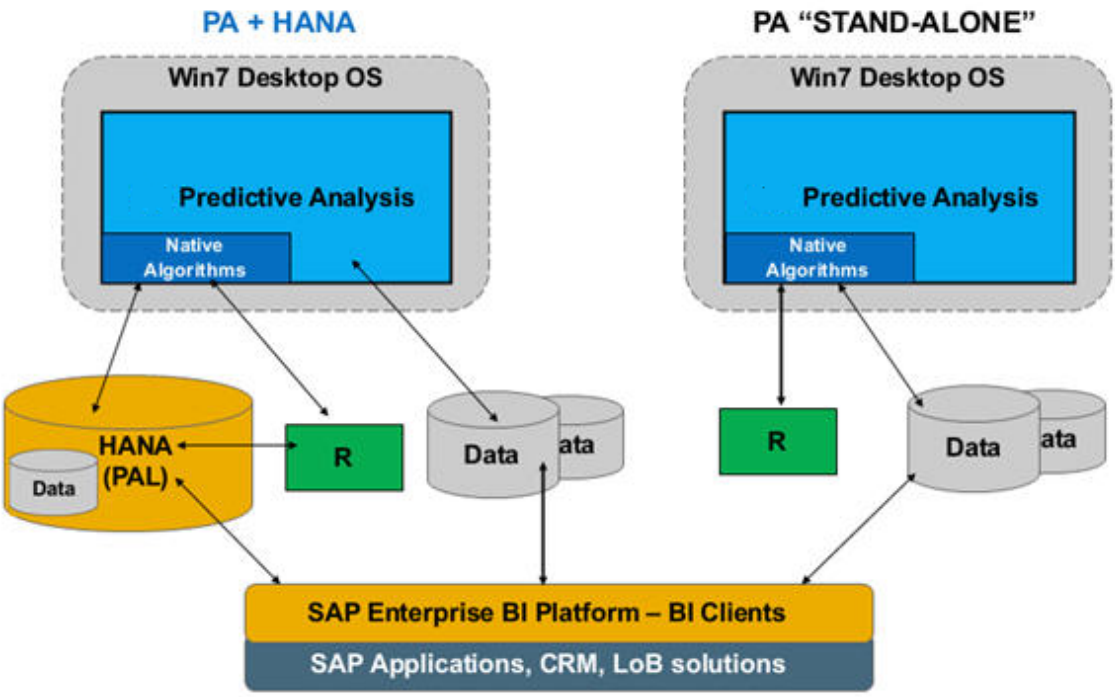


Abbildung 6: Nutzungsvarianten von SAP PA[Gra13]

Weitere Geschäftsdaten bezieht die Lösung aus angeschlossenen Datenbanken wie SAP Sybase IQ [Gra13]. Sie werden über die offene Schnittstelle JDBC angebunden [Gra13]. Darüber ist auch die Integration von Lösungen aus dem SAP-BusinessObjects-Portfolio (Frontend und Backend) möglich [Gra13]. In SAP Predictive Analysis erstellte Analysemodelle oder auch Visualisierungen lassen sich so direkt in Reports einbetten oder weiterverarbeiten [Gra13]. Die Visualisierungen erzeugt SAP Predictive Analysis mithilfe des integrierten SAP-eigenen Charting Visualization Object Models (CVOM), das auch in SAP-BusinessObjects-Lösungen eingesetzt wird [Gra13]. Informationen lassen sich damit in zahlreichen Formen darstellen, von einfachen Diagrammen bis hin zu komplexen Korrelationsplots [Gra13].

Im Gegensatz zu Prognoselösungen anderer Anbieter lässt sich SAP Predictive Analysis direkt in SAP HANA integrieren (siehe Abbildung 6) [Gra13]. SAP HANA verfügt dazu seit Version 1.0 Service Pack 3 über eine vordefinierte Predictive Analysis Library (PAL) [Gra13]. Sie enthält Algorithmen für die prognostische Analyse großer Datenmen-

gen [Gra13]. SAP Predictive Analysis spricht diese Algorithmen direkt an und orchestriert deren Aufruf direkt auf der SAP HANA Plattform selbst [Gra13]. Während SAP Predictive Analysis also auf einem Desktop-Client läuft, erfolgt die eigentliche Analyse in der In-Memory-Datenbank [Gra13].

Microsoft Excel Microsoft Excel ist eine Tabellenkalkulation vom Softwarehersteller Microsoft und ist Teil des Microsoft Office Paketes [Wal13][Sch07]. Excel erlaubt durch verschiedene Hilfsmittel in Form von statistischen Funktionen und Assistenten auch Personen, die nicht mit umfangreichen mathematischen Kenntnissen ausgestattet sind, die transparente Erzeugung von komplizierten Statistiken[Sch07]. Die verarbeitbaren Datentypen sind nur im Bezug auf die für den Kontext genutzten Hilfsmittel begrenzt[Sch07]. Je nach Kenntnisstand kann der Nutzer einen Assistenten oder die Excel Syntax verwenden, um Formeln zu nutzen und Statistiken zu erstellen[Sch07]. Weiterhin bietet Excel verschiedene Visualisierungsmöglichkeiten zur besseren Veranschaulichung [Sch07][Wal13].

6 Anforderungsanalyse

Es soll innerhalb der Projektgruppe festgestellt werden, ob mit Hilfe der In Memory Datenbank SAP HANA eine Optimierung der mittelfristigen Unternehmensplanung für Energieunternehmen erreicht werden kann. Die Haupthypothese „Je mehr Daten- und Datenfeatures für die Prognose und Simulation zur Verfügung stehen, desto höher die Prognosegenauigkeit“, soll im Rahmen des Projektes verifiziert werden. Insbesondere soll evaluiert werden, welche Kombination von Daten und welcher Algorithmus die höchste Prognosegenauigkeit für die kurz- bis mittelfristige Planung bietet. Dieser kurz- bis mittelfristige Zeitraum bezeichnet in der Energieplanung einen Zeitraum von ca. einem Monat. In den folgenden Abschnitten werden die Anforderungen für die Durchführung des Projektes definiert. Hierzu zeigen die Tabellen 6, 7, 8 und 9 die erstellten Muss-, Wunsch-, Abgrenzungs-, und Qualitätsanforderungen. Die Musskriterien sind für die Durchführung des Projektes unabdingbar. Die Wunschkriterien sind nicht unabdingbar, sollten jedoch so gut wie möglich realisiert werden. Die Abgrenzungskriterien sollen deutlich machen, welche Ziele mit dem Produkt *nicht* realisiert werden. Die Qualitätsanforderungen definieren Qualitätsaspekte zur Durchführung des Projektes.

Musskriterien	
Nummer	**Kriterium**
1.1	Identifikation relevanter Daten für die Prognoseerstelltung.
1.2	Lokalisierung potentieller (heterogener) Datenquellen.
1.3	Die Prognosen basieren auf historischen Daten und die Prognose selbst wird ebenfalls für einen bereits vergangenen Zeitraum durchgeführt.
1.4	Vorverarbeitung der Daten für den Datenimport in SAP HANA.
1.5	Identifikation geeigneter Algorithmen für die Prognoseerstellung.
1.6	Transformation der Daten für die Anwendung von Algorithmen in SAP HANA.
1.7	Durchführung der Prognosen in SAP HANA.
1.8	Messung der Zuverlässigkeit/Genauigkeit der Prognose.
1.9	Visualisierung/Vergleich der Ergebnisse mit Realdaten.

Tabelle 6: Musskriterien

Wunschkriterien	
Nummer	**Kriterium**
2.1	Automatisierung des gesamten ETL-Prozesses.
2.2	Export von Auswertungen nach Excel.
2.3	Automatisierte Visualisierung der Ergebnisse.

Tabelle 7: Wunschkriterien

Abgrenzungskriterien	
Nummer	**Kriterium**
3.1	Es werden keine Echtzeitanalysen durchgeführt.
3.2	Auf die Entwicklung eines Web-Front-Ends wird verzichtet.

Tabelle 8: Abgrenzungskriterien

Qualitätsanforderungen	
Nummer	**Kriterium**
4.1	Die Prädiktionsgenauigkeit der Algorithmen wird anhand von festgelegten Fehlerkennzahlen evaluiert.
4.2	Der korrekte Datenimport der Rohdaten in das System muss sichergestellt werden.
4.3	Es muss sichergestellt werden, dass die Daten vor der Anwendung der Algorithmen bereinigt werden.

Tabelle 9: Qualitätsanforderungen

Software- Server- und Entwicklungsumgebung Die folgenden Komponenten werden benötigt, um das Projekt zu realisieren.

- Server-Software: SAP HANA Datenbank SP-07, -08, -09
- SAP HANA Studio
- Tabellenkalulationsprogramm
- OpenVPN-Client
- Eclipse
- SAP HANA Studio
- JDK
- SVN-Client
- SFTP-Client
- Jira
- Confluence
- Predictive Analysis

Produkteinsatz Ziel dieses Projektes ist die Beantwortung der Frage, welche Kombination von Daten unter Verwendung welcher Algorithmen die beste Prognose für den kurz- bis mittelfristigen Stromverbrauch erstellt. Dabei handelt es sich nicht um eine klassische Softwareentwicklung bei der am Ende ein fertiges Produkt entsteht. Vielmehr werden Softwareartefakte entwickelt, die zur Beantwortung der Fragestellung benötigt werden. Mit den Ergebnissen dieser Fragestellungen ist anschließend eine Realisierung als produktive Software denkbar.

Anwendungsbereiche und Zielgruppen Ein denkbarer Anwendungsbereich befindet sich in der regionalen beziehungsweise überregionalen operativen und taktischen Versorgungsplanung mit Energie. Potentielle Unternehmen könnten zum Beispiel einzelne Energieversorger sein. Ebenfalls ist eine Verwendung der Ergebnisse an einer Strombörse denkbar.

Die Anforderungsanalyse ist hiermit abgeschlossen. Im nächsten Kapitel werden die gefundenen potentiellen Datenlieferanten und die Beschaffenheit der Quelldaten erläutert.

7 Beschreibung der Datenbasis

In diesem Kapitel werden die für diese Arbeit verwendeten Datenlieferanten näher beschrieben. Folgende Datenquellen und -lieferanten werden verwendet:

- Astro-Daten (z.B. Sonnenaufgangs- und Untergangsdaten)
- Strompreise für Deutschland (Haushalts- und Industriestrompreise)
- Stromverbrauchsdaten
- Temperaturdaten
- Zeitdimension

Auf das Handlungsfeld „Stromkonsum“ wirkt eine Vielzahl von Faktoren. Zunächst hat sich die Projektgruppe auf die folgenden Einflussfaktoren konzentriert: Temperatur, Wochentage, Stromverbrauchsdaten sowie den Strompreis. In den folgenden Abschnitten erfolgt eine Beschreibung der bisher verwendeten Datenquellen.

7.1 Astro-Daten

Die Astro Daten stammen von der Internetseite `www.galupki.de`. Diese Seite enthält Informationen über Sonnenaufgangs-, Sonnenuntergangsdaten, Dämmerungszeiten, Blaue Stunde, Goldene Stunde, Mondphase, Mondaufgang, Monduntergang berechnen, Kalenderdaten (z.b. Feiertagskalender), die sich im CSV-Format exportieren lassen und damit direkt in andere Anwendungen importiert werden können. Aus der Datei mit den Astrodaten werden zunächst die Kalender- und Feiertagsangaben für die Prognosen als Inputdaten verwendet. Die Datenquelle kann unter `http://galupki.de/kalender/sunmoon.php` gefunden werden.

7.2 Strompreise

Die Strompreise stammen von Eurostat. Dies ist der führende Anbieter hochwertiger Statistiken für Europa. Eurostat ist das statistische Amt der Europäischen Union in Luxemburg. Seine Aufgabe ist es, verschiedene Statistiken auf europäischer Ebene zu liefern, die Vergleiche zwischen Ländern und Regionen ermöglichen. Die Strompreise sind aufgeteilt nach Haushalten und Industrie, bezogen auf das jeweilige Land. Die Datenquelle kann unter `http://ec.europa.eu/eurostat/data/database` gefunden werden.

7.3 Stromverbrauchsdaten

Die Quelle der Stromverbrauchsdaten ist das Datenportal des European Network of Transmission System Operators for Electricity (ENTSO-E). Die ENTSO-E als statis-

tische Datenbank umfasst eine Reihe von historischen Datensätzen über Energiesysteme von ENTSO-E-Mitglieder. Die Stromverbrauchdaten umfassen: Elektrizitätsversorgung, Energieverbrauch nach Wirtschaftszweig, gesamter Energieverbrauch oder Energieverwendung der Betriebe des verarbeitenden Gewerbes. Die Quelle der Daten ist unter `https://www.entsoe.eu/data/data-portal/consumption/Pages/default.aspx` zu erreichen.

7.4 Temperatur-Daten

Die Temperatur-Daten stammen von Deutschen Wetterdienst (DWD). Der Deutsche Wetterdienst ist eine teilrechtsfähige Anstalt des öffentlichen Rechts im Geschäftsbereich des Bundesministeriums für Verkehr und digitale Infrastruktur. Seine Aufgabe ist die Erfüllung der meteorologischen Erfordernisse aller Wirtschafts- und Gesellschaftsbereiche in Deutschland. Die Quelle für die Temperaturdaten ist unter `http://www.dwd.de/bvbw/appmanager/bvbw/dwdwwwDesktop?_nfpb=true&_pageLabel=dwdwww_wir_ueberuns&_nfls=false` verfügbar.

7.5 Zeitdimension

Die Zeitdimension ist eine interne SAP HANA-Tabelle, die durch SAP HANA generiert werden kann. Dies kann komfortabel über einen Assistenten erledigt werden. Hierbei wird unter anderem festgelegt, wie hoch die Granularität dieser Tabelle sein kann (z.B. auf Sekunden-, Minuten- oder Stundenbasis). Hier wurde die Granularität auf Stundenbasis ausgewählt, da alle weiteren Datenquellen auf Stundenbasis vorliegen. Abbildung 7 zeigt den Konfigurationsassistent der Zeittabelle und Abbildung 8 zeigt einen Ausschnitt aus der erstellten Tabelle.

Die Vorstellung der Datenlieferanten ist hiermit beendet. Im folgenden Kapitel werden die Datenstrukturen und die daraus abgeleiteten Kennzahlen näher beschrieben.

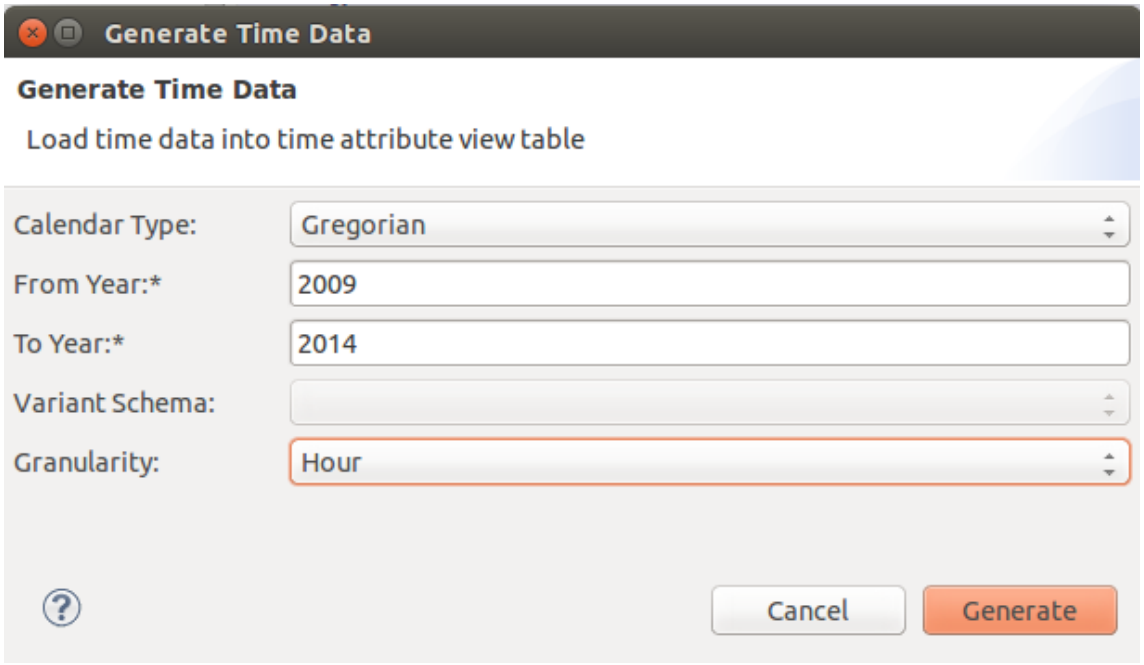


Abbildung 7: Der Konfigurationsassistent zur Zeittabelle in SAP HANA

SQL Result

```
select * from "_SYS_BI"."M_TIME_DIMENSION" WHERE "YEAR" = 2009 order by "DATE_SQL" asc
```

	DATETIMESTAMP	DATE_SQL	YEAR	QUARTER	MONTH	WEEK	WEEK_YEAR	DAY_OF_WEEK	DAY	HOUR	MINUTE	SECOND	CALQUARTER	CALM
1	01.01.2009 00:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	00	00	00	20091	20090
2	01.01.2009 01:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	01	00	00	20091	20090
3	01.01.2009 02:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	02	00	00	20091	20090
4	01.01.2009 03:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	03	00	00	20091	20090
5	01.01.2009 04:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	04	00	00	20091	20090
6	01.01.2009 05:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	05	00	00	20091	20090
7	01.01.2009 06:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	06	00	00	20091	20090
8	01.01.2009 07:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	07	00	00	20091	20090
9	01.01.2009 08:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	08	00	00	20091	20090
10	01.01.2009 09:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	09	00	00	20091	20090
11	01.01.2009 10:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	10	00	00	20091	20090
12	01.01.2009 11:00:00.0	01.01.2009	2009	01	01	01	2009	03	01	11	00	00	20091	20090

Abbildung 8: Ausschnitt aus der SAP HANA Zeittabelle

8 Datenstruktur

In diesem Kapitel werden die Strukturen der genutzten Daten beschrieben. Diese Tabellenstrukturen werden im Zuge des Datenimports nach HANA in das Schema **olimp** eingetragen.

8.1 EEX-Daten

Die EEX-Daten beinhalten Transparenzdaten des europäischen Marktes. Sie sind folgendermaßen aufgeteilt:

8.1.1 EEX-Ex-Ante-Daten

Dies sind die Planungsdaten der EEX. Sie eignen sich in der Form nicht für die Planung. Denkbar wäre jedoch die eignen Prognosen mit den Prognosen der EEX verglichen werden, um einen Rückschluss auf die eigene Prognosegüte zu bekommen. Die Beschreibung erfolgt über die Tabellen 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 und 23.

Planned Generation Solar

	Geplante Generation von Solarenergie pro Stunde	
Bedeutung	Beschreibung	Geplante stündliche Generation von Solarenergie in MegaWatt.
	Dimensionen	Typ, Region, Zeit, Solarenergie,
	Exemplarische Analysefragen	Inwieweit weicht im Jahre 2011 die geplante Generation von Solarenergie von den tatsächlichen Werten ab?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 10: Kennzahlensteckbrief erwartete Solarenergie

Type	Connecting-Area	Timestamp	Expected-SolarEnergy	Publication-Timestamp	Modification-Timestamp
ESPL					

Tabelle 11: EEX_Ex_ante_Planned_Generation_Solar

Planned Generation Wind

	Geplante Generation von Windenergie pro Stunde	
Bedeutung	Beschreibung	Geplante stündliche Generation von Windenergie in MegaWatt
	Dimensionen	Typ, Region, Zeit, Windenergie,
	Exemplarische Analysefragen	Inwieweit weicht im Jahre 2011 die geplante Generation von Windenergie von den tatsächlichen Werten ab?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 12: Kennzahlensteckbrief erwartete Windenergie

Type	Connecting-Area	Timestamp	Expected-Wind-Energy	Publication-Timestamp	Modification-Timestamp
EWPL					

Tabelle 13: EEX_Ex_ante_Planned_Generation_Wind

Non usuability Generation

	Geplante Nichtverfügbarkeitskapazität pro Stunde	
Bedeutung	Beschreibung	Geplante Nichtverfügbarkeiten von Stromherstellern in MegaWatt.
	Dimensionen	Typ, Region, Quelle, Zeit, Nichtverfügbarkeitskapazität, Status
	Exemplarische Analysefragen	Zu welchen Zeiten war die geplante Nichtverfügbarkeitskapazität 2011 am höchsten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 14: Kennzahlensteckbrief erwartete Nichtverfügbarkeitskapazität

Type	Country	Source	NUM-Start-Date	NUM-End-Date	Num-Capacity	Timestamp	Status	Publication-Timestamp	Modification-Timestamp
NU-GL									

Tabelle 15: EEX_Ex_ante_Non_usability_Generation

Available Capacity

	Geplante Kapazität aller Einheiten pro Stunde	
Bedeutung	Beschreibung	Geplante stündliche Kapazität aller Einheiten zur Stromproduktion in MegaWatt.
	Dimensionen	Typ, Quelle, Region, Zeit, verfügbare Kapazität
	Exemplarische Analysefragen	Welche Quellen weisten laut Plan 2011 die höchste Kapazität auf?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 16: Kennzahlensteckbrief geplante Kapazität aller Einheiten

Type	Source	Country	Timestamp	Available-Capacity	Publication-Timestamp	Modification-Timestamp
ACIL						

Tabelle 17: EEX_Ex_ante_Available_Capacity

Planned Energy

	Geplante maximale Energieerzeugung pro Stunde	
Bedeutung	Beschreibung	Geplante maximale stündliche Erzeugung von Energie in MegaWatt.
	Dimensionen	Typ, Einheit, Zeit, Energie
	Exemplarische Analysefragen	Inwieweit verändern sich die Prognosewerte der geplanten maximalen Erzeugungsenergie von 2009 zu 2010?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 18: Kennzahlensteckbrief geplante maximale Energieerzeugung

Type	UnitID	Timestamp	Planned-Energy	Publication-Timestamp	Modification-Timestamp
PLEL					

Tabelle 19: EEX_Ex_ante_Planned_Energy

Planned Generation

	Geplante Energieerzeugung pro Stunde	
Bedeutung	Beschreibung	Geplante stündliche Erzeugung von Energie auf das Land bezogen in MegaWatt.
	Dimensionen	Typ, Region, Zeit, Energie
	Exemplarische Analysefragen	Inwieweit verändern sich die Prognosewerte der geplanten Erzeugungsenergie von 2009 zu 2010?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 20: Kennzahlensteckbrief geplante Energieerzeugung bezogen aufs Land

Type	Country	Timestamp	Planned-Generation	Publication-Timestamp	Modification-Timestamp
CPGL					

Tabelle 21: EEX_Ex_ante_Planned_Generation

Sum Installed Capacity

	Summe installierter Kapazitäten pro Stunde	
Bedeutung	Beschreibung	Summe der installierten Kapazitäten über 100 MegaWatt in MegaWatt.
	Dimensionen	Typ, Quelle, Region, Zeit, Summierte Kapazität
	Exemplarische Analysefragen	Inwieweit verändern sich die Prognosewerte der summierten installierten Kapazitäten von 2009 zu 2010?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Nicht Relevant
	Prognose	Nicht Relevant
	Evaluation	Relevant

Tabelle 22: Kennzahlensteckbrief Summe installierter Kapazitäten über 100 MW

Type	Source	Con-necting-Area	Timestamp	SumInstall-edCapacity	Pub-lication-Timestamp	Modi-fication-Timestamp
SICL						

Tabelle 23: EEX_Ex_ante_Sum_installed_capacity

8.1.2 EEX-Ex-Post-Daten

Das sind die tatsächlichen Produktions- und Verbrauchsdaten. Mithilfe dieser Daten können entsprechend Planungen beziehungsweise Prognosen erstellt werden. Die Beschreibung erfolgt über die Tabellen 24, 25, 26, 27, 28, 29, 30, 31, 32 und 33.

Actual Generation

	Energieerzeugung pro Stunde	
Bedeutung	Beschreibung	Stündliche Erzeugung von Energie in MegaWatt.
	Dimensionen	Typ, Region, Zeit, Energie
	Exemplarische Analysefragen	Inwiefern unterscheiden sich die prognostizierten Werte aus dem Jahre 2012 von den tatsächlichen Werten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 24: Kennzahlensteckbrief tatsächliche Energieerzeugung

Type	Country	Timestamp	Actual-Generation	Publication-Timestamp	Modi-fication-Timestamp
APGL					

Tabelle 25: EEX_Ex_post_Actual_Generation

Generation Wind

	Windenergieerzeugung pro Stunde	
Bedeutung	Beschreibung	Stündliche Erzeugung von Windenergie in MegaWatt.
	Dimensionen	Typ, Region, Zeit, Energie
	Exemplarische Analysefragen	Inwiefern unterscheiden sich die prognostizierten Werte aus dem Jahre 2012 von den tatsächlichen Werten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 26: Kennzahlensteckbrief tatsächliche Windenergieerzeugung

Type	Country	Timestamp	Actual-Generation	Publication-Timestamp	Modification-Timestamp
AWPL					

Tabelle 27: EEX_Ex_post_Generation_Wind

Generation Solar

	Solarenergieerzeugung pro Stunde	
Bedeutung	Beschreibung	Stündliche Erzeugung von Solarenergie in MegaWatt.
	Dimensionen	Typ, Region, Zeit, Energie
	Exemplarische Analysefragen	Inwiefern unterscheiden sich die prognostizierten Werte aus dem Jahre 2012 von den tatsächlichen Werten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 28: Kennzahlensteckbrief tatsächliche Solarenergieerzeugung

Type	Country	Timestamp	Actual-Generation	Publication-Timestamp	Modification-Timestamp
ASPL					

Tabelle 29: EEX_Ex_post_Generation_Solar

Non usuability Generation

	Nichtverfügbarkeitskapazität pro Stunde	
Bedeutung	Beschreibung	Nichtverfügbarkeiten von Stromherstellern in MegaWatt.
	Dimensionen	Typ, Region, Quelle, Zeit, Nichtverfügbarkeitskapazität, Status
	Exemplarische Analysefragen	Inwieweit unterscheiden sich im Jahre 2011 die geplanten Werte für die Nichtverfügbarkeitskapazität von den tatsächlichen Werten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 30: Kennzahlensteckbrief tatsächliche Nichtverfügbarkeitskapazität

Type	Country	Source	NUM-Start-Date	NUM-End-Date	Num-Capacity	Time-stamp	Status	Publi-cat-ion-Time-stamp	Mod-ificat-ion-Time-stamp
NU-GL									

Tabelle 31: EEX_Ex_post_Non_usability_Generation

Previous Day Generation

	Energieerzeugung des Vortages pro Stunde	
Bedeutung	Beschreibung	Stündliche Erzeugung von Energie des Vortages in MegaWatt.
	Dimensionen	Typ, Region, Quelle, Zeit, Energie
	Exemplarische Analysefragen	Inwiefern unterscheiden sich die prognostizierten Werte aus dem Jahre 2012 von den tatsächlichen gestrigen Werten?
Datenquelle	Herkunft	EEX
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 32: Kennzahlensteckbrief tatsächliche Energieerzeugung des Vortages

Type	Country	Source	Timestamp	Previous-Day-Generation	Publication-Timestamp	Modi fication-Timestamp
PDGL						

Tabelle 33: EEX_Ex_post_Previous_Day_Generation

8.2 Verbrauchsdaten von Entso-E

Die Tabelle 34 enthält stündliche Stromverbrauchsdaten für die Jahre 2009 bis etwa August 2014 für Deutschland. Für jeden Tag existieren entsprechend 24 Datensätze. Die Beschreibung erfolgt über die Tabellen 34 und 35.

	Stromverbrauch pro Stunde	
Bedeutung	Beschreibung	Stündliche Stromverbrauchsdaten in MegaWatt.
	Dimensionen	Ort, Zeit, Verbrauch
	Exemplarische Analysefragen	Wann waren beim Stromverbrauch im Jahre 2011 die größten Stromverbräuche zu verzeichnen?
Datenquelle	Herkunft	Entso-E
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 34: Kennzahlensteckbrief Stromverbrauch

Country	Timestamp	Consumption
DE	2009-11-23 00:00:00	52689

Tabelle 35: Entsoe_Power_Consumption

8.3 Wetterdaten

Diese Wetterdaten stammen vom Deutschen Wetterdienst (DWD) und beinhalten neben verschiedenen Temperaturangaben auch Winddaten sowie Angaben zum Luftdruck und dem Bewölkungsgrad. Die Beschreibung erfolgt über die Tabellen 36, 37, 38, 39, 40, 41, 42, 43, 44 und 45.

	Lufttemperatur pro Stunde	
Bedeutung	Beschreibung	Stündliche Lufttemperaturmessungen in Grad Celsius.
	Dimensionen	Ort, Zeit, Lufttemperatur
	Exemplarische Analysefragen	Wann waren im Jahre 2011 Extremwerte zu verzeichnen?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 36: Kennzahlensteckbrief Lufttemperatur

Timestamp	AirTemp	Lat	Lon
2009-11-23 00:00:00			

Tabelle 37: DWD_Weather_AirTemp

	Bewölkungsgrad pro Stunde	
Bedeutung	Beschreibung	Stündliche Bewölkungsgradmessungen in Achteln.
	Dimensionen	Ort, Zeit, Bedeckungsgrad
	Exemplarische Analysefragen	Wann waren im Jahre 2011 Extremwerte zu verzeichnen?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 38: Kennzahlensteckbrief Bewölkungsgrad

Timestamp	TotalCloudAmount	Lat	Lon
2009-11-23 00:00:00			

Tabelle 39: DWD_Weather_Cloudiness

	Luftdruck pro Stunde	
Bedeutung	Beschreibung	Stündliche Luftdruckmessungen in bar.
	Dimensionen	Ort, Zeit, Luftdruck
	Exemplarische Analysefragen	Wann waren im Jahre 2011 Extremwerte zu verzeichnen?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 40: Kennzahlensteckbrief Luftdruck

Timestamp	Pressure	Lat	Lon
2009-11-23 00:00:00			

Tabelle 41: DWD_Weather_Pressure

	Bodentemperatur pro Stunde	
Bedeutung	Beschreibung	Stündliche Bodentemperaturmessungen in Grad Celsius.
	Dimensionen	Ort, Zeit, Bodentiefe, Bodentemperatur
	Exemplarische Analysefragen	Wann waren im Jahre 2011 Extremwerte zu verzeichnen?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 42: Kennzahlensteckbrief Bodentemperatur

Timestamp	MeasureDepth	SoilTemp	Lat	Lon
2009-11-23 00:00:00				

Tabelle 43: DWD_Weather_SoilTemp

	Winddaten pro Stunde	
Bedeutung	Beschreibung	Stündliche Winddaten mit Geschwindigkeitsangaben in $\frac{km}{h}$ und Windrichtung.
	Dimensionen	Ort, Zeit, Geschwindigkeit, Richtung
	Exemplarische Analysefragen	Wann waren im Jahre 2011 Windstille zu verzeichnen?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 44: Kennzahlensteckbrief Winddaten

Timestamp	Speed	Direction	Lat	Lon
2009-11-23 00:00:00				

Tabelle 45: DWD_Weather_Wind

8.4 Astro-Daten

Hierzu gehören Daten zu Sonnenaufgang/-untergang, Mondphasen, Arbeitstagen und Feiertagen. Die Beschreibung erfolgt über die Tabellen 46, 47, 48, 49, 50, 51, 52, 53 und 54.

	Sonnenaufgang und -Untergang	
Bedeutung	Beschreibung	Tägliche Zeiten für den Sonnenaufgang und -untergang.
	Dimensionen	Ort (nur DE), Zeit
	Exemplarische Analysefragen	Welchen Einfluss hatten Sonenaufgang/-untergang auf die Stromproduktion im Jahre 2009?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 46: Kennzahlensteckbrief Sonnenaufgang/-untergang

	Mondphasen	
Bedeutung	Beschreibung	Tägliche Zeiten für die Mondphasen.
	Dimensionen	Ort (nur DE), Zeit, Mondphasen
	Exemplarische Analysefragen	Welchen Einfluss hatten die Mondphasen auf die Stromproduktion im Jahre 2009?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 47: Kennzahlensteckbrief Mondphasen

	Arbeitszeitfaktor	
Bedeutung	Beschreibung	Arbeitstage in Deutschland bezogen auf Datum.
	Dimensionen	Ort (nur DE), Zeit, Arbeitszeitfaktor
	Exemplarische Analysefragen	Welchen Einfluss hatte die Arbeitszeit auf den Stromverbrauch im Jahre 2010?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 48: Kennzahlensteckbrief Arbeitzeitfaktor

	Feiertage	
Bedeutung	Beschreibung	Feiertage in Deutschland bezogen auf Datum.
	Dimensionen	Ort (nur DE), Zeit, Feiertage
	Exemplarische Analysefragen	Welchen Einfluss hatten die Feiertage auf den Stromverbrauch im Jahre 2010?
Datenquelle	Herkunft	Deutscher Wetterdienst (DWD)
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 49: Kennzahlensteckbrief Feiertage

Datum	WoTag	Woche	lfdTag	JulTag	DiffUTC

Tabelle 50: DWD_Astro Referenzspalten Zeit

Datum	...	SAastro	SUastro	zusätzliche SA/SU-Spalten

Tabelle 51: DWD_Astro Referenzspalten Sonnenaufgang/-untergang

Datum	...	MondTag	MondProzent	Mondphase

Tabelle 52: DWD_Astro Referenzspalten Mondphasen

Datum	WoTag	...	ArbZeitFaktor

Tabelle 53: DWD_Astro Referenzspalten Arbeitszeitfaktor

Datum	...	Feiertage

Tabelle 54: DWD_Astro Referenzspalten Feiertage

8.5 Stompreisdaten für Haushalt und Industrie

Diese Strompreisdaten stammen von Eurostat und enthalten Strompreisdaten für Haushalt und Industrie auf Jahresbasis. Der Beschreibung der Daten erfolgt anhand der Tabelle 55, 56 und 57.

	Strompreisdaten pro Jahr	
Bedeutung	Beschreibung	Jährliche Strompreisdaten für Haushalt/Industrie in Euro per Kilowattstunde.
	Dimensionen	Zeit, Strompreis
	Exemplarische Analysefragen	Wie stark veränderte sich der Strompreis seit 2009?
Datenquelle	Herkunft	Eurostat
Anwendung	Analyse	Relevant
	Prognose	Relevant
	Evaluation	Relevant

Tabelle 55: Kennzahlensteckbrief Stompreise

Year	Strompreisdaten
2009	0.123

Tabelle 56: Strom-Preis_Deutschland_Haushalt

Year	Strompreisdaten
2009	0.246

Tabelle 57: Strom-Preis_Deutschland_Industrie

Die Beschreibung der Datenstrukturen ist hiermit abgeschlossen. Im folgende Kapitel wird der entwickelte Softwareentwurf näher beschrieben.

9 Architektur-Entwurf

In diesem Kapitel erfolgt der Entwurf einer Software-Architektur, die die Grundlage des zu entwickelnden Softwaresystems bildet. Im ersten Unterkapitel werden grundsätzliche Begriffe wie Architektur und Architekturmuster definiert. Im folgenden Kapitel erfolgt die Typisierung des Softwaresystems. Zuletzt wird die Bausteinsicht, die Verteilungssicht sowie die Laufzeitsicht der Architektur näher beschrieben. Es besteht mit steigendem Detaillierungsgrad ein fließender Übergang in das Software-Design.

9.1 Definitionen

Die Architektur einer Software unterteilt sich in Komponenten und definiert deren Beziehungen untereinander und zur Umgebung. Eine Architektur beschreibt nicht den Software-Entwurf bis ins kleinste Detail. Aufgabe einer Architektur ist es, den Zusammenhang zwischen den Anforderungen an ein Softwaresystem und dessen Entwurfsentscheidungen zu beschreiben. Ein wesentliches Ziel einer Softwarearchitektur besteht darin, das Projektrisiko mit fortschreitendem Verlauf zu minimieren. Dies wird ermöglicht, indem Risiken frühzeitig erkannt und Lösungen dafür beschrieben werden. Architekturen fördern Arbeitsteilung und stellen eine wichtige Kommunikationsbasis dar. Außerdem werden sie dokumentiert (zum Beispiel über die Unified Modeling Language (UML). Sie speichern somit gewonnenes Wissen [RH06, Vgl. S. 1 f.].

Der Entwurf einer Software-Architektur verbessert die langfristige Wartbarkeit und Verständlichkeit des zu entwickelnden Softwaresystems. So wird auch nach langer Zeit der Überblick über die Strukturen und Zusammenhänge innerhalb des Systems bewahrt [SH11, Vgl. S. 2ff.].

Eine Architektur betrachtet ein Softwaresystem typischerweise aus drei verschiedenen Perspektiven. Die Bausteinsicht beschreibt die statische Struktur. Das System wird in Implementierungsbestandteile zerlegt. Die Laufzeitsicht illustriert die Dynamik des Systems, indem die Zusammenarbeit von Bausteinen zur Laufzeit beschrieben werden. Die Verteilungssicht zeigt die möglichen Ausführungsumgebungen des Systems [SH11, Vgl. S. 50f.]. Folgend wird zuerst eine Typisierung der Architektur vorgenommen, die Bausteinsicht beschrieben, anschließend die Verteilung näher beschrieben und zuletzt ein Blick auf die Laufzeitsicht geworfen.

Architekturmuster helfen bei der Zerlegung des Systems in Komponenten (Abstraktionen von Quellcode [Sta11, Vgl. S. 85]) und bei der Verteilung der Verantwortlichkeiten. Anders als Entwurfsmuster (engl. Design Patterns) stellen Architekturmuster die Systemstruktur als Ganzes dar. Es wird dabei nicht auf die Struktur einzelner Komponenten eingegangen [Sta11, Vgl. S. 95].

9.2 Typisierung des Softwaresystems

Bei der zugrundeliegenden Vision sollen Features (zum Beispiel die Außentemperatur) unterschiedlicher Datenquellen hinsichtlich ihres Einflusses auf eine zu prognostizierende Größe (Energieverbrauch) bewertet werden. Ein DWH integriert Informationen aus unterschiedlichen Datenquellen in einer für die Entscheidungsfindung optimierten Datenbank [Far06, Vgl. S. 5].

Folglich handelt es sich vordergründig um die Erstellung einer Data Warehouse Architektur. Die klassischen Aktivitäten beim Data Warehousing sind Datenextraktion-/ Datensammeln, Datenbereinigung (Data Cleansing), Datentransformation und Daten laden (LOAD). In einer für die Abfragen optimierten multidimensionalen Datenstruktur (für Online Analytical Processing (OLAP) optimiert) werden die Daten im Data Warehouse abgelegt. Auf dieser integrierten Datenbasis können Analyse- und Auswertungsmethoden ausgeführt werden [Far06, Vgl. S. 9]. Aufgrund der hohen Datenmengen, die in das Data Warehouse transferiert werden, eignet sich ein klassischer ETL-Prozess (Extraktion, Transformation und Laden) nur bedingt. Stattdessen findet ein sogenannter ELTA-Prozess (Extraktion, Laden, Transformation und Analysieren) statt. Dabei wird die Extraktion und das Laden von einer separaten Anwendung durchgeführt. Nur geringfügige Transformationen (bspw. Filter) finden bis dahin statt. Die eigentliche Transformation und Analyse wird in SAP HANA durchgeführt.

Bei der sogenannten multidimensionalen Modellierung sind Dimensionen Datenstrukturen, durch die verschiedene Aspekte der zu analysierenden Daten dargestellt werden. Beispiele für diese Aspekte sind der Ort, die Zeit oder ein Produkt. Dimensionen können auch hierarchisch angeordnet werden. Beispielsweise kann ein Ort sich in einem Bundesland befinden, das sich in einem Land befindet. Dimensionen sind somit hierarchisch organisierte Datenstrukturen, die sowohl eine Aggregation der Daten sowie ein Navigieren anhand von Operatoren ermöglichen. Sie stellen den qualifizierenden Anteil eines multidimensionalen Datenmodells dar [Far06, Vgl. S. 13f.]. Die in der Vision genannten Features werden als Bestandteile von Dimensionen dargestellt. Beispielsweise könnte das Feature Außentemperatur eine Eigenschaft einer Dimension *Wetter* sein.

Fakten repräsentieren den quantifizierenden Anteil des multidimensionalen Datenmodells. Sie sind Gegenstand der Analyse und Auswertung. Sogenannte Kennzahlen sind zum Teil verdichtete numerische Messgrößen, die betriebswirtschaftliche Sachverhalte wie beispielsweise Gewinn, Umsatz, Verlust darstellen [Far06, Vgl. S. 19]. Die in der Vision genannte zu prognostizierende Größe ist im multidimensionalen Datenmodell eine Kennzahl.

9.3 Bausteinsicht

Wie bereits in Kapitel 9.1 beschrieben wird mit Hilfe der Bausteinsicht die statische Struktur des Systems beschrieben. Im ersten Unterkapitel wird eine allgemeine Schichtenarchi-

tektur definiert, die als Leitfaden für den Aufbau der zu entwickelnden Programme dient. Anschließend erfolgt beispielhaft das Design eines Programms zur Überführung von Wetterdaten in SAP HANA, das sich nach der zuvor definierten Schichtenarchitektur richtet. Das letzte Unterkapitel zeigt verschiedene Data Warehouse Entwurfsmuster. Es wird außerdem eine Auswahl für die Verwendung eines angemessenen Musters getroffen.

9.3.1 Allgemeine Schichtenarchitektur

Das Schichtenarchitektur-Muster strukturiert das System in Schichten. Jede Schicht bekommt ihren eigenen Aufgabenbereich zugeordnet und stellt ihrer übergelagerten Schicht Dienste zur Verfügung [Som12, Vgl. S. 194]. Der wesentliche Vorteil einer Schichtenarchitektur ist, dass mit ihr komplexe Problemstellungen in überschaubare Größen zerteilt werden, die beherrschbar sind (teile und herrsche) [Sta11, Vgl. S. 137]. Die Dienste einer übergeordneten Schicht sollten nur in Ausnahmefällen genutzt werden. Diese Art der Nutzung des Schichtenmodells schränkt die Unabhängigkeit einer Schicht von den darunterliegenden ein. Bei einer wechselseitigen Nutzung von Komponenten sollten diese sich auf derselben Schicht befinden. Schichten können unabhängig voneinander erstellt, betrieben und ausgetauscht werden [Sta11, Vgl. S. 144].

Der wesentliche Nachteil der Schichtenbildung ist die mögliche Beeinträchtigung der Performanz des Gesamtsystems. Eine Anfrage muss unter Umständen über mehrere Schichten weitergeleitet werden, bis sie in ihrer Zielschicht angekommen ist. Dadurch entsteht eine zusätzliche Last für das Softwaresystem. Dieser Nachteil kann über eine Layer-Bridge überwunden werden. Die Layer-Bridge sorgt dafür, dass gezielt Zwischenschichten übersprungen werden können. Ein weiterer Nachteil ist die Durchführung von schichtenübergreifenden Änderungen, wie zum Beispiel das Hinzufügen eines neuen Datenfeldes. Dabei muss die Anpassung auf mehreren Schichten erfolgen, was einen erhöhten Anpassungsaufwand bedeutet [Sta11, Vgl. S. 144f.].

Es besteht die Empfehlung alle Komponenten des Softwaresystems nach dem in Abbildung 9 gezeigten Schichtenarchitekturmuster zu bilden. Dieses gliedert das Softwaresystem in vier Schichten. Die *User Interface*-Schicht ist für die Interaktion mit dem Benutzer verantwortlich. Sie nutzt die Services der Schicht *Domain*, um die notwendigen Informationen (zum Beispiel aggregierte Analysedaten) zu erhalten. *Domain* stellt von der zugrundeliegenden Technik unabhängige Services bereit. *DataAccess* enthält die Repositories und das objektrelationale Mapping, um einen gebündelten Zugriff auf die *Infrastruktur*-Schicht zur Verfügung zu stellen. In der *Infrastructure*-Schicht werden Daten-Definitions und-Manipulations-Funktionen für Persistenz, sowie Analysefunktionalität, Prognosefunktionalität und weiter allgemeine Funktionalitäten wie beispielsweise das Sitzungsmanagement bereitgestellt. Viele der Aufgaben der *Infrastructure*-Schicht werden im Vordergrund auf der SAP HANA Plattform bereitgestellt.

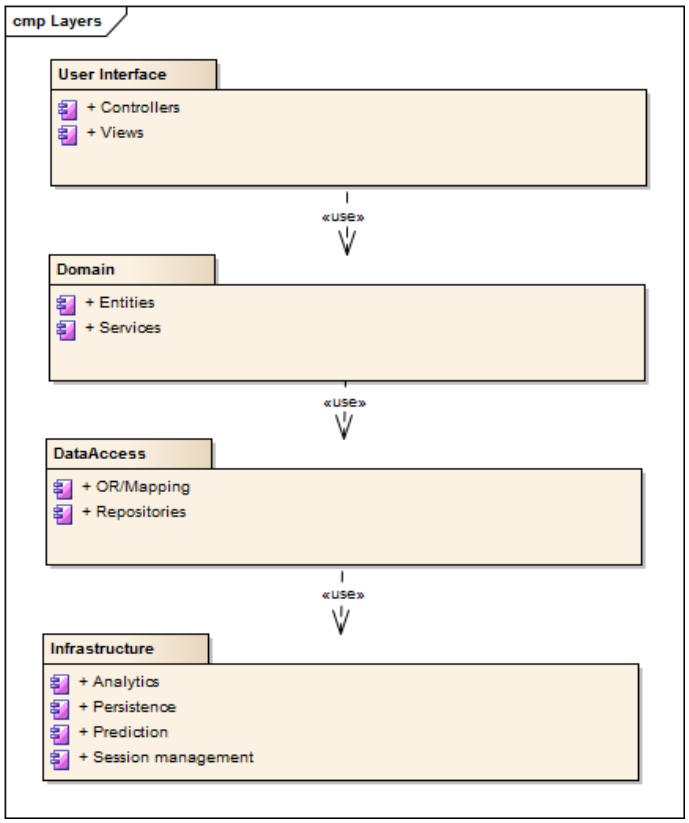


Abbildung 9: Architektur Schichten

9.3.2 EL-Prozess Design

Wie in Kapitel 9.2 genannt findet kein klassischer ETL-Prozess sondern ein ELTA-Prozess statt. In diesem Kapitel erfolgt das Design einer Anwendung zwecks Extraktion und Laden von Daten (EL) in SAP HANA. Es werden die dazu nötigen Pakete und Abhängigkeiten betrachtet. Die zuvor definierte Schichtenarchitektur wird bei der Definition der Abhängigkeiten beachtet. Abbildung 10 zeigt beispielhaft auf der rechten Seite die Paketstruktur der allgemeinen Funktionalität, die von anderen Paketen geteilt wird (shared) und auf der linken Seite die Paketstruktur zu domänenspezifischen Inhalten (in dem Fall das Wetter).

Die Pakete mit den Endungen *.ui* sind der Schicht *User interface* zugeordnet. Pakete mit den Endungen *.logic* sind in der Schicht *Domain* angesiedelt. Die Pakete mit den Endungen *.dataaccess* und *.model* liegen in der zuvor definierten Schicht *DataAccess*. Eine Besonderheit stellen Pakete mit der Endung *.util* dar. Es handelt sich dabei um Pakete, die Funktionalität bereitstellen, die tendenziell auf verschiedenen Schichten verwendet wird. Diese sogenannten Utilities sind keiner spezifischen Schicht zuzuordnen.

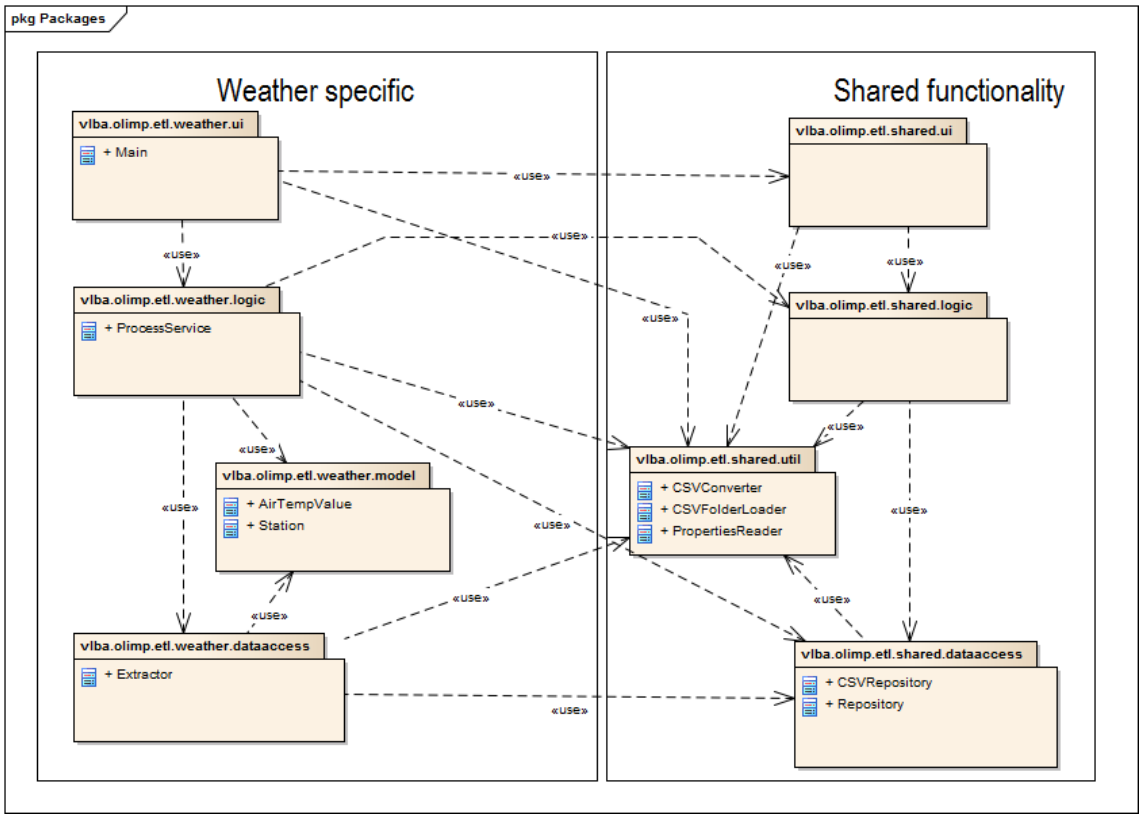


Abbildung 10: Architektur Pakete EL-Programm

9.3.3 Data Warehouse Design

Im folgenden wird auf Design-Spezialitäten des Data Warehouse Systems eingegangen. Bei der Wahl des Schemas für das Data Warehouse existieren diverse mögliche Varianten. Bekannte Vertreter sind das Star-Schema, Snowflake-Schema, Galaxie-Schema oder Mischformen[Far06, Vgl. S. 29ff.]. Ein Galaxie-Schema wird vorab ausgeschlossen, da nur eine Faktentabelle notwendig ist. Die Wahl zwischen dem Star-Schema und Snowflake-Schema hängt vom Anwendungsthema ab. Tendenziell ist die Beantwortung von Anfragen bei einem Star-Schema effizienter, da keine JOIN-Operatoren benötigt werden. Außerdem weist ein Star-Schema eine weniger komplexe Struktur auf, was die Erstellung und Wartbarkeit vereinfacht. Die durch ein Star-Schema entstehende Redundanz verursacht zwar einen Mehrbedarf an Speicherkapazitäten, jedoch wird dieser als unkritisch betrachtet. Der Grund dafür ist, dass typischerweise die Dimensionstabellen im Vergleich zur Faktentabelle relativ klein sind [Far06, Vgl. S. 31.f].

Abbildung 11 zeigt das Data Warehouse in Form eines Snowflake-Schemas [Far06, Vgl. S. 30f.] Abbildung 12 zeigt das Data Warehouse in Form eines Star-Schemas [Far06, Vgl. S. 30f.]. Wenn die Entscheidung des Designs auf die Umsetzung eines klassischen Data Warehouse fällt, ist ein Star Schema aufgrund der genannten Vorteile zu wählen. Die

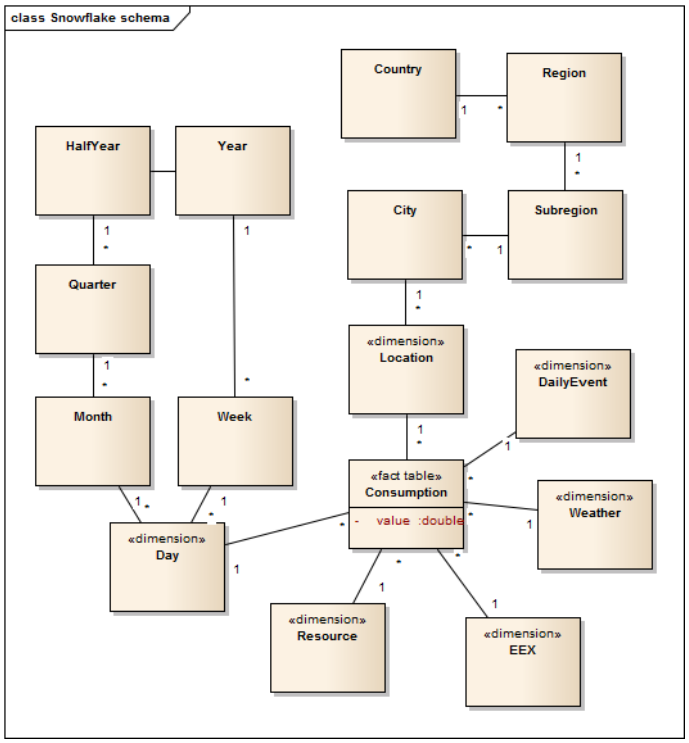


Abbildung 11: Snowflake-Schema

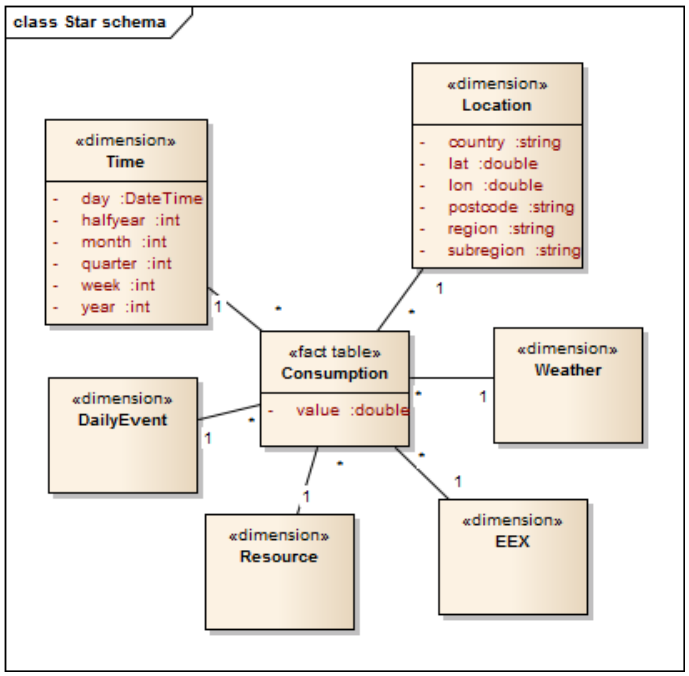


Abbildung 12: Star-Schema

zu testenden Hypothesen stellen klare Anforderungen an die zugrundeliegenden Datenstrukturen. Deshalb muss nicht eine Flexibilität bereitgestellt werden, wie sie in einem

klassischen Data Warehouse Design realisiert wird. Es wird die Entscheidung getroffen, die Analyse-Datenstrukturen in Form von Views und einfachen Tabellen umzusetzen, ohne ein klassisches Data Warehouse Design zu verwenden. Die wesentlichen Gründe dafür sind, dass der Implementationsaufwand als geringer und die Flexibilität als adäquat geschätzt wird.

9.4 Verteilungssicht

Abbildung 13 zeigt die Verteilung des Systems. Zentrum dessen ist das Device *OL*. Auf diesem Server wird eine Java-Anwendung als Komponente des Systems ausgeführt, die die Schritte Extraktion und Laden des in Kapitel 9.2 definierten ELTA-Prozesses für die Datenquellen durchführt. Ein Beispiel für eine mögliche Datenquelle ist der Deutsche Wetterdienst. Auf die Datenquellen wird über einen Pull-Mechanismus zugegriffen. Je nachdem, ob ein kontinuierlicher Fluss oder ein einmaliges Laden der Datenquelle notwendig ist, wird der Pull-Mechanismus pro Datenquelle in Zeitintervallen oder einmalig angestoßen. Nach der Extraktion erfolgen auf dem Device *OL* geringfügige Transformationen (z.B. Filter). Die Daten werden jetzt in das Device *HPI* über einen Push-Mechanismus geladen. Dort findet die eigentliche Transformation innerhalb der SAP HANA Datenbank zwecks Analyse statt. Eine mögliche UI-Kompontente sorgt für die Visualisierung der notwendigen Auswertungen.

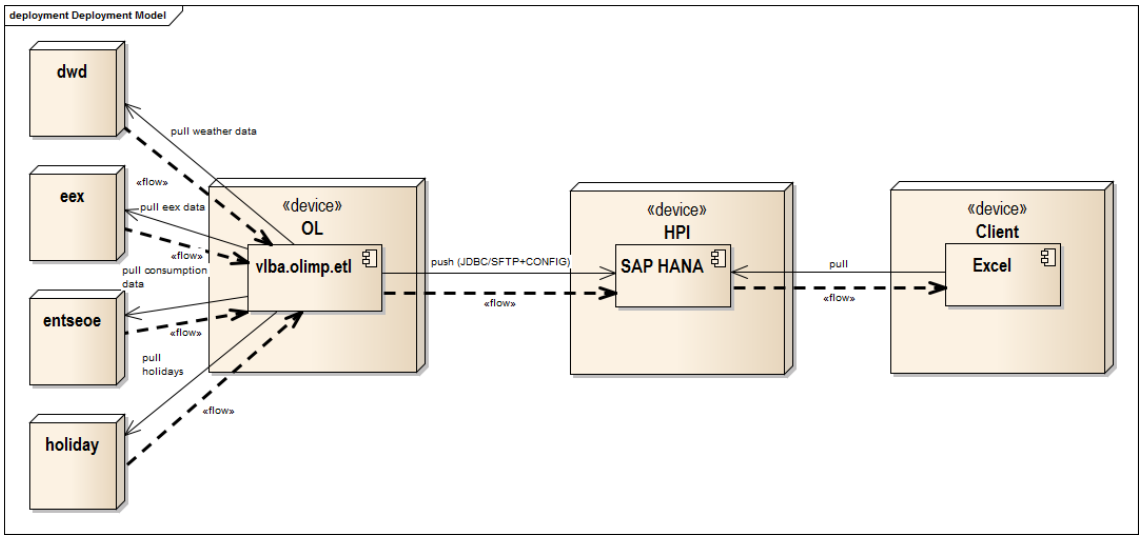


Abbildung 13: Architektur Verteilungssicht

9.5 Laufzeitsicht

Abbildung 14 zeigt beispielhaft die Interaktion verschiedener Akteure (Instanzen von Klassen) zum Extrahieren und Laden von Daten aus einer beliebigen Datenquelle. Der Vorgang

ist generisch beschrieben, damit er für verschiedene Datenquellen wiederverwendet werden kann. Er beginnt damit, dass eine Instanz der Klasse *Main* (*User Interface* Schicht) eine neue Instanz der Klasse *ProcessService* erstellt und die Methode *process* aufruft. *ProcessService* ist in der Schicht *Domain* angesiedelt und koordiniert den Vorgang. Jetzt erzeugt *ProcessService* eine Instanz der Klasse *Repository* aus der *DataAccess* Schicht. Mit *begin* wird ein neuer Übertragungsvorgang vorbereitet. Danach erzeugt der *ProcessService* eine neue Instanz eines *Extractors*. Der *Extractor* ist in der *DataAccess* Schicht angesiedelt. Er ist für das Extrahieren der Daten aus der Datenquelle verantwortlich und stellt ein einheitliches Interface bereit. Jetzt führt der *ProcessService* eine Schleife aus, die bis zum EoF (End of File) des Extraktors iteriert. In ihr wird über *getNextValue* der Nächste zu verarbeitende Wert gelesen und mit *newRecord* in das *Repository* geschrieben. Am Ende der Schleife führt der *ProcessService* ein *commit* beim *Repository* durch. Es erfolgt die gebündelte Übertragung der Daten an SAP HANA. Nach Abschluss gibt der *ProcessService* an *Main* eine Rückmeldung in der Form eines Reports. Der Report enthält Angaben über den Erfolg der Extraktion und des Ladens.

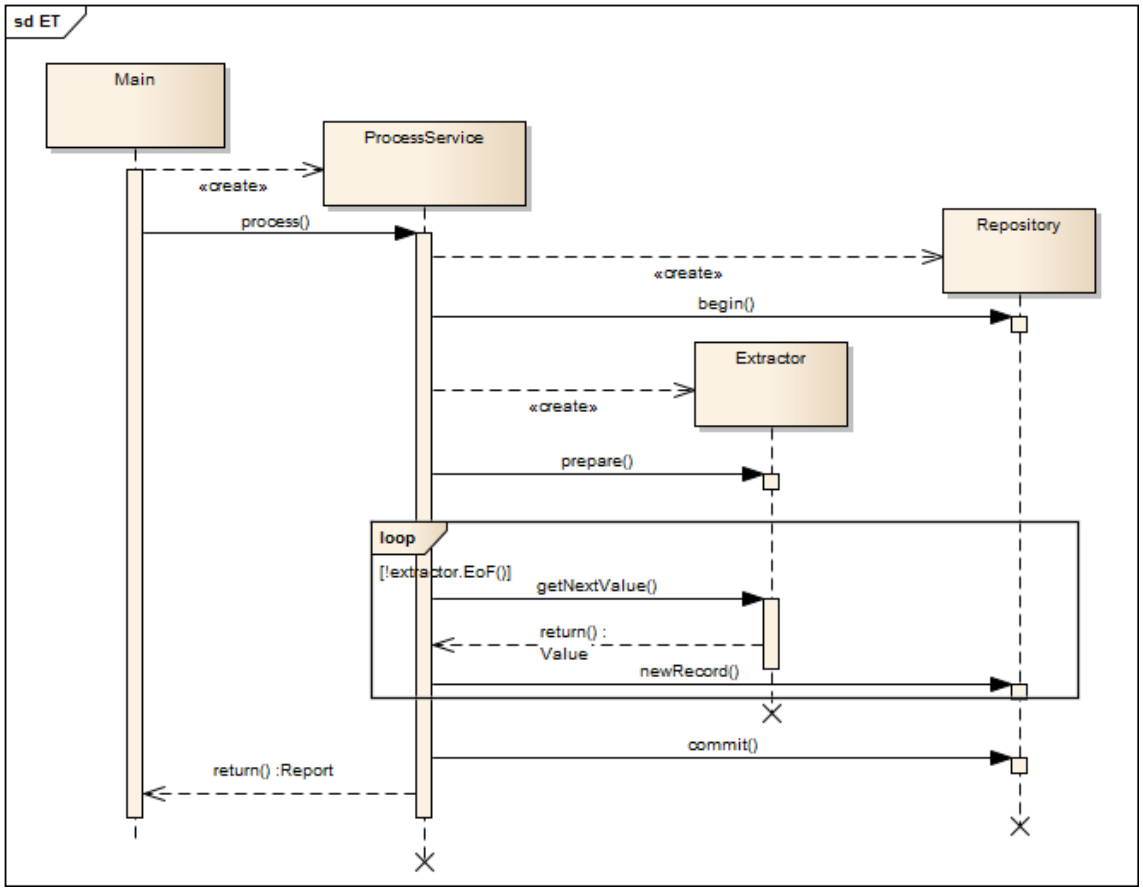


Abbildung 14: Sequenzdiagramm für einen beispielhaften Extraktions- und Ladeprozess

10 Implementierung

In diesem Kapitel werden die entwickelten Softwareartefakte und die zur Lösung der Aufgabenstellung beitragen, näher beschrieben. Hierzu werden in den beiden folgenden Abschnitten zunächst Code-Conventions für die Entwicklung der Artefakte in Java und SAP HANA dargestellt. Anschließend wird auf die erstellte Java Dokumentation verwiesen. Zuletzt erfolgt die Beschreibung des ELTA-Prozesses, der den Gesamtablauf vom Extrahieren, Laden, Transformieren und Analysieren der Daten darstellt.

10.1 Code Conventions

Code Conventions stellen Konventionen über Benennungen, zu verwendende Strukturen und Muster dar [Ber08]. Diese tragen zur Verbesserung der Lesbarkeit und Wartbarkeit der zu entwickelnden Quellcode-Artefakte bei. Zusätzlich können Code Conventions dazu beitragen, die Fehleranfälligkeit der Programmcodes zu verringern.

Die Entwicklung findet einerseits in Java und andererseits in SAP HANA statt. Um auf die Besonderheiten und Quasi-Standards der beiden Paradigmen einzugehen, werden zwei Code Conventions definiert, die je nach Entwicklungsumgebung einzuhalten sind.

10.1.1 Java

Zuerst werden die Konventionen für die Softwareentwicklung in Java definiert(siehe Tabelle 58). Grundlage dessen ist das Dokument *Code Conventions for the Java TM Programming Language* der Firma Oracle [Mic99] aus dem Jahr 1999.

10.1.2 SAP HANA

Für die Entwicklung in SAP HANA sind die Code Conventions innerhalb des *SAP HANA Developer Guide* [SAP14a] einzuhalten(siehe Tabelle 59).

Der Datenfluss innerhalb der SAP HANA Instanz ist wie folgt definiert: Nicht-transformierte Daten, die noch nicht für die Analyse verwendet werden können, sind im Schema *OLIMP* abzulegen. Damit sind jene Daten gemeint, die durch Extraktion und Laden (siehe auch Kapitel 9.3.2) in das System überführt wurden. Transformierte Daten sind im Schema *PRE* abzulegen. Sie bilden die Grundlage für die Analyse. Die Modellbildung und Ausführung der Algorithmen erfolgt im Schema *PAL*. Insbesondere werden im Schema *PAL* die Views hinterlegt, welche die Trainingsdaten für die Algorithmen darstellen. Die von den Algorithmen erstellten Tabellen und Prozeduren für die Prediktionen werden ebenfalls im Schema *PAL* hinterlegt.

Konvention	**Beschreibung**
Englisch als Entwicklungssprache	Der Quelltext ist in der Sprache Englisch zu verfassen.
Deutsch als Kommentarsprache	Kommentare und die *javadoc* werden zur besseren Verständlichkeit in Deutsch verfasst.
CamelCase-Notation	Für Klassen-, Variablen- und Funktionsnamen wird die Camel-Case-Notation verwendet, das heißt jedes Folgewort wird mit einem Großbuchstaben begonnen. Beispiel: getAbsoltuePathOfCsvData().
Namespace	Alle zu entwickelnden Pakete befinden sich im Namespace *vlba.olimp*. *vlba* steht für die Abteilung der Universität Oldenburg, *olimp* für die Angabe der Projektgruppe.
Repository	Der erstellte Programmcode ist zeitnah in dem dafür vorgesehenen SVN-Repository zu hinterlegen.
Zeichensatz	Der zu verwendende Zeichensatz für Quelltext-Artefakte ist UTF-8 festgelegt.
Java-Version	Es wird die Laufzeitversion 1.7 der *Java Runtime Environment* (JRE) für die Entwicklung festgelegt.
Entwurfsmuster	Wenn Verständlichkeit und/oder Wiederverwendbarkeit davon profitieren, sind Entwurfsmuster im Design einzusetzen [GHJV15].

Tabelle 58: Entwicklungsrichtlinien für Java

Konvention	**Beschreibung**
Englisch als Entwicklungssprache	Der Quelltext ist in Englisch zu verfassen.
Deutsch als Kommentarsprache	Kommentare und Dokumentation sind in Deutsch zu verfassen.
Zeichensatz	Für Artefakte ist der Zeichensatz UTF-8 festgelegt.
Namespace	Tabellen, die Daten aus einer externen Datenquelle beinhalten sind im Präfix mit einem Kürzel der Herkunft zu bezeichnen. Ein Beispiel: *DWD_WEATHER*. Der Präfix *DWD* deutet dabei auf die Herkunft vom Deutschen Wetterdienst.
Repository	Tabellen- und Schemadefinitionen sind im SVN-Repository im Format *.hdbtable* und *.hdbschema* abzuspeichern.

Tabelle 59: Entwicklungsrichtlinien für SAP HANA

10.2 Javadoc des Programmcodes

Im Verlaufe der Programmierung wurde zu jeder Klasse und jeder Methode JavaDoc geschrieben. Das generierte JavaDoc hierzu ist auf der beigelegten CD unter *Quellcode/olimp/doc/index.html* zu finden.

10.3 ELTA-Prozess

In diesem Abschnitt wird der in Kapitel 9.5 entworfene abstrakte Extraktions- und Ladeprozess konkret beschrieben. Mit Hilfe dieses Prozesses werden die Daten von aus den verschiedenen Datenquellen extrahiert und verarbeitet, anschließend in das SAP HANA System mit Hilfe von Secure File Transfer Protocol (SFTP) geladen und dort transformiert. Zuletzt erfolgt eine Modellbildung und Prognose (Analyse).

Rohdaten und Vorverarbeitung Rohdaten sind die direkten Daten der Datenlieferanten, zum Beispiel die Stromverbrauchsdaten der ENTSO-E. Diese Daten liegen in der Regel in einem Format vor, welches mit SAP HANA nicht kompatibel ist. Solche Dateiformate können zum Beispiel das Excel-Datei (XLS), CSV oder XML Dateiformat sein[5]. Oftmals liegen diese Rohdaten auch nicht in einer einzelnen Datei vor, sondern in vielen einzelnen Dateien, welches den Datenimport zusätzlich erschwert. Die ENTSO-E Stromverbrauchsdaten liegen beispielsweise für jeden Monat der Jahre 2009 - 2014 in einzelnen Dateien vor. Aus diesem Grund werden die Daten von verschiedenen Java-Programmen vorverarbeitet, um einen möglichst einfachen und fehlerfreien Datenimport in SAP HANA zu gewährleisten. Grundsätzlich wird die Vorverarbeitung dazu verwendet, aus vielen einzelnen Dateien eine einzelne CSV Datei zu erstellen, die für den Datenimport in SAP HANA geeignet ist. Dazu wird folgendermaßen vorgegangen: Die Java-Klasse `CSVFolderLoader` lädt alle Dateien eines spezifizierten Datenformates in eine interne Datenstruktur. Die Java-Klasse `CSVConverter` schreibt die Datensätze aus den verschiedenen Dateien in eine einzelne CSV Datei, die anschließend als Import-Datei für SAP HANA verwendet wird[6]. Um dieses Zusammenfügen von Datensätzen aus verschiedenen Datenquellen zu dokumentieren, wird die Java-Klasse `Report` verwendet, welche das Zusammenfügen der einzelnen Datensätze dokumentiert. Die Dokumentation des Zusammenfügens veranschaulicht dabei der folgende Pseudocode:

```
1 for each single_file
2         for each dataset in single_file
3                         increase read_lines;
4                         if  line_is_skipped then
5                                 increase skipped_lines;
6                                 add_reason;
7                         else
8                                 write line_in_csv_file;
9                                 increase written_lines;
```

Abbildung 15: Pseudocode des Reports auf Rohdaten- und Vorverarbeitungsebene

Dabei wird wie folgt vorgegangen: Zunächst wird über jeden Datensatz in jeder Rohdatendatei iteriert (Zeile 1 und 2). Anschließend wird der Zähler für die gelesenen Zeilen (Zeile

[5] Prinzipiell sind hier alle Datenformate möglich
[6] Für einzelne Datenquellen kann das Zusammenfügen der einzelnen Rohdaten differieren

3) um den Wert 1 erhöht. Wenn dieser Datensatz ignoriert werden sollte (weil er beispielsweise leer ist), wird der Zähler für die ausgelassenen Zeilen um den Wert 1 erhöht (Zeile 5). In diesem Fall kann zusätzlich ein Grund in Form eines Strings für das Auslassen eines Datensatzes hinzugefügt werden (Zeile 6). Wenn der Datensatz in die Ziel-CSV-Datei geschrieben wird (Zeile 8) wird der Zähler für die geschriebenen Zeilen um den Wert 1 erhöht (Zeile 9). Durch diese Vorgehensweise kann nachvollzogen werden, ob die Anzahl zwischen gelesenen und geschriebenen Zeilen übereinstimmt. Zusätzlich lässt sich mit dieser Vorgehensweise überprüfen, aus welchen Gründen die Anzahl der gelesenen und geschriebenen Zielen differiert.

Datenimport in SAP HANA Die im ersten Schritt erstellte CSV Datei wird im Anschluss per SFTP automatisiert auf dem Server des HPI in Potsdam hochgeladen. Hierzu ist neben der CSV Datei selbst auch eine Control File (CTL) Datei notwendig. Diese Datei enthält die notwendigen Informationen für den Datenimport in die SAP HANA Instanz. Die CTL Datei wird im Laufe des automatisierten Befüllens der SAP HANA Datenbank ebenfalls erstellt und per SFTP auf dem Server des HPI hochgeladen. Der folgende Pseudocode zeigt den prinzipiellen Aufbau einer CTL Datei:

```
1 IMPORT DATA
2 INTO TABLE "<SCHEMANAME>" . "<TABLENAME>"
3 FROM '<PATH_TO_CSV_FILE_ON_SERVER>'
4 FIELDS DELIMITED BY '<FIELDDELMITER>'
5 ERROR LOG '<PATH_TO_ERROR_LOG_ON_SERVER>'
```

Abbildung 16: Pseudocode der CTL-Datei

Zeile 2 definiert das Schema und die Tabelle in der die Daten der CSV-Datei, die in Ziele 3 spezifiziert wird, geladen werden. Zusätzlich muss angegeben werden, mit welchem Zeichen die einzelnen Felder getrennt werden (Zeile 4). Zeile 5 besagt, das eine Fehlerdatei für den Datenimport geführt wird. In dieser Fehlerdatei werden alle (möglichen) Fehler des Datenimportes geschrieben. Diese (minimale) Konfiguration wurde für jeden Datenimport verwendet[7]. Das bedeutet, für jeden Datenimport liegt eine Fehlerdatei vor. Die Qualität des Datenimportes ist somit nachvollziehbar.

Sobald sich beide Dateien auf dem Server befinden kann der Datenimport per SQL-Befehl gestartet werden. Der Code hierzu lautet:

Dieser Befehl (siehe Listing 17) wird per JDBC auf dem lokalen Server ausgeführt uns veranlasst die SAP HANA Instanz den Datenimport mit den in der CTL Datei (Zeile 1) hinterlegten Synataxparametern zu starten. Optional kann in Zeile 2 angegeben werden, mit wie vielen Threads (Anzahl der parallel laufenden Prozesse) und Batches (Anzahl der

[7] Für vereinzelte Datenimporte wurden ggf. zusätzliche Parameter verwendet. Die vollständige Syntax kann unter `http://help.sap.com/saphelp_hanaplatform/helpdata/en/20/f712e175191014907393741fadcb97/content.htm` eingesehen werden

```
1 IMPORT FROM '<PATH_TO_CTL_FILE_ON_SERVER>'
2 WITH THREADS 64 BATCH 200000;
```

Abbildung 17: Pseudocode des Import-Befehls

geschriebenen Datensätze je commit) der Datenimport durchgeführt werden soll. Da in der Minimalkonfiguration definiert ist, dass in jedem Fall alle Fehler mitgeschrieben werden sollen, erstellt das System zu jeden Datenimport eine Fehlerdatei, in welcher alle Fehler, die während des Datenimportes aufgetreten sind, protokolliert werden. Das folgende Beispiel zeigt den möglichen Inhalt einer Fehlerdatei:

```
1 Parsing error: incorrect delimiter for the next column of
      ExpectedSolarEnergy field : 01.
2 ESPL;10YDE-ENBW-----N;2011-03-27T06:00:00+02:00;01. Feb;2011-03-26T18
      :00:13+01:00;2011-03-26T16:46:07+01:00
```

Abbildung 18: Pseudocode der Fehlerdatei

Der Pseudocode in Listing 18 zeigt ein Beispiel für einen fehlerhaften Datenimport. Die Fehlerdatei zeigt hierzu die betreffende Zeile der Quelldatei und den Fehlergrund. Im Beispiel ist der Grund ein fehlerhaftes Trennzeichen in der Quelldatei. Hierzu ist dann ein manuelles korrigieren der fehlerhaften Datensätze denkbar.

Modellbildung Nachdem die Daten in SAP HANA in der richtigen Form und Menge für die Simulation zur Verfügung stehen, kommt die nächste Herausforderung: die Modellbildung. Ein Modell besteht aus 3 Komponenten, das als eine Art Formel verstanden werden kann:

Algorithmus + Datenmenge + Konfiguration = Modell

Es folgt eine Beschreibung der einzelnen Komponenten des Modells:

1. Algorithmus:
   In SAP HANA Studio ist die PAL integriert. PAL definiert Funktionen (*Algorithmen*), die entweder mit Hilfe von SQLScript-Prozeduren oder manuell über das eingebaute analytische Tool AFM aufgerufen werden können. Mit den Algorithmen werden Data-Mining-Funktionen auf die Daten angewendet, um darin verborgene Trends aufzudecken. Die im Projekt verwendete Version von PAL beinhaltet klassische sowie universelle PA-Algorithmen, die in 9 Data-Mining-Kategorien aufgeteilt sind:
   - Clustering
   - Classification

- Regression
- Association
- Time Series
- Preprocessing
- Statistics
- Social Network Analysis
- Miscellaneous

Die wichtigste Aufgabe des Analysten ist dabei, einen für den Anwendungsfall geeigneten Algorithmus zu finden. In der Projektvorbereitung wurden hierzu zahlreiche Recherchen durchgeführt, die in Abschnitt 11 zu finden sind.

2. Datenmenge:
   Unter der Datenmenge werden die je nach Hypothese ausgewählten Datensätze für die Prognose verstanden. Eine detaillierte Beschreibung hierzu ist im Abschnitt 10.3 zu finden.

3. Konfiguration:
   Die *Konfiguration* umfasst die möglichen Einstellungen der Parametern eines bestimmten Algorithmus.[8] Die Predictive Analysis Library erlaubt hierzu die Justierung der entsprechenden Parameter für jeden Algorithmus. Die Justierung der Parameter des Algorithmus kann erforderlich werden, um eine höhere Prognosequalität zu erzielen. Äußerst wichtig ist hierzu anzumerken, dass jede Änderung einer der drei oben genannten Komponenten zur Bildung eines neuen Modells führt.

Prognose Wenn das Modell konfiguriert ist und das System keine Fehlermeldungen anzeigt, wird die Prognose entweder manuell über den AFM oder durch die SQL Prozedur ausgeführt.

[8] Die möglichen Parametereinstellungen differieren je Algorithmus

11 Auswahl der Algorithmen

In diesem Kapitel werden die von der SAP HANA PAL-Bibliothek zur Verfügung gestellten Algorithmen näher beschrieben. Es wird auf die allgemeine Beschreibung der Funktion, der mathematischen Formel sowie der PAL-Einstellungen bezüglich der Input/Output-Tabellen eingegangen. Hierzu werden zunächst im Abschnitt 11.1 die verfügbaren Algorithmen zur Zeitreihenanalyse vorgestellt. Der darauffolgende Abschnitt 11.2 befasst sich mit den Algorithmen zur Regression. Anschließend wird im Abschnitt 11.3 das Thema *Support Vector Machine* behandelt. Das Kapitel zur Auswahl der Algorithmen endet mit dem Abschnitt 11.4 Messkriterien. Mit den in diesem Unterkapitel vorgestellten Messkriterien lässt sich die Prädiktionsqualität verschiedener Algorithmen auf gleicher Datenbasis miteinander vergleichen.

11.1 Zeitreihenanalyse

In diesem Abschnitt werden die verschiedenen Algorithmen zur Zeitreihenanalyse vorgestellt. Hierzu wird jeweils der entsprechende Algorithmus sowie dessen Funktionsweise vorgestellt.

11.1.1 Linear Regression with damped Trend and seasonal Adjust

Die lineare Regression liefert eine Prognose mit einem konstanten Trend (zunehmend oder abnehmend). Um solche Trends dahingehend anzupassen, dass die Prognose in der Zukunft immer zuverlässig bleibt, sollte ein Anpassungsparameter in Betracht gezogen werden. Mit dem *Damped Parameter* bietet der Algorithmus *Linear Regression with Damped Trend and Seasonal Adjust* die Möglichkeit, den Anpassungsgrad des Trends anzugeben. Zusätzlich ist es möglich, die *Seasonality* einzugeben. *Seasonality* bedeutet dabei, dass die Daten von periodischen Einflüssen wie z.B. Quartalen oder Monaten beeinflusst werden. Diese kann der Algorithmus auch automatisch erkennen [SAP15, S. 269].

Einstellungen in PAL Der Algorithmus akzeptiert keine Null-Werte oder fehlende Daten im Input. Ansonsten sollten die Daten numerisch sein. Es sollte außerdem immer eine *Forecast length* eingegeben werden. Diese repräsentiert die Anzahl der vorherzusagenden Perioden. Der Trend ist der *Damping Factor* und dieser sollte zwischen 0 und 1 liegen, wobei 1 der Default-Wert ist. Dieser Faktor ist für die Minimierung der Fehler in der Vorhersage des Trends zuständig. Es ist auch möglich zu bestimmen, ob der Trend nur die Zukunft beeinflussen soll oder ebenfalls die Vergangenheitsdaten. Der *Seasonality Factor* hat den Default-Wert 0. Dies bedeutet, dass keine Seasonality betrachtet wird. Der Wert 1 bedeutet, dass der *Seasonality-Einfluss* existiert und dass die Anzahl der Perioden, die eine *Season* bilden, manuell im Parameter *PERIODS* eingegeben werden soll. Falls der

Wert 2 eingegeben wurde, so heißt dies, dass die *Seasonality* automatisch erkannt werden soll. In diesem Fall sollte im Parameter *SEASONAL_HANDLE_METHOD* eingegeben werden, ob der Wert für den Parameter *PERIODS* automatisch mit der Mittelwert-Methode (Default-Wert 0) oder mit der *Fitting Linear Method* (Wert 1) zu ermitteln ist. Der Parameter *MEASURE_NAME* ermöglicht es anzugeben, welche Fehlerkennzahlen ausgegeben werden sollen. Die Tabelle *Result* enthält zwei Spalten: *Name* und *Value*. Hier werden die verschiedenen Kennzahlen mit ihren Werten aufgelistet. Die Tabelle *Forecast* enthält zwei Spalten: *Timestamp* und *Value*. An dieser Stelle wird jeder Periode eine Prognose zugeordnet [SAP15, S.270].

11.1.2 ARIMA

ARIMA (*Auto Regressive Integrated Moving Average*) ist ein Algorithmus der Zeitreihengruppe und kann als (p, d, q) geschrieben werden, wobei p sich auf die *auto regressive order* bezieht, d auf die *integrated order* und q auf die *moving average order* [SAP15, S. 225].

Mathematischer Hintergrund Ein *ARIMA Model* ist eine Universalisierung des *auto regressive moving average (ARMA) models*. Der integrierte Teil ist dafür da, aus Daten, die nicht stationär sind, stationäre Beziehungen zu induzieren. Ein *ARIMA (p, d, q) Model* kann wie folgt ausgedrückt werden:

$$\Phi(B)(1-B)^d(Y_t - c) = \Theta(B)\varepsilon_t, t \in Z \tag{1}$$

Ein *ARMA (p, q) Model* kann auf diese Weise ausgedrückt werden:

$$\Phi(B)(Y_t - c) = \Theta(B)\varepsilon_t, t \in Z\varepsilon_t - i.i.d.N(0, \sigma^2) \tag{2}$$

Wobei B ein *lag operator* und c der Durchschnitt der Zeitreihen-Daten ist [Bus14, S. 172] und [SAP15, S. 226]:

$$\Phi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p, p \geq 0\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q, q \geq 0 \tag{3}$$

ARIMATRAIN *ARIMATRAIN* konvertiert in PAL zuerst die originalen nicht stationären Daten zu stationären und fügt die stationären Zeitreihen-Daten in ein *ARMA*

Model ein. Die Input-Tabelle enthält zwei Spalten: *Timestamp* und die *Raw Daten*, die vorherzusagen sind. Die Vorhersage erfolgt anhand von Vergangenheitsdaten. Die Werte für p,q und d können in PAL bestimmt werden und es kann zudem festgelegt werden, ob das Ergebnis stationär sein soll oder nicht.

Die Output-Tabelle enthält die Namen für die verschiedenen Parameter, die aus dem *ARIMATRAIN* resultieren, wie z.B. *ARParameter* und *MAParameter* und einen dazugehörigen Wert [SAP15, S.229].

ARIMAFORECAST Zunächst kann *ARIMA* zur Vorhersage zukünftiger Perioden beitragen. Die Input-Tabelle enthält dabei die zwei Spalten (*Name* und *Wert*). Die Anzahl der vorherzusagenden Perioden kann als Parameter eingegeben werden. Die Output-Tabelle enthält sechs Spalten; eine davon repräsentiert den Timestamp und die anderen die vorhergesagten Werte [SAP15, S.232].

11.1.3 Forecast Smoothing

Forecast Smoothing kann benutzt werden, um optimale Parameter aus einer Reihe von Glättungsfunktionen in PAL zu berechnen, einschließlich Single Exponential Smoothing, Double Exponential Smoothing und Triple Exponential Smoothing. Die Funktion gibt außerdem die Prognoseergebnisse auf der Grundlage dieser optimalen Parameter aus. Diese Optimierung wird durch die Exploration des Parameterraums, der alle möglichen Parameterkombinationen enthält, berechnet. Die Qualitätsbewertung findet mittels Vergleich von historischen und prognostizierten Werten statt. In PAL wird der *MSE* (*Mean Squared Error*) verwendet, um die Qualität der Parameter zu bewerten [SAP15, S. 254].

Mathematischer Hintergrund Die Parameteroptimierung basiert auf globalen und lokalen Suchalgorithmen. Der globale Suchalgorithmus, der in dieser Funktion verwendet wird, ist *Simulated Annealing*. Der lokale Suchalgorithmus hingegen ist *Nelder Mead*. Diese Algorithmen ermöglichen einen effizienten Suchprozess [SAP15, S. 254].

Einstellungen in PAL *Forecast Smoothing* akzeptiert keine Null-Werte oder fehlende Daten im Input. Außerdem sollten die Daten numerisch sein.

Die Input-Tabelle enthält eine Spalte mit dem Zeitstempel-Index und eine weitere mit den Rohdaten.

Die Funktion berechnet die optimalen Parameter und Output-Prognoseergebnisse. Mit den verschiedenen in PAL verfügbaren Einstellungen kann u.a. das gewünschte statistische Modell für die Prognoseberechnung gewählt werden (*Single*, *Double* oder *Triple Exponential Smoothing*) und der Glättungsfaktor.

Die Output-Tabelle enthält den Zeitstempel-Index, das Ergebnis der Glättungsprognose und die Fehlerquote des jeweiligen prognostizierten Wertes [SAP15, S. 255- 258].

11.1.4 Single Exponential Smoothing

Single Exponential Smoothing eignet sich für die Modellierung von Zeitreihen ohne Betrachtung von Trends und *Seasonality*. Die angepasste Vorhersage ist die gewichtete Summe des letzten angepassten Werts und des letzten tatsächlichen Wertes [SAP15, S. 275].

Mathematischer Hintergrund Sei St der angepasste Wert für die t-Periode. Mathematisch:

$$S_1 = x_0 \tag{4}$$

$$S_t = \alpha x_{t-1} + (1-a)S_{t-1} \tag{5}$$

Wobei $\alpha \in (0,1)$ der Fehler-Anpassungsfaktor ist. Der Forecast erfolgt dann wie folgt:

$$S_{T+1} = \alpha x_T + (1-a)S_T \tag{6}$$

Einstellungen in PAL Der Algorithmus akzeptiert keine Null-Werte oder fehlende Daten im Input. Ansonsten sollten die Daten numerisch sein.

Die Input-Tabelle enthält zwei Spalten: *ID* und die *Raw Daten*, die vorherzusagen sind. Die Vorhersage erfolgt anhand von Vergangenheitsdaten. Mit *Alpha* als Anpassungsfaktor, dessen Wert zwischen 0 und 1 liegen sollte, versucht der Algorithmus die Fehlerquote zu minimieren, um genauere Vorhersagen zu treffen.

Durch die verschiedenen Einstellungen in PAL kann die Anzahl der vorherzusagenden Perioden eingegeben werden. Die Output-Tabelle enthält die ID entsprechend der vorherzusagenden Perioden und die vorhergesagten Werte

Der Parameter *MEASURE_NAME* ermöglicht es anzugeben, welche Fehlerkennzahlen ausgegeben werden sollen [SAP15, S. 277 - 280].

11.1.5 Double Exponential Smoothing

Double Exponential Smoothing eignet sich für die Modellierung von Zeitreihen mit Betrachtung von Trends in den Daten und der *Seasonality*. Zwei Anpassungsfaktoren werden im Algorithmus verwendet: *Alpha* für das Signal und *Beta* für den Trend [SAP15, S.281].

Mathematischer Hintergrund

$$S_0 = x_0 \tag{7}$$

$$b_0 = x_1 - x_0 \tag{8}$$

$$S_t = \alpha x_t + (1 - \alpha)(S_t - 1 + b_t - 1) \tag{9}$$

$$b_t = \beta(S_t - St - 1) + (1 - \beta)b_t - 1 \tag{10}$$

Wobei α und $\beta \in$(0,1) zwei Anpassungsparameter sind. Das Modell kann als zwei Paare des *Single Exponential Smoothing* verstanden werden. Der Forecast erfolgt wie folgt:

$$F_{T+m} = S_T + mb_T \tag{11}$$

Einstellungen in PAL Der Algorithmus akzeptiert keine Null-Werte oder fehlende Daten im Input. Ansonsten sollten die Daten numerisch sein.

Die Input-Tabelle enthält zwei Spalten: *ID* und die *Raw Daten* die vorherzusagen sind. Die Vorhersage erfolgt anhand von Vergangenheitsdaten und unter Betrachtung von Trends in den Daten. Dies erfolgt mit Hilfe der Anpassungsfaktoren, deren Werte zwischen 0 und 1 liegen: *Alpha* für die Fehlerkorrektur und *Beta* für den Trend.

Durch die verschiedenen Einstellungen in PAL kann die Anzahl der vorherzusagenden Perioden eingeben werden. Die Output-Tabelle enthält die ID entsprechend der vorherzusagenden Perioden und die vorhergesagten Werte.

Der Parameter *MEASURE_NAME* ermöglicht es anzugeben, welche Fehlerkennzahlen ausgegeben werden sollen [SAP15, S.282-285].

11.1.6 Triple Exponential Smoothing

Triple Exponential Smoothing eignet sich für die Modellierung von Zeitreihen mit Betrachtung von Trends und *Seasonality* in den Daten. Drei Anpassungsfaktoren werden im Algorithmus verwendet: *Alpha* für das Signal, *Beta* für den Trend und *Gamma* für die *Seasonality* [SAP15, S.286].

Mathematischer Hintergrund

$$S_t = \alpha \times \frac{X_t}{C_{t-L}} + (1 - \alpha) \times (S_{t-1} + B_{t-1}) \tag{12}$$

$$B_t = \beta \times (S_t - S_{t-1} + (1 - \beta) \times B_{t-1}) \tag{13}$$

$$C_t = \gamma \times \frac{X_t}{S_t} + (1 - \gamma) \times C_{t-L} \tag{14}$$

$$F_{t+m} = (S_t + m \times B_t) \times C_{t-L+1+((m-1)modL)} \tag{15}$$

Einstellungen in PAL Der Algorithmus akzeptiert keine Null-Werte oder fehlende Daten im Input. Ansonsten sollten die Daten numerisch sein.

Die Input-Tabelle enthält zwei Spalten: *ID* und die *Raw Daten*, die vorherzusagen sind. Die Vorhersage erfolgt anhand von Vergangenheitsdaten und unter Betrachtung von Trends und Seasonality in den Daten. Dies erfolgt mit den Anpassungsfaktoren, deren Werte zwischen 0 und 1 liegen: *Alpha* für die Fehlerkorrektur, *Beta* für den Trend und *Gamma* für die Seasonality.

Durch die verschiedenen Einstellungen in PAL können die Anzahl der vorherzusagenden Perioden sowie die Zyklen für die Seasonality, z.B. 12 Monate oder 4 Quartale, eingegeben werden.

Die Output-Tabelle enthält die ID entsprechend der vorherzusagenden Perioden und die vorhergesagten Werte.

Der Parameter *MEASURE_NAME* ermöglicht es anzugeben, welche Fehlerkennzahlen ausgegeben werden sollen. [SAP15, S.289-293].

11.2 Regressionsanalyse

In diesem Abschnitt werden die verschiedenen Algorithmen zur Regressionsanalyse vorgestellt. Hierzu wird der entsprechende Algorithmus sowie dessen Funktionsweise vorgestellt.

11.2.1 Bi-variate natürliche logarithmische Regression

Die *bi-variate natürliche logarithmische Regression* ist ein Ansatz zur Modellierung der Beziehung zwischen einer skalaren Variable y und einer Variablen X. Durch die natürliche logarithmische Regression werden Daten mit Hilfe von natürlichen logarithmischen Funktionen modelliert. Unbekannte Modellparameter werden aus den Daten geschätzt. Solche Modelle werden *natürliche logarithmische Modelle* genannt [SAP15, S.164].

Mathematischer Hintergrund In PAL wird die natürliche logarithmische Regression durchgeführt, indem eine Umwandlung in die lineare Regression stattfindet und diese dann gelöst wird:

$$y = \beta_1 ln(x) + \beta_0 \tag{16}$$

β_0 und β_1 sind hier die Parameter, die berechnet werden müssen. Es sei $x' = \ln(x)$. Dann ist

$$y = \beta_0 + \beta_1 \times x' \tag{17}$$

Somit haben y und x' eine lineare Beziehung und können mit dem linearen Regressionsverfahren gelöst werden.

Die Implementierung unterstützt auch die Berechnung des F-Werts und R^2, um die statistische Signifikanz zu bestimmen [SAP15, S.164].

Einstellungen in PAL Der Algorithmus akzeptiert keine Null-Werte oder fehlende Daten im Input. Außerdem sollten die Daten numerisch sein. Angesichts der Struktur von Y und X müssen zudem mehr als 2 Datensätze für die Analyse zur Verfügung stehen.

PAL beinhaltet zum einen die logarithmische Regressionsfunktion und zum anderen eine auf dem Ergebnis der logarithmischen Regression basierende Vorhersagefunktion.

Die Input-Tabelle der logarithmischen Regressionsfunktion enthält eine Spalte mit der ID sowie jeweils eine Spalte mit der skalaren Variablen y und der Variablen X.

Mit Hilfe der in PAL verfügbaren Parameter kann die Regressionsfunktion genauer eingestellt werden.

Für die Funktion existieren vier verschiedene Output-Tabellen, die neben der ID jeweils eine weitere Spalte enthalten. Die *Result*-Tabelle listet die entsprechenden Koeffizienten (A_i, Beta-Koeffizient) auf. Die zweite Tabelle *Fitted Data* beinhaltet die ID und den vorhergesagten Wert Y_i. In der Tabelle *Significance* befinden sich die Messkriterien bezüglich der Zuverlässigkeit und Genauigkeit (R2/ F). In der vierten Tabelle ist das logarithmische Regressionsmodell im Predictive Model Markup Language (PMML)-Format enthalten [SAP15, S.165 - 167].

Die auf dem Ergebnis der logarithmischen Regression basierende Vorhersagefunktion besteht zwei Input-Tabellen. Die erste Tabelle enthält die Vorhersagedaten (ID, Variable X) und die zweite beinhaltet den Koeffizienten (A_i-Wert oder PMML-Modell).

Durch die in PAL verfügbaren Parameter kann auch diese Funktion genauer eingestellt werden.

Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert [SAP15, S. 167 - 169].

11.2.2 Bi-variate geometrische Regression

Die *geometrische Regression* wird als Ansatz verwendet, um die Beziehung zwischen einer skalaren Variablen y und einer Variablen X zu modellieren. Im geometrischen Regressionsmodell werden durch geometrische Funktionen Daten modelliert. Unbekannte Modellparameter werden aus den Daten geschätzt. Solche Modelle werden *geometrische Modelle* genannt [SAP15, S. 156].

Mathematischer Hintergrund In PAL wird die geometrische Regression durchgeführt, indem eine Umwandlung in die lineare Regression stattfindet und diese dann gelöst wird:

$$y = \beta_0 \times x^{\beta_1} \tag{18}$$

β_0 und β_1 sind hier die Parameter, die berechnet werden müssen.

Die Schritte sind:

1. Setzen der natürlichen logarithmischen Operation auf beide Seiten:
   ln (y) = ln ($\beta_0 \times x^{\beta_1}$)
2. Umwandlung in: ln (y) = ln (β_0) + β_1 × ln (x)
3. Es sei $y' = ln(y)$, $x' = ln(x)$, $\beta_0' = ln(\beta_0)$ $y' = \beta_0' + \beta_1 \times x'$

Somit haben y und x' eine lineare Beziehung und können mit dem linearen Regressionsverfahren gelöst werden. Die Implementierung unterstützt auch die Berechnung des F-Werts und R^2, um die statistische Signifikanz zu bestimmen [SAP15, S. 156].

Einstellungen in PAL Die geometrische Regressionsfunktion akzeptiert keine Null-Werte oder fehlende Daten im Input. Außerdem sollten die Daten numerisch sein.

PAL beinhaltet zum einen die geometrische Regressionsfunktion (*GEOREGRESSION*) und zum anderen eine auf dem Ergebnis der geometrischen Regression basierende Vorhersagefunktion (*FORECASTWITHGEOR*).

Die Input-Tabelle der geometrischen Regressionsfunktion enthält eine Spalte mit der ID sowie jeweils eine Spalte mit der skalaren Variablen y und der Variablen x.

Mit Hilfe der in PAL verfügbaren Parameter kann die Regressionsfunktion genauer eingestellt werden.

Für die Funktion existieren vier verschiedene Output-Tabellen, die neben der ID jeweils eine weitere Spalte enthalten. Die *Result*-Tabelle listet die entsprechenden Koeffizienten (A_i, Beta-Koeffizient) auf. Die zweite Tabelle *Fitted Data* beinhaltet die ID und den vorhergesagten Wert Y. In der Tabelle *Significance* befinden sich die Messkriterien bezüglich der Zuverlässigkeit und Genauigkeit (R^2/ F). In der vierten Tabelle ist das logarithmische Regressionsmodell im PMML-Format enthalten [SAP15, S.158 - 160].

Die auf dem Ergebnis der geometrischen Regression basierende Vorhersagefunktion besteht aus zwei Input-Tabellen. Die erste Tabelle enthält die Vorhersagedaten (ID, Variable X) und die zweite beinhaltet den Koeffizienten (Ai-Wert oder PMML-Modell).

Durch die in PAL verfügbaren Parameter kann auch diese Funktion genauer eingestellt werden.

Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert [SAP15, S.120 - 121].

11.2.3 Multilineare Regression

Die multilineare Regression kann bei der Untersuchung linearer Beziehungen zwischen einer abhängigen Variable Y und mehreren unabhängigen Variablen, die als Prädikatoren (*X1*, *X2*, *X3*, *Xn*) bezeichnet werden, genutzt werden. Sie ist eine Erweiterung der linearen Regression, wobei zwischen einem abhängigen Merkmal (y, Ziel der Untersuchung) und einem unabhängigen Merkmal (x, Beeinflussungsvariable) bivariate Zusammenhänge betrachtet werden.

Mathematischer Hintergrund Die mathematische Funktion bildet eine Gleichung, in der die Werte der X-Variablen zur Schätzung der Y-Werte linear kombiniert werden [SAP15, S.180]:

$$Y = \beta_0 + \beta X_1 + \beta X_2 + \beta X_3 \dots \beta X_n + e_i \tag{19}$$

Wobei Y die Zielvariable darstellt. β_0 ist ein konstantes Glied, wenn X_1 und $X_2 = 0$ sind. β_1, β_2, β_n = geben die Gewichtung des Prädikators bei der Vorhersage an. X_1, X_2,...,X_n sind beobachtete unabhängige Variablen und e_1 ist der Rest-/Prognosefehler. Die multilineare Regression ermöglicht, kausale Abhängigkeiten zwischen bestimmten Merkmalen zu untersuchen und hilft insbesondere festzustellen, wie gut die Auswahl einer Reihe unabhängiger Variablen zusammen die abhängige Variable vorhersagt. Auch ist es möglich festzustellen, wie die bestimmte Variable *X1* die Vorhersage einer abhängigen Variable y verbessern kann, wenn gleichzeitig mehrere unabhängige Variablen beim Forecast genutzt werden.

Einstellungen in PAL Der Algorithmus funktioniert in PAL unter der Bedingung, dass keine Null-Werte oder fehlenden Daten im Input angegeben werden. Die Daten sollten numerisch und nicht kategorisch sein [SAP15, S.181].

PAL beinhaltet zum einen die multiple lineare Regressionsfunktion und zum anderen eine auf dem Ergebnis der multiplen linearen Regression basierende Vorhersagefunktion. Die Input-Tabelle enthält drei Spalten: ID, Y-Zielvariable (die vorherzusagen ist) und je nach Anzahl der unabhängigen Variablen X für jede eine Spalte. Die Vorhersage erfolgt anhand der Vergangenheitsdaten.

Mit Hilfe der in PAL verfügbaren Parameter kann die multiple lineare Regressionsfunktion genauer eingestellt werden.

Die Output-Tabelle enthält dann die Spalte ID mit den entsprechenden vorherzusagenden Perioden und Koeffizienten (A_i, Beta-Koeffizient) für die Funktion. In der Tabelle *Fitted* gibt es eine Spalte mit der ID und den vorherzusagenden Werten. In der Tabelle *Significance* befinden sich die Messkriterien bezüglich der Zuverlässigkeit und Genauigkeit R2 und F. Die auf dem Ergebnis der multiplen linearen Regression basierende Vorhersagefunktion besteht aus zwei Input-Tabellen. Die erste Tabelle enthält die Vorhersagedaten (ID, Variable X) und die zweite beinhaltet den Koeffizienten (Ai-Wert oder PMML-Modell). Die Einstellungen bei der Prognose können auch an dieser Stelle durch die in PAL verfügbaren Parameter eingestellt werden. Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert [SAP15, S.182-184].

11.2.4 Polynomiale Regression

Falls es sich um einen nicht linearen Zusammenhang zwischen einer abhängigen Zielvariablen Y und einer unabhängigen Variablen x handelt, besteht die Option, das Modell mit quadratischen oder Termen höherer Ordnung von x zu optimieren, um bessere Ergebnisse bei der Vorhersage zu gewinnen. Die Abhängigkeiten sind dann auch auf einer gekrümmten Geraden feststellbar. Dies kann unter Verwendung der polynomialen Regression erreicht werden [SAP15, S.190].

Mathematischer Hintergrund Die mathematische Funktion bildet folgende Gleichung:

$$y = \beta_0 + \beta_1 \times x + \beta_2 \times x^2 + \ldots + \beta_n \times x^n \tag{20}$$

$\beta_0 ... \beta_n$ sind die Parameter, die berechnet werden müssen [SAP15, S.190].

Als Grad des Polynoms wird die vorkommende Potenz n von x bezeichnet, Polynome zweiten Grades werden z.B. quadratisch genannt, solche von Grad 3 kubisch usw.

y stellt die zu berechnende Zielvariable dar. x^1, x^2,...,x^n sind beobachtete unabhängige Variablen. n ist die Anzahl Polynome und β stellt den Koeffizienten dar. Unter der Annahme, dass verschiedene Potenzen von x als eigenständige erklärende Variablen angesehen werden, könnte die polynomiale Regression als Spezialfall der multiplen Regression interpretiert werden mit $x_1 = x^1 +, x_2 = x^2, x_3 = x^n$.

Einstellungen in PAL Der Algorithmus funktioniert in PAL unter der Bedingung, dass keine Null-Werte oder fehlenden Daten im Input angegeben werden. Die Daten sollten numerisch und nicht kategorisch sein.

PAL beinhaltet zum einen die polynomiale Regressionsfunktion (*POLYNOMIALREGRESSION*) und zum anderen eine auf dem Ergebnis der geometrischen Regression ba-

sierende Vorhersagefunktion (*FORECASTWITHPOLYNOMIALR*). Die Input-Tabelle der polynomialen Regressionsfunktion enthält eine Spalte mit der ID sowie jeweils eine Spalte mit der skalaren Variablen y und der Variablen x.

Mit Hilfe der in PAL verfügbaren Parameter kann die polynomiale Regressionsfunktion genauer eingestellt werden.

Für die Funktion existieren vier verschiedene Output-Tabellen, die neben der ID jeweils eine weitere Spalte enthalten. Die Output-Tabelle enthält dann die Spalte ID mit den entsprechenden vorherzusagenden Perioden und Koeffizienten (A_i, Beta-Koeffizient) für die Funktion. In der Tabelle *Fitted* gibt es eine Spalte mit der ID und den vorherzusagenden Werten. In der Tabelle *Significance* befinden sich die Messkriterien bezüglich der Zuverlässigkeit und Genauigkeit R^2 und F. In der vierten Tabelle ist das polynomiale Regressionsmodell im PMML-Format enthalten. Die auf dem Ergebnis der polynomialen Regression basierende Vorhersagefunktion besteht aus zwei Input-Tabellen. Die erste Tabelle enthält die Vorhersagedaten (ID, Variable X) und die zweite beinhaltet den Koeffizienten (Ai-Wert oder PMML-Modell). Die Einstellungen bei der Prognose können auch an dieser Stelle durch die in PAL verfügbaren Parameter eingestellt werden. Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert. [SAP15, S.191 - 193].

11.2.5 Exponentielle Regression

Mit diesem Algorithmus ist es möglich, verborgene Daten in den Datentrends aufzufinden. Der Algorithmus führt eine eindimensionale Regressionsanalyse aus. Zur Ermittlung des Einflusses einer einzelnen Variablen auf eine andere Variable wird eine Exponentialfunktion mit der Methode des kleinsten Quadrates verwendet [SAP15, S.172].

Mathematischer Hintergrund Die mathematische Funktion bildet folgende Gleichung:

$$y = \beta_0 \times exp(\beta_1 \times x^1 + \beta_2 \times x^2 + \ldots + \beta_n \times x^n) \tag{21}$$

$\beta_0 \ldots \beta_n$ sind die Parameter, die berechnet werden müssen.

Im Gegensatz zur polynomialen Regression wird die exponentielle im PAL in die lineare Regression umgewandelt und gelöst.

Die Schritte sind

1. Logarithmieren der Gleichung ln(y) = ln
   $(\beta_0 \times \exp(\beta_1 \times x^1 + \beta_2 \times x^2 + \ldots + \beta_n \times x^n))$

2. Umformung in ln(y) = ln
   $(\beta_0) + \beta_1 \times x^1 + \beta_2 \times x^2 + \ldots + \beta_n \times x^n$
3. Gleichsetzung y' = ln(y),
   $\beta_0' = \ln(\beta_0)$y' $= \beta_0' + \beta_1 \times x^1 + \beta_2 \times x^2 + \ldots + \beta_n \times x^n$

Somit bilden die y', $x^1, \ldots, x^n$ eine lineare Beziehung, die gelöst werden kann [SAP15, S.173].

Einstellungen in PAL Der Algorithmus funktioniert in PAL unter der Bedingung, dass keine Null-Werte oder fehlenden Daten im Input angegeben werden. Die Daten sollten numerisch und nicht kategorisch sein.

PAL beinhaltet zum einen die exponentielle Regressionsfunktion (*EXPREGRESSION*) und zum anderen eine auf dem Ergebnis der exponentiellen Regression basierende Vorhersagefunktion (*FORECASTWITHEXPR*). Die Input-Tabelle der exponentiellenen Regressionsfunktion enthält eine Spalte mit der ID sowie jeweils eine Spalte mit der skalaren Variablen y und der Variablen x.

Mit Hilfe der in PAL verfügbaren Parameter kann die Regressionsfunktion genauer eingestellt werden.

Für die Funktion existieren vier verschiedene Output-Tabellen, die neben der ID jeweils eine weitere Spalte enthalten. Die Output-Tabelle enthält dann die Spalte ID mit den entsprechenden vorherzusagenden Perioden. In der Tabelle *Fitted* gibt es eine Spalte mit der ID und den vorherzusagenden Werten. In der Tabelle *Significance* befinden sich die Messkriterien bezüglich der Zuverlässigkeit und Genauigkeit R^2 und F. In der vierten Tabelle ist das exponentielle Regressionsmodell im PMML-Format enthalten. Die auf dem Ergebnis der exponentiellen Regression basierende Vorhersagefunktion besteht aus zwei Input-Tabellen. Die erste Tabelle enthält die Vorhersagedaten (ID, Variable X) und die zweite beinhaltet den Koeffizienten (Ai-Wert oder PMML-Modell). Die Einstellungen bei der Prognose können auch an dieser Stelle durch die in PAL verfügbaren Parameter eingestellt werden. Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert [SAP15, S.172 - 179].

11.3 Support Vector Machine (SVM)

Die Support Vector Machine (SVM) ist ursprünglich ein Klassifikationsalgorithmus, der auch für eine Regression verwendet werden kann. Im ersten Abschnitt wird kurz die Funktionsweise allgemein innerhalb der Klassifikation erklärt. Der zweite Abschnitt erklärt die Regression. Anschließend werden die Anwendung sowie Vor- und Nachteile beleuchtet.

Funktionsweise allgemein Die Support Vector Machine (SVM) ist ein Klassifikationsalgorithmus, der von den 90er Jahren bis heute stark an Popularität gewonnen hat. Sie basiert auf einem einfachen und intuitiven Klassifikator, der *maximal margin classifier* genannt wird [JWHT13, Vgl. S. 337ff.]. Mit Hilfe verschiedener Techniken ist die Support Vector Machine in der Lage, lineare und nicht lineare Klassifikationen vorzunehmen [JWHT13, Vgl. S. 337ff.].

Die Klassifikation findet in einem p-dimensionalen Raum statt. Eine sogenannte *Hyperebene* ist ein Teilraum mit der Dimensionalität p - 1. Eine Hyperebene bestehend aus zwei Dimensionen ist eine Fläche. Eine Hyperebene bestehend aus einer Dimension ist eine Linie. Eine zweidimensionale Hyperebene wird beispielsweise in (22) abgebildet [JWHT13, Vgl. S. 338].

$$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 = 0 \tag{22}$$

Folgend die allgemeine Definition einer p-dimensionalen Hyperebene:

$$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p = 0 \tag{23}$$

Wenn die Gleichung (25) für einen Punkt

$$X = (X_1, X_2, ..., X_p) \tag{24}$$

gilt, hat dies die Bedeutung, dass der Punkt sich auf einer Seite der Hyperebene befindet. Im Falle von (26) befindet sich der Punkt auf der anderen Seite der Hyperebene [JWHT13, Vgl. S. 338].

$$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p > 0 \tag{25}$$

$$\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_p * X_p < 0 \tag{26}$$

Die zu klassifizierenden Daten werden in Form einer

$$n \times p \tag{27}$$

Matrix dargestellt. Eine Klassifikation wird, wie definiert, mit Hilfe einer *separierenden*

Hyperebene durchgeführt. Abbildung 19 zeigt eine solche Hyperebene, die zwei Klassen von Daten korrekt klassifiziert [JWHT13, Vgl. S. 338ff].

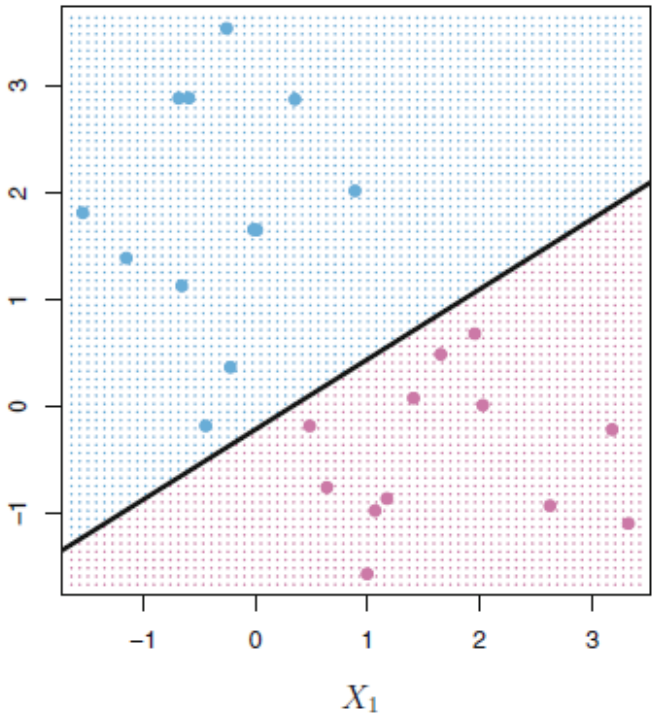


Abbildung 19: Separierende eindimensionale Hyperebene [JWHT13, Vgl. S. 340]

Theoretisch können die Daten durch unendlich viele separierende Hyperebenen klassifiziert werden. Eine spezielle Form einer separierenden Hyperebene ist die *maximal margin hyperplane*. Diese Hyperebene ist jene, die den größten Abstand zu den Punkten der Trainingsdaten hat. Abbildung 20 stellt eine solche Hyperebene grafisch dar [JWHT13, Vgl. S. 441].

Folgend die formale Definition. (28) definiert die Beobachtungswerte der Trainingsmenge für das Modell. (29) definiert zwei verschiedene Klassen, in die ein Punkt eingeteilt werden kann.

$$x_1 ... x_n \in \mathbb{R}^p \tag{28}$$

$$y_1 ... y_n \in \{-1, 1\} \tag{29}$$

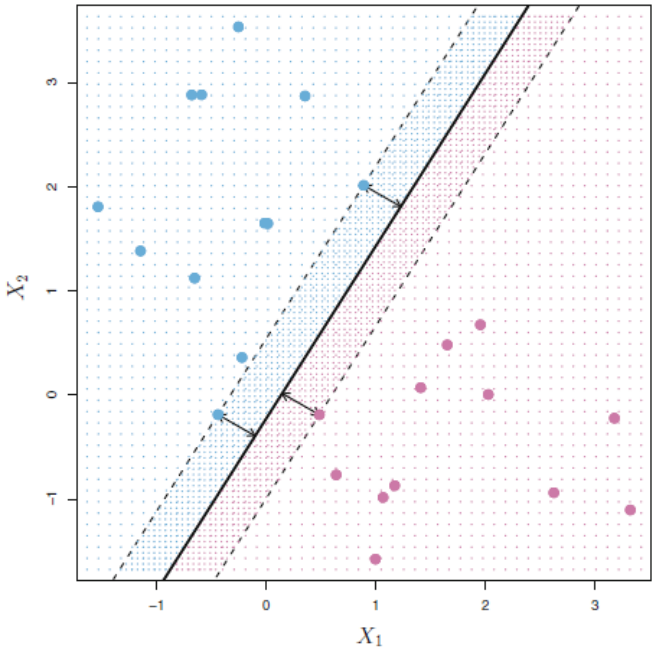


Abbildung 20: maximum margin hyperplane [JWHT13, Vgl. S. 342]

Bei der Erzeugung der Hyperebene wird dieses Optimierungsproblem gelöst:

$$\begin{aligned} &maximize\ M \\ &\beta_0, \beta_1, ..., \beta_p \end{aligned} \tag{30}$$

$$subject\ to \sum_{j=1}^{p} \beta_j^2 = 1 \tag{31}$$

$$y_i(\beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_p * x_{ip}) \geq M\ \forall\ i = 1, ..., n \tag{32}$$

Es werden also

$$\beta_0, \beta_1, ..., \beta_p \tag{33}$$

ermittelt, um den Abstand M zu maximieren [JWHT13, Vgl. S. 342f.].

Der Vorteil einer solchen Hyperebene ist, dass bei der Klassifikation weiterer Daten die Wahrscheinlichkeit größer ist, dass diese korrekt klassifiziert werden, als bei einer Hyperebene mit geringerem Abstand. Der Klassifikator wird als *maximum margin classifier* bezeichnet [JWHT13, Vgl. S. 441].

Tatsächlich wird innerhalb einer Support Vector Machine ein Klassifikator verwendet, der

weniger anfällig gegenüber Ausreißern in den Trainingsdaten ist, mit dem die meisten Beobachtungswerte korrekt klassifiziert werden. Dieser Klassifikator wird *support vector classifier* genannt. Der wesentliche Unterschied zum *maximum margin classifier* besteht darin, dass weitere Optimierungsparameter (beispielsweise C) hinzukommen, die näherungsweise das Modell optimieren [JWHT13, Vgl. S. 445 ff.].

Anstelle der Nutzung von linearen Klassifikatoren (*p features*) können auch nichtlineare Klassifikatoren (beispielsweise *2p features*) verwendet werden, um nichtlineare Klassifikationsaufgaben zu lösen [JWHT13, Vgl. S. 350].

Funktionsweise SVM-Kernel Ein *Support Vector Classifier* ist eine Klassifikationsmethode für eine 2-Klassen-Einstellung, in der die Grenze zwischen zwei Klassen linear ist. Jedoch ist in der Praxis oft eine nicht-lineare Klassentrennung aufzufinden, weshalb der Support Vector Classifier um diese Funktion erweitert werden muss. Die *Support Vector Machine* ist eben diese Erweiterung des Support Vector Classifiers, die aus der spezifizierten Vergrößerung des Merkmalsraums durch Hinzunahme von *Kerneln* entsteht. Ein Kernel $K(x_i, x_{i'})$ ist in diesem Zusammenhang eine Funktion, die die Ähnlichkeit zweier Beobachtungen quantifiziert.

Ein linearer Kernel, welcher uns den *Support Vector Classifier* zurückgibt, hat dabei die Form

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{34}$$

[JWHT13][Vgl. S. 352] und befähigt uns die Ähnlichkeit eines Beobachtungspaares durch eine standardisierte *Pearson Korrelation* zu untersuchen. Der polynomiale Kernel wiederum hat die Form

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d, \tag{35}$$

[JWHT13][Vgl. S. 352] wobei d eine positive Zahl ist, die den Grad der Funktion darstellt. Ein solcher nicht zwangsläufig linearer Kernel erlaubt flexiblere Entscheidungen bezüglich der Grenzen der Klassifikation. Dieser Kernel führt dazu, dass ein *Support Vector Classifier* nun mithilfe von polynomialen Funktionen eines bestimmten Grades auf höher-dimensionale Räume angewendet werden kann. Ein letzter wichtiger Kernel ist der *radiale Kernel* in der Form:

$$K(x_i, x_{i'}) = exp(-\gamma * \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2), \tag{36}$$

[JWHT13][Vgl. S. 352] Hierbei ist γ eine positive Konstante. Der Kernel vergleicht im

Grunde die eigentlichen Daten mit den Trainingsdaten und der Einfluss der einzelnen Werte wird dabei an der *euklidischen Distanz* dieser zu den Trainingsdaten festgelegt. So kann trotz Ausreißern bei den Werten eine gute Klassifizierung erreicht werden.

Funktionsweise bezüglich Regressionen Im Laufe der Zeit haben sich zwischen SVMs und klassischen statistischen Methoden stärkere Verbindungen entwickelt. So können die nötigen Kriterien passend für einen *Support Vector Classifier* im Stile von $f(X) = \beta_0 + \beta_1 * X_1 + ... + \beta_p * X_p$ [JWHT13][Vgl. S. 356] folgendermaßen umgeschrieben werden:

$$minimize_{\beta_0,\beta_1,...,\beta_p} \left\{ \sum_{i=1}^{n} max[0, 1 - y_i * f(x_i)] + \lambda * \sum_{j=1}^{p} \beta_j^2 \right\} \tag{37}$$

[JWHT13][Vgl. S. 356]

Dabei wird das Modell durch β_p parametrisiert und fungiert so als die Koeffizienten der polynomialen Funktion. λ ist ein nicht-negativer Anpassungsparameter. Bei einem größeren λ ist $\beta_1, ..., \beta_p$ klein, wodurch mehr Verstöße für den festgelegten Abstand (margin) erlaubt werden. Dies resultiert in einem Klassifikator, der eine niedrige Varianz und eine hohe Verzerrung aufweist. Ist λ klein und damit $\beta_1, ..., \beta_p$ groß, so werden weniger Verstöße erlaubt und der Klassifikator weist eine hohe Varianz und eine niedrige Verzerrung auf. λ steht dabei in direkter Relation zum nicht-negativen Anpassungsparameter C, in dem die Schlupfvariablen $\epsilon_1, ..., \epsilon_n$ zusammengefasst werden. ϵ umfasst die individuellen Beobachtungen der Werte, die auf der falschen Seite des Abstands (margin) sind. Dabei wird diese Formel zum besseren Verständnis als *Loss + Penalty* -Formel vereinfacht:

$$minimize_{\beta_0,\beta_1,...,\beta_p} \{L(\mathbf{X}, \mathbf{y}, \beta) + \lambda * P(\beta)\} \tag{38}$$

[JWHT13][Vgl. S. 357]

Unter $L(\mathbf{X}, \mathbf{y}, \beta)$ ist die Verlustfunktion (Loss) zu verstehen, die das Ausmaß der Eignung des Moduls in Bezug auf die genutzten Daten quantifiziert. $P(\beta)$ ist die Straffunktion (Penalty), die den Bereich, in dem ein Verstoß registriert wird, festlegt.

SVM Anwendung Nun werden in der Praxis für die verschiedenen Kernel nicht zwangsläufig dieselben Parameter verwendet. In allen Kerneln ist die Definition des Anpassungsparameters C 11.3 und der Schlupfvariablen ϵ 11.3 wichtig. Bei dem polynomialen Kernel kommt der Anpassungsparameter λ 11.3 sowie der Grad d 11.3 der Polyminialfunktion hinzu. Für den radialen Kernel kommt außerdem noch die positive Konstante γ 11.3 hinzu.

Einstellungen in PAL Der Algorithmus funktioniert in PAL unter der Bedingung, dass keine Null-Werte oder fehlenden Daten im Input angegeben werden. Die Daten sollten numerisch und nicht kategorisch sein.

PAL beinhaltet zum einen die Funktion (*SVMTRAIN*), die das Modell generiert und zum anderen eine auf dem Ergebnis der (*SVMTRAIN*) basierende Vorhersagefunktion (*SVMPREDICT*). Die Input-Tabelle der SVMTRAIN-Funktion enthält eine Spalte mit der ID sowie jeweils eine Spalte mit der Datenwerten. Mit Hilfe der in PAL verfügbaren Parameter kann die SVMTRAIN-Funktion genauer eingestellt werden. Für die Funktion existieren zwei verschiedene Output-Tabellen. Die Tabelle Model Result (part one) beinhaltet 2 Spalten, Name und Wert der Parameter. Die Tabelle Model Result (part two) beinhaltet ID und berechneten Alpha-Koeffizient. Die auf dem Ergebnis der SVMTRAIN basierende Vorhersagefunktion besteht aus einer Input-Tabelle, die 3 Spalten beinhalten, in denen sich Vorheragedaten wie z.B. (ID) befinden. Die Einstellungen bei der Prognose können auch an dieser Stelle durch die in PAL verfügbaren Parameter eingestellt werden. Die Output-Tabelle der Funktion, *Fitted Result*, enthält neben der ID den vorhergesagten Y_i-Wert [SAP15, S.142 - 149].

Vor- und Nachteile Die SVM bietet eine hohe Generalisierungsfähigkeit, mit der reale Aufgabenstellungen gut gelöst werden können. Außerdem kann mit einem höherdimensionalen Raum umgegangen werden [JWHT13][Vgl. S. 337ff.]. Für weitere Eingabedaten ist ein neues Training erforderlich. Der jeweils für die Aufgabenstellung passende Kernel muss empirisch gesucht werden [JWHT13][Vgl. S. 337ff.].

11.4 Messkriterien

Um das Ziel des Projekts genauer zu erfüllen, werden verschiedene Messkriterien für die binären Systeme beschrieben, die bei der Bewertung helfen und anhand von Beispielen erklärt werden. Nach der Ausführung der prädiktiven Modelle sollten jeweils die Genauigkeit und die Zuverlässigkeit des Modells evaluiert werden. Dafür sind die folgenden Methoden und Maßnahmen relevant:

11.4.1 R-Squared

Der Determinationskoeffizient, auch Bestimmtheitsmaß oder R^2 genannt, bestimmt die Güte der Anpassung des Algorithmus und des erstellten Modells an die Daten selbst [Lei13]. Das Bestimmtheitsmaß ist wie folgt definiert:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)}{\sum_{i=1}^{N}(Y_i - \bar{Y})} \tag{39}$$

Dabei ist Y_i der Erhebungswert an i-ter Stelle, $\hat{Y}_i$ der Prognosewert an i-ter Stelle und $\bar{Y}$ das arithmetische Mittel der Erhebungsmenge. Durch die zugrunde liegende Streuungszerlegungsformel lässt sich schließen, dass die Anpassung des Modells an die Daten mit steigendem R-Squared verbessert wird und bei $R^2 = 1$ perfekt ist, da in diesem Fall alle Residuen gleich Null sind. Für $R^2 = 0$ hingegen ist die Anpassung des Modells sehr schlecht, d. h. die Vorhersage entspricht stets dem Mittelwert der Zielgröße und es besteht keine Abhängigkeit zu den erklärenden Variablen [Lei13].

11.4.2 R-Squared adjusted

Je mehr unabhängige Variablen in der Prognose verwendet werden, desto größer ist der R-Squared-Wert. Dies bedeutet aber nicht, dass die Genauigkeit unbedingt höher ist. Auch mit zufälligen Werten und Variablen, die keinen Bezug auf die Prognose haben, wird R-squared größer. Im Gegensatz zu R-squared ist die angepasste Version *R squared adjusted* von Bedeutung um die Genauigkeit zu ermitteln, wenn mehr als eine Spalte in der Prognose betrachtet wird. R-squared adjusted reduziert sich, wenn der „Betrag“ der neu eingefügten Spalte nicht größer als eine zufällige Spalte mit zufälligen Werten ist. Somit wird klar, wie viele Variablen für die Prognose optimal sind, um die genausten Ergebnisse zu ermitteln [GM05].

$$R^2 adjusted = 1 - \frac{(1 - R^2)(N - 1)}{(N - p - 1)} \tag{40}$$

11.4.3 MAE (Mean Absolute Error)

Mit Hilfe des Mean Absolute Error (MAE) wird der durchschnittliche Betrag der Abweichung von den n Vorhersagewerten $\hat{y}^i$ zu den tatsächlich gemessenen Zielwerten y^i berechnet [Lei13, S. 40].

$$MAE = [n^{-1} \sum_{i=1}^{n} |e_i|] \tag{41}$$

[WM05]

11.4.4 RMSE (Root Mean Squared Error)

Eine der bedeutendsten Fehlerkennzahlen bei der Untersuchung der Qualität einer Prognose ist das Fehlerbewertungsmaß Root Mean Squared Error (RMSE). Als Differenz zu den tätsächlichen Daten berechnet der RMSE die Streuung der vorhergesagten Daten. Der

RMSE stellt ein strengeres Fehlerbewertungsmaß als die Standardabweichung dar, da er im Gegensatz dazu nicht mittelwertbereinigt ist [Lei13, S. 40].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{42}$$

Wobei Y_i und Y^i die tatsächlichen und die vorhergesagten Werte $n = i$ mit $n \geq i \geq 1$ repräsentieren [CD14].

11.4.5 Variationskoeffizienten CV(MAE) und CV(RMSE)

Der Variationskoeffizient wird als eine Kennzahl bezeichnet, die die Streuung eines Merkmals beschreibt.

Die Variationskoeffizienten *CV(MAE)* und *CV(RMSE)* sind statistische Kenngrößen und sind relativ, weil bei der Berechnung eine Normalisierung des mittleren absoluten Fehlers (MAE) bzw. der Wurzel des mittleren quadratischen Fehlers(RMSE) mit dem Mittelwert $\bar{y}$ der Zielwerte erfolgt. Aus diesem Grund sind beide Bewertungskriterien nicht abhängig von den Eingabevariablen. Auf diese Art und Weise gelingt es, die Resultate verschiedener Algorithmen zu vergleichen.

$$CV(MAE) = \frac{MAE}{\overline{y}} \tag{43}$$

$$CV(RMSE) = \frac{RMSE}{\overline{y}} \tag{44}$$

11.5 Korrelation zwischen den Faktoren

In der Statistik definiert die Korrelation einen bestehenden Zusammenhang zwischen zwei quantitativen Merkmalen. Bei einer Korrelation wird die Stärke einer statistischen Beziehung anhand von 2 Variablen (Faktoren) gemessen [Sta14].

Die Stärke des Zusammenhangs wird als eine Zahl ausgedrückt. Diese Maßzahl der statistischen Beziehung beträgt zwischen 0 und 1. Liegt der Grad der Beziehung bei „0“, so handelt es sich um *keine* Beziehung, während bei einem Wert von „1“ von der perfekten Beziehung zwischen den Faktoren ausgegangen wird.

Es existieren positive und negative Korrelationen [Hai00]. Bei positiven Korrelationen gehört zu einem hohen Wert eines Faktors ein zweiter hoher Wert des anderen Faktors. Im Gegensatz dazu liegt bei negativen Korrelationen bei einem hohen Wert eines Faktors in

der Regel ein niedriger Wert des anderen Faktors vor. Ein Beispiel für eine negative Korrelation wäre etwa eine statistische Beziehung zwischen der Variablen „aktuelles Alter“ und „verbleibende Lebenserwartung“. Es kann gesagt werden, dass je höher das aktuelle Alter einer Person, desto niedriger ist die durchschnittliche verbleibende Lebenserwartung. Ein Beispiel für eine positive Korrelation wäre eine statistische Beziehung zwischen den Variablen „Körpergröße“ und „Gewicht“. Es kann gesagt werden, dass je höher die Körpergröße ist, desto höher das durchschnittliche Gewicht des Menschen[Sta14].

Bei der Messung einer Korrelation der Faktoren liegt keine Information darüber vor, welcher Faktor den anderen beeinflusst. Aus diesem Grund folgt, dass die untersuchten Faktoren gleichberechtigt sind [Hai00].

Abschließend lässt sich sagen, dass je größer der Zusammenhang zwischen den beiden Faktoren ist, desto präziser lassen sich Prognosen von einem auf den anderen Faktor machen.

12 Test und Evaluation

In diesem Kapitel wird zunächst das von der Projektgruppe erstellte Test- und Evaluationskonzept vorgestellt. Dieses Konzept soll nicht nur die programmierte Software, sondern auch die Ergebnisse der Analysen, Simulationen und Prognosen der Hypothesen verifizieren. Zuerst erfolgt die Vorstellung des Konzepts. Anschließend werden Testergebnisse vorgestellt.

12.1 Konzept

Abbildung 21 zeigt das entwickelte Testkonzept mit dessen Hilfe das gesamte Ergebnis der Projektgruppe getestet und evaluiert wird. Das entwickelte Konzept wird im Folgenden für die einzelnen Ebenen beschrieben. Es dient der ganzheitlichen Qualitätssicherung vom Datenursprung (Quelle) bis hin zur Analyse.

Rohdaten und Vorverarbeitung Wie bereits in Kapitel 10.3 genannt, wird in der Form eines Reports für jede Datenquelle die Anzahl der gelesenen und übersprungenen Zeilen mit Begründung dokumentiert. Die Differenz (gelesen - übersprungen) beziffert die Anzahl von Zeilen, die in SAP HANA geschrieben werden sollen.

Datenimport in SAP HANA Nachdem die Daten wie in Kapitel 10.3 erläutert in SAP HANA überführt worden sind, kann mit Hilfe des bereits genannten Reports geprüft werden, ob die zu schreibende Anzahl von Zeilen mit der tatsächlichen übereinstimmt. Die tatsächlich geschriebene Anzahl von Zeilen wird in den Report mit aufgenommen. Auch wird die in Kapitel 10.3 benannte Fehlerdatei nach Einträgen analysiert.

Tabellen in HANA Sobald der Datenimport abgeschlossen ist, sind die Datensätze in die spezifizierte Tabelle geschrieben. Von hier aus können weitere Tests auf die Korrektheit der Daten durchgeführt werden. Es erfolgt die Hinzunahme dieser Kennzahlen:

- Min-/Max-Werte der einzelnen Spalten
- Test auf leere Datensätze
- Zählen der gesamten Datensätze
- Prüfung auf ungültige Werte

Beispiel: Die Datentabelle `DWD_WEATHER_AIRTEMP` beinhaltet die Spalte `Air_Temp`, in der die Temperaturen der verschiedenen Messstationen gespeichert sind. Auf dieser Spalte kann ein Test auf Min-/Max-Werte durchgeführt werden. Damit erfolgt die Identifikation von Werten, die der Spezifikation entsprechen[9].

[9] Eine Lufttemperatur von größer 40 Grad Celsius ist in Deutschland nicht realistisch [Wet14].

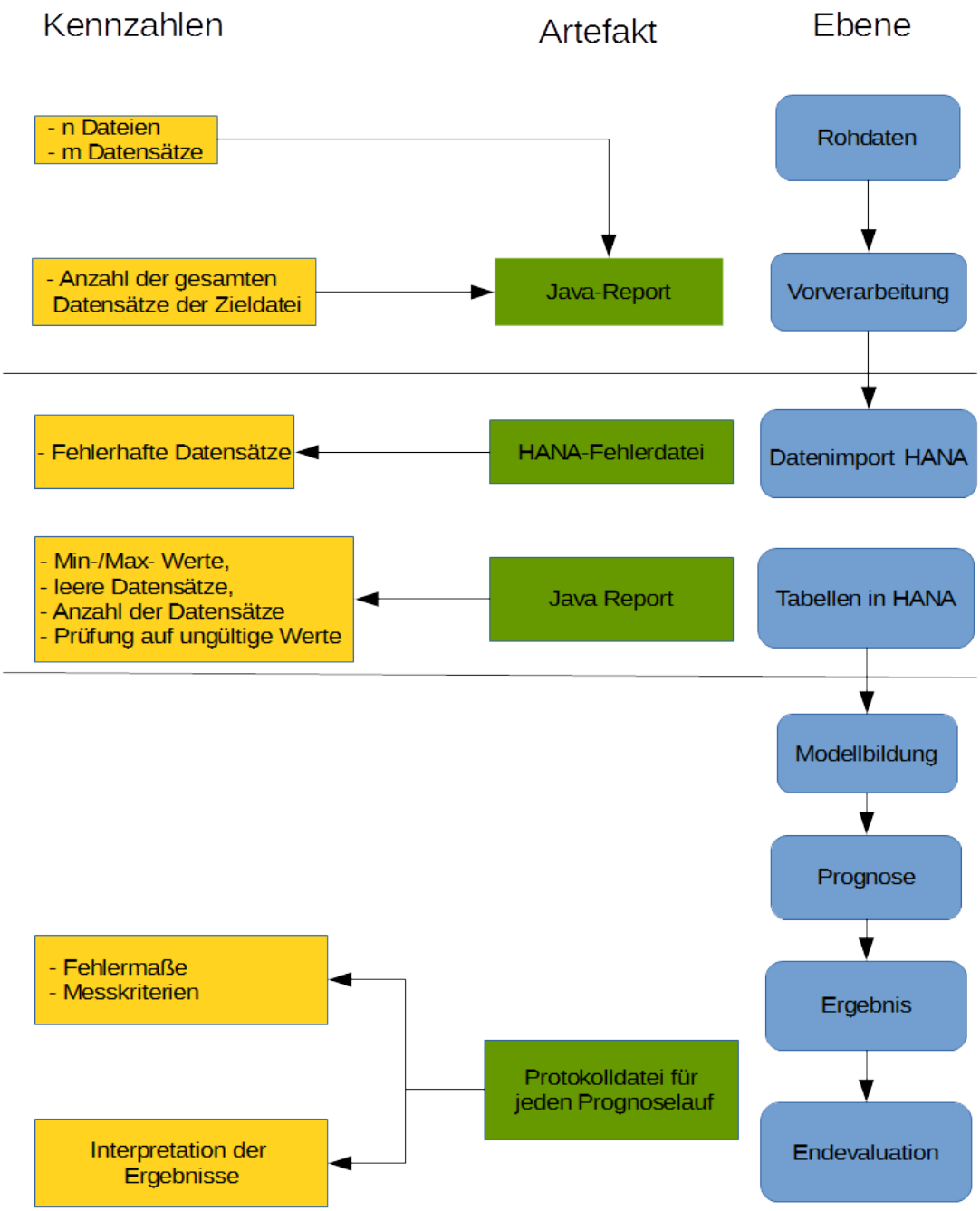


Abbildung 21: Testkonzept

Ergebnis Für jede Hypothese und jeden Algorithmus wird ein eigenständiger Ordner erstellt, in dem die Ergebnisse festgehalten werden. Das Ergebnis der Prognose (siehe auch Kapitel 10.3) wird in einem Protokoll festgehalten. Abbildung 22 zeigt beispielhaft ein solches Protokoll. Innerhalb des Dateinamens erfolgt die Benennung der Rahmenbedingungen der Modellbildung bestehend aus Trainingsdaten und Parametern. Ein Beispiel: Der Dateiname *household_industry_g0_001#c1000.xlsx* besagt, dass in den Trainingsdaten Haushalts- und Industriepreise enthalten sind und dass ein Parameter $\gamma = 0,001$ sowie ein Parameter $C = 1000$ gesetzt ist.

Innerhalb des Protokolls wird für jede Stunde innerhalb des vorherzusagenden Zeitraums der tatsächliche Verbrauch dem prognostizierten gegenübergestellt. Diese Gegenüberstellung ist Basis für die Berechnung der Messkriterien. Auf dem Protokoll werden für die Vorhersage die Messkriterien R-Squared, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) und die Variationskoeffizienten MAE und RMSE ausgegeben. Weitere Informationen zu den Messkriterien sind aus Kapitel 11.4 zu entnehmen.

tat. Verbrauch X	Forecast Y	Y-X	|(Y-X)|	$(Y-X)^2$			
46719,00	42527,46	-4191,54	4191,54	17569007,57		Anzahl der Datensätze	745
46065,00	40819,75	-5245,25	5245,25	27512647,56			
44399,00	39127,26	-5271,74	5271,74	27791242,63		Summe Y-X	*-3768749,47*
42637,00	37899,66	-4737,34	4737,34	22442390,28		Summe |Y-X|	5893778,59
41576,00	37334,17	-4241,83	4241,83	17993121,75		Summe $(Y-X)^2$	*62848557818,04*
40798,00	36575,61	-4222,39	4222,39	17828577,31		Mittelwert X	*61570,54*
39851,00	35035,33	-4815,67	4815,67	23190677,55			
38102,00	35801,83	-2300,17	2300,17	5290782,03			
38397,00	37178,97	-1218,03	1218,03	1483597,08			
39013,00	39475,97	462,97	462,97	214341,22			
41423,00	41605,22	182,22	182,22	33204,13			
43640,00	43800,47	160,47	160,47	25750,62			
46097,00	44835,97	-1261,03	1261,03	1590196,66		R-squared/R^2/Determinationskoeffizient	*0,45890547*
47080,00	44398,04	-2681,96	2681,96	7192909,44		Mean Absolute Error(MAE)	*7911,112201*
46481,00	43710,79	-2770,21	2770,21	7674063,44		Root Mean Square Error(RMSE)	*9184,796149*
46090,00	43759,43	-2330,57	2330,57	5431556,52		Variantionskoeffizienten = CV(MAE)	*0,128488599*
46631,00	46104,25	-526,75	526,75	277465,56		Variantionskoeffizienten = CV(RMSE)	*0,14917518*

Abbildung 22: Protokoll zur Prognose

Neben der Erstellung eines Protokolls inklusive Messkriterien wird die Prognose in einem orthogonalen Koordinatensystem visualisiert und dem tatsächlichen Stromverbrauch gegenübergestellt. Auf der X-Achse wird die Zeit dargestellt, auf der Y-Achse jeweils der prognostizierte und tatsächliche Verbrauch. Abbildung 23 zeigt beispielhaft eine solche Visualisierung.

Interpretation Anhand der in Kapitel 11.4 beschriebenen Messkriterien wird das Ergebnis der Vorhersage ausgewertet und interpretiert. Diese Messkriterien werden bei der Analyse bezüglich Genauigkeit und Fehlerquote evaluiert. Dabei gilt, dass der R-Squared möglichst hoch sein soll (maximal 1), und dass MAE und RMSE möglichst minimal sind. Außerdem erfolgt eine Sichtung und qualitative Bewertung des grafischen Prognoseverlaufs (siehe auch beispielhaft Abbildung 23). Qualitativ bedeutet, dass neben den Messkriterien auch visuell individuelle Merkmale erfasst werden. Beispielsweise kann dies bedeuten,

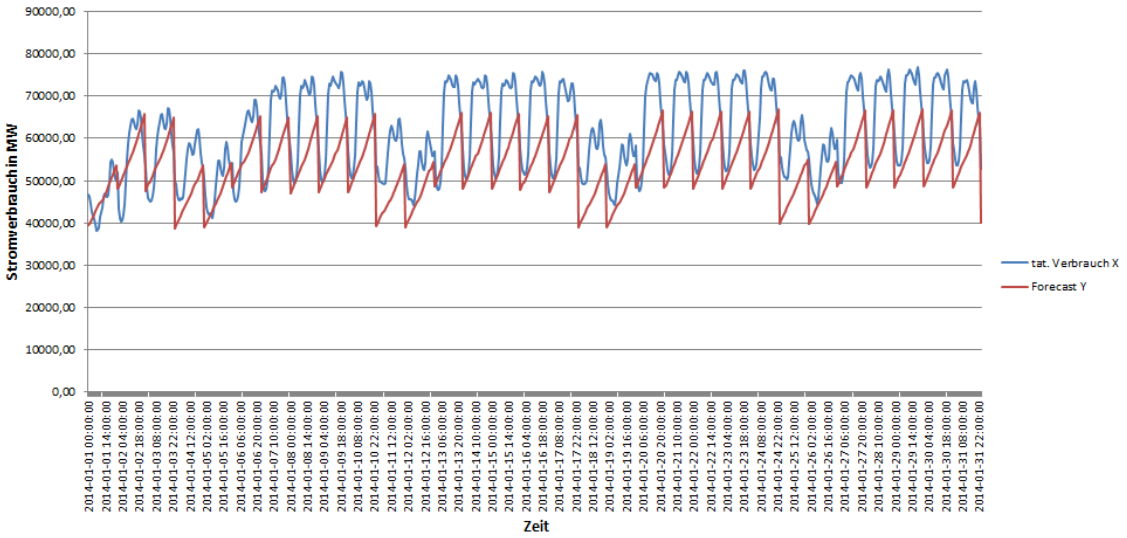


Abbildung 23: Visualisierter Verlauf

dass die Prognose zwar quantitativ schlechter ist als eine andere, jedoch Leistungsspitzen besser vorhersagt. Qualitativ hat sie so einen höheren Stellenwert.

12.2 Testergebnisse des Datenimports

Dieser Abschnitt dokumentiert die Testergebnisse des Datenimports. Hierzu wird das vorgestellte Testkonzept für die importierten Datenmengen durchgeführt. Es folgen die erstellten Reports für das Extrahieren und Laden der Daten aus verschiedenen Datenquellen in SAP HANA. Daraus ist zu entnehmen, dass eine vollständige Überführung stattgefunden hat.

DWD Temperatur Import Report
Count of read lines: 106987541

Total count of skipped lines: 80328414

Reasons for skipped lines:
Year is before 2009: 80165211
recent data and Year is before 2014: 156479
Air temp is under -100 or over 100: 6724

Count of written lines on SAP HANA: 26659127

Calculated written lines (readLines-skippedLines): 26659127

DWD Luftdruck Import Report
Count of read lines: 64127517

Total count of skipped lines: 40759385

Reasons for skipped lines:
Year is before 2009: 40616864
recent data and Year is before 2014: 142521

Count of written lines on SAP HANA: 23368132

Calculated written lines (readLines-skippedLines): 23368132

DWD Bewölkungsgrad Import Report
Count of read lines: 72479217

Total count of skipped lines: 63746710

Reasons for skipped lines:
Year is before 2009: 48304338
recent data and Year is before 2014: 142521
Air pressure is under 0: 15299851

Count of written lines on SAP HANA: 8732507

Calculated written lines (readLines-skippedLines): 8732507

DWD Bodentemperatur Import Report
Count of read lines: 200912650

Total count of skipped lines: 139355723

Reasons for skipped lines:
Measure Depth is over 10 or under 0: 15421293
Year is before 2009: 123482605
recent data and Year is before 2014: 451825

Count of written lines on SAP HANA: 61556927

Calculated written lines (readLines-skippedLines): 61556927

DWD Windgeschwindigkeit Import Report
Count of read lines: 99389149

Total count of skipped lines: 84968153

Reasons for skipped lines:
Wind speed is under 0: 19056
Wind direction is under 0 or over 360: 23847
Year is before 2009: 84847345
recent data and Year is before 2014: 77905

Count of written lines on SAP HANA: 14420996

Calculated written lines (readLines-skippedLines): 14420996

Entsoe Power Consumption Report
Count of read lines: 49500

Total count of skipped lines: 0

Count of written lines on SAP HANA: 49500

Calculated written lines (readLines-skippedLines): 49500

12.3 Ergebnisse der Vorhersagen

Dieses Kapitel befasst sich mit den Testergebnissen der verschiedenen Algorithmen zu den jeweiligen Hypothesen. Im folgenden Abschnitt werden zunächst die Prädiktionen der verschiedenen Algorithmen für die Datenbasis gebildet. Das bedeutet, zunächst wird der Energieverbrauch ausschließlich anhand der historischen Energieverbrauchsdaten vorhergesagt. Diese Vorhersagen bilden die Vergleichsbasis für die Überprüfung der verschiedenen Hypothesen.

12.3.1 Datenbasis

Ausgangslage für alle weiteren Hypothesen und den damit verbundenen Vorhersagen ist die Datenbasis. Die Datenbasis besteht aus den historischen Energieverbrauchsdaten der Entso-E vom 01.01.2009 bis zum 31.12.2013. Mit Hilfe dieser Daten soll der Stromverbrauch für den Zeitraum 01.01.2014 bis zum 31.01.2014 vorhergesagt werden. Mit dieser Datenbasis, bestehend also lediglich aus den Energieverbrauchsdaten, werden zunächst die verschiedenen von der SAP-PAL-Bibliothek angebotenen Algorithmen benutzt um den Energieverbrauch für den oben genannten Zeitraum vorherzusagen. Zunächst wird der Datenfluss der Rohdaten zu den Zieldaten erläutert. Die Ergebnisse zu den einzelnen Algorithmen werden im darauf folgenden Abschnitt dokumentiert.

Datenfluss In diesen Abschnitt wird beschrieben, wie die Quelldaten zu den Zieldaten verarbeitet werden. Die Quelltabelle ist die Tabelle `OLIMP.ENTOSE_POWER_CONSUMPTION`. Die Datenfelder der Tabelle sind in Tabelle 60 beschrieben.

Country	Timestamp	Consumption
DE	01.01.2009 00:00:00	44.835
DE	01.01.2009 01:00:00	46.380
...	...	...

Tabelle 60: Tabelle `OLIMP.ENTSOE_POWER_CONSUMPTION`

Die Spalte `Country` ist eine zweistellige String-Respräsentation des Landes, in dem der Stromverbrauch aufgezeichnet wurde. Die Spalte `Timestamp` ist der Zeitpunkt der Aufzeichnung. Die Spalte `Consumption` enthält die Stromverbrauchdaten als Integer-Repräsentation. Die Tabelle enthält stündliche Angaben zum Stromverbrauch in Deutschland für den Zeitraum vom 01.01.2009 bis zum Monat Juli 2014. Hierbei gilt zu beachten, dass es sich bei diesen Daten um Rohdaten handelt, so dass eine Vorverarbeitung erforderlich ist. Insbesondere muss an dieser Stelle gewährleistet werden, dass keine leeren Datensätze in den Daten vorhanden sind. Hierzu wird zunächst eine View erstellt, welche mit Hilfe der SAP HANA Tabelle `SYS_BI_M_TIME_DIMENSION` sicherstellt, dass für den

erwähnten Zeitraum jeweils ein stündlicher Eintrag existiert. Der SQL-Code für die erstellte View ist in Listing 24 zu finden.

```
1 CREATE VIEW "OLIMP"."VIEW_CONSUMPTION_ALL_DIRTY" ( "ID",
2             "CONSUMPTION" ) AS SELECT
3             (t.HOUR_COUNT-8783) as ID,
4             TO_INTEGER(c."Consumption") as CONSUMPTION
5 FROM "_SYS_BI"."M_TIME_DIMENSION" t
6 LEFT OUTER JOIN "OLIMP"."OLIMP::Entsoe_Power_Consumption" c ON c."TimeStamp"
      = t.DATETIMESTAMP
7 WHERE t.YEAR_INT >= 2009
8 AND t.YEAR_INT <= 2013
9 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 24: View `VIEW_CONSUMPTION_ALL_DIRTY`

Durch die Verwendung eines Left Outer Joins der beiden Tabellen `_SYS_BI.M_TIME_DIMENSION` auf die Tabelle `OLIMP.ENTSOE_POWER_CONSUMPTION` wird gewährleistet, dass stündliche Einträge für den Zeitraum vom 01.01.2009 00:00:00 Uhr bis zum 31.12.2013 23:00:00 Uhr existieren. Im nächsten Schritt werden nun fehlende Stromverbrauchsangaben ergänzt. Dies wird durch die Prozedur in Listing 25 realisiert.

Dabei hat die Prozedur die folgende Aufgabe: Um sicherzustellen, dass in der Datenspalte `Consumption` keine `NULL`-Werte vorliegen, durchsucht die Prozedur die gesamte Spalte nach `NULL`-Werten. Sobald ein solcher Wert vorliegt, wird diese Zeile mit dem arithmetischen Mittel des vorherigen Wertes und des nachfolgenden Wertes - ausgehend von der `NULL`-Zeile - gefüllt. Hierbei wird eine neue Tabelle mit den nun bereinigten Stromverbräuchen angelegt. Diese Tabelle hat den Namen `PRE.CONSUMPTION_ALL_CLEAN` und hat den gleichen Aufbau wie die View `VIEW_CONSUMPTION_ALL_DIRTY`, mit dem Unterschied, dass für jeden Zeitpunkt ein entsprechender Stromverbrauch vorliegt. Mit Hilfe dieser Tabelle kann nun wiederum eine View erstellt werden, die den Erfordernissen der verschiedenen SAP-PAL-Algorithmen entspricht. Grundsätzlich benötigt jeder in der SAP-PAL-Bibliothek implementierter Algorithmus mindestens die folgenden Datenfelder:

ID	**Variable X**	**Variablen N**
0	44.385	...
1	44.685	...
...	...	...

Tabelle 61: Erforderliche Datenfelder für die SAP-PAL-Algorithmen

Die Spalte `ID` repräsentiert dabei eine fortlaufende ID beginnend bei 0. Diese Vorgabe wird von allen in der SAP-PAL-Bibliothek implementieren Algorithmen vorausgesetzt. Dabei repräsentiert die ID mit dem Wert 0 den Zeitpunkt 01.01.2009, 00:00 Uhr. Die ID mit dem Wert 1 repräsentiert den Zeitpunkt 01.01.2009, 01:00 Uhr. Dieses Verhalten setzt sich entsprechend fort und repräsentiert die Zeit als fortlaufende Nummerierung. Die meisten

```
1  CREATE PROCEDURE OLIMP.CLEAN_ENTOSE_POWER_CONSUMPTION LANGUAGE SQLSCRIPT
2  AS
3          CURSOR c_cursor1 FOR SELECT * FROM "OLIMP"."
               VIEW_CONSUMPTION_ALL_DIRTY";
4          v_prev_consumption INTEGER;
5          v_next_consumption INTEGER;
6          v_arith INTEGER;
7  BEGIN
8          v_prev_consumption := 0;
9          OPEN c_cursor1();
10
11         call OLIMP.DROP_TABLE_IF_EXISTS('CONSUMPTION_ALL_CLEAN', 'PRE');
12         CREATE COLUMN TABLE PRE.CONSUMPTION_ALL_CLEAN(ID INTEGER NOT NULL,
               CONSUMPTION INTEGER NOT NULL);
13
14         IF c_cursor1::ISCLOSED
15         THEN
16                 CALL ins_msg_proc('WRONG: cursor not open');
17         ELSE
18                 FOR cur_row as c_cursor1 DO
19                         IF cur_row.CONSUMPTION IS NULL
20                         THEN
21                                 SELECT CONSUMPTION INTO v_next_consumption
                                       FROM "OLIMP"."VIEW_CONSUMPTION_ALL_DIRTY"
                                        WHERE ID > cur_row.ID AND CONSUMPTION IS
                                        NOT NULL LIMIT 1;
22
23                                 v_arith := (v_prev_consumption +
                                       v_next_consumption)/2;
24                                 CALL ins_msg_proc('Consumption for ID ' ||
                                       cur_row.ID || ' is null, prev: ' ||
                                       v_prev_consumption || ' next: ' ||
                                       v_next_consumption || ' Arith: ' ||
                                       v_arith);
25                                 INSERT INTO PRE.CONSUMPTION_ALL_CLEAN VALUES
                                       (cur_row.ID, v_arith);
26                         ELSE
27                                 INSERT INTO PRE.CONSUMPTION_ALL_CLEAN VALUES
                                       (cur_row.ID, cur_row.CONSUMPTION);
28                                 v_prev_consumption := cur_row.CONSUMPTION;
29                         END IF;
30                 END FOR;
31                 CLOSE c_cursor1;
32         END IF;
33 END
```

Abbildung 25: Bereinigung der Entsoe-Stromdaten

SAP-PAL-Algorithmen verzichten damit auf den Datentyp `Timestamp` als Repräsentation der Zeit. Dadurch wird auch die von vielen Algorithmen geforderte Stationarität von Trainingsdaten erzwungen. Die Spalte `Variable X` definiert die zu berechnende abhängige Variable. Die Spalte(n) `Variable N` (optional) definieren N unabhängige Variablen. Hierbei gilt zu beachten, dass diese Spalte(n) bei den Algorithmen

- mindestens einmalig vorhanden sein muss (z.B. Lineare Regression).

- nicht vorhanden sein darf (z. B. Arima-Zeitreihe).
- mehrfach vorhanden sein darf (z.B. exponentielle-, polynomiale Regression).

Mit diesen Informationen kann nun die entsprechende View erstellt werden. Listing 26 zeigt den SQL-Code für die erstellte View:

```
1 CREATE VIEW "OLIMP"."VIEW_CONSUMPTION" ( "ID",
2             "CONSUMPTION" ) AS SELECT
3               ID, TO_INTEGER(c."Consumption") as CONSUMPTION
4 FROM "PRE.CONSUMPTION_ALL_CLEAN"
5 WHERE ID <= 43824
6 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 26: View für die Trainingsdaten der Stromverbräuche

Die View bezieht generell die Daten aus der zuvor erzeugten Tabelle `PRE.CONSUMPTION_ALL_CLEAN`. Diese View ist mit den Stromverbrauchsdaten vom 01.01.2009 00:00:00 Uhr bis zum 31.12.2013 23:00:00 Uhr gefüllt, was einer fortlaufenden ID von 0 bis 43824 entspricht. Die daraus resultierenden Daten sind beispielhaft in Tabelle 62 dargestellt.

ID	**Consumption**
0	46719
1	42637
2	...

Tabelle 62: View „CONSUMPTION“

Für die zu prognostizierenden Daten wird ebenfalls eine View erstellt, welche die fortlaufende ID und den tatsächlichen Verbrauch ab dem 01.01.2014 00:00:00 enthält. Der ID-Zahlenbereich von 43825 bis 44569 entspricht den Zeitraum vom 01.01.2014 00:00 Uhr bis zum 31.01.2014 23:00 Uhr. Für diesen Zeitraum sollen entsprechende Prädiktionen durch die Algorithmen erstellt werden. Der SQL-Code für diese View ist in Listing 27 ersichtlich.

```
1 CREATE VIEW "OLIMP"."VIEW_CONSUMPTION_FORECAST" ( "ID",
2             "CONSUMPTION" ) AS SELECT
3               ID, TO_INTEGER(c."Consumption") as CONSUMPTION
4 FROM "PRE.CONSUMPTION_ALL_CLEAN"
5 WHERE ID >= 43825
6 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 27: View für die Testdaten der Stromverbräuche

Hiermit endet die Beschreibung des Datenflusses für die Datenbasis. Die nächsten Abschnitte befassen sich mit den Ergebnissen der Algorithmen, die auf die beschriebenen Views in Listing 26 (Trainingsdaten) und Listing 27 (Testdaten) angewendet wurden.

Arima - 1. Durchlauf Für den ersten Durchlauf werden die Parameter des Arima-Modells wie folgt gewählt: $P = 1, Q = 0, D = 0$ Mit diesen Parametern wird das Arima-Modell errechnet, welches anschließend die Grundlage für die Vorhersagen bildet. Das errechnete Modell ist in Tabelle 63 ersichtlich. Betrachtet man auf dieser Grundlage die erstellten Prädiktionswerte, so wird ersichtlich das die Modellbildung aufgrund der gewählten Parameter schlecht ist.

Parameter	Wert
ARParameters	0,961857
MAParameters	
d	0
Intercept	53968,7
Sigma2	7,49506e+06
logLikelihood	-409046
SeriesData	
DeltaSeriesData	44763
EPS	

Tabelle 63: Modellbildung des ersten Durchlaufes. Algorithmus: Arima

Grafisch wird dies in Abbildung 28 abgebildet. Dabei zeigt die y-Achse den Verbrauch in Megawatt und die x-Achse den gewählten Prädiktionzeitraum Januar 2014 auf stündlicher Basis. Diese Beschreibung gilt ebenfalls für die Abbildungen 29, 30 und 31. Aus der Grafik wird ersichtlich, dass die vorhergesagten Werte für den gesamten Zeitraum Januar 2014 konstant sind und somit keine nutzbaren Vorhersagen ermöglicht.

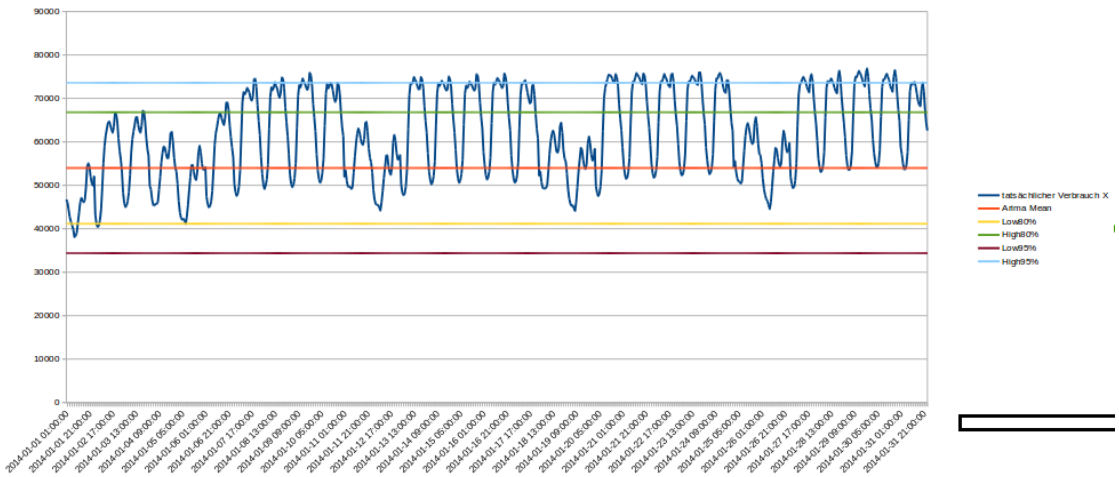


Abbildung 28: Grafische Darstellung der Vorhersagen

2.Durchlauf Für den zweiten Durchlauf werden die Parameter des Arima-Modells wie folgt gewählt: $P = 0, Q = 1, D = 0$ Mit diesen Parametern wird das Arima-Modell er-

rechnet, welches anschließend die Grundlage für die Vorhersagen bildet. Das errechnete Modell ist in Tabelle 64 ersichtlich.

Parameter	**Wert**
ARParameters	
MAParameters	0,911074
d	0
Intercept	53968,7
Sigma2	3,05629e+07
logLikelihood	-439844
SeriesData	
DeltaSeriesData	
EPS	-4512,21

Tabelle 64: Modellbildung des zweiten Durchlaufes. Algorithmus: Arima

Betrachtet man auf dieser Grundlage die erstellten Prädiktionswerte in Abbildung 29, so wird ersichtlich das die Modellbildung aufgrund der gewählten Parameter schlecht ist. Trotz des geänderten Parameters werden keine adäquaten Vorhersagen berechnet.

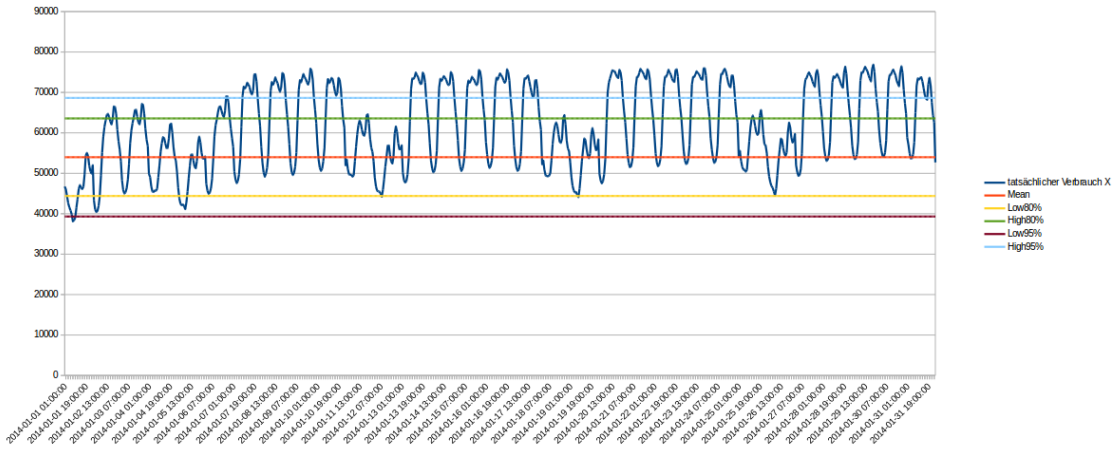


Abbildung 29: Grafische Darstellung des zweiten Durchlaufes

Arima - 3. Durchlauf Für den dritten Durchlauf werden die Parameter des Arima-Modells wie folgt gewählt: $P = 1, Q = 1, D = 0$ Mit diesen Parametern wird das Arima-Modell errechnet, welches anschließend die Grundlage für die Vorhersagen bildet. Das errechnete Modell ist in Tabelle 65 ersichtlich.

Betrachtet man auf dieser Grundlage die erstellten Prädiktionswerte in Abbildung 30, so wird ersichtlich das die Modellbildung aufgrund der gewählten Parameter schlecht ist. Trotz der geänderten Parameter des Modells werden keine adäquaten Vorhersagen berechnet.

Parameter	Wert
ARParameters	0,94016
MAParameters	0,509876
d	0
Intercept	53968,7
Sigma2	5,23952e+06
logLikelihood	-401201
SeriesData	
DeltaSeriesData	44763
EPS	49,2691

Tabelle 65: Modellbildung des dritten Durchlaufes. Algorithmus: Arima

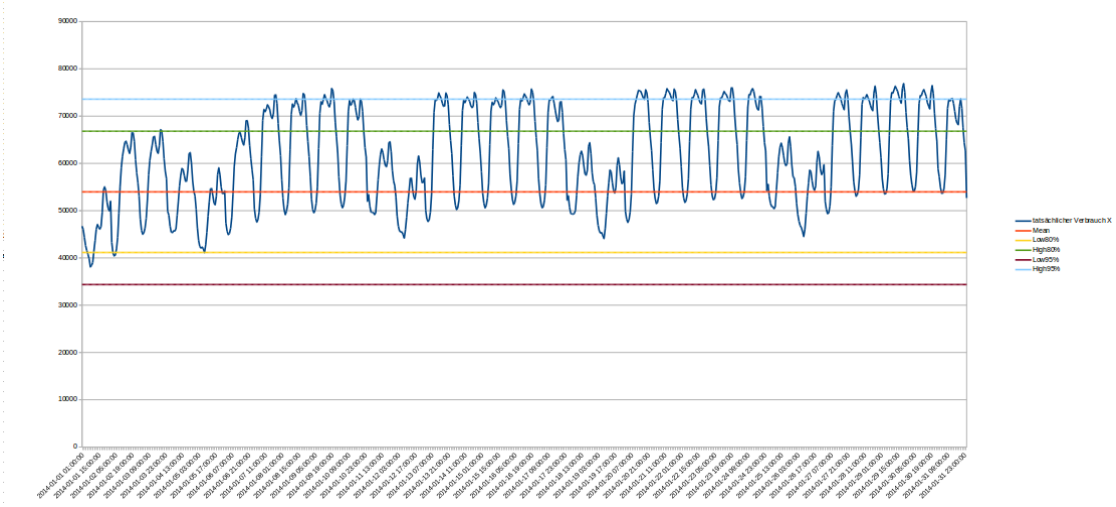


Abbildung 30: Grafische Darstellung des dritten Durchlaufes

Arima - 4. Durchlauf Für den vierten Durchlauf wurden die Parameter des Arima-Modells wie folgt gewählt: $P = 2, Q = 1, D = 0$ Mit diesen Parametern wird das Arima-Modell errechnet, welches anschließend die Grundlage für die Vorhersagen bildet. Das errechnete Modell ist in Tabelle 66 ersichtlich.

Parameter	Wert
ARParameters	1,66753;-0,727652
MAParameters	-0,136124
d	0
Intercept	53968,7
Sigma2	4,26622e+06
logLikelihood	-396683
SeriesData	
DeltaSeriesData	44547;44763
EPS	706,426

Tabelle 66: Modellbildung des vierten Durchlaufes. Algorithmus: Arima

Betrachtet man auf dieser Grundlage die erstellten Vorhersagewerte in Abbildung 31, so wird ersichtlich das die Modellbildung aufgrund der gewählten Parameter schlecht ist. Trotz der geänderten Parameter des Modells werden keine adäquaten Vorhersagen berechnet.

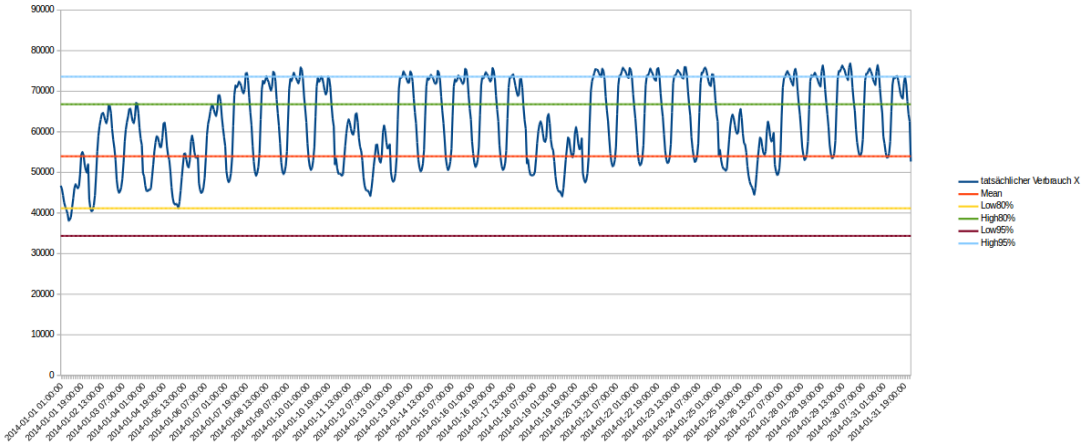


Abbildung 31: Grafische Darstellung des vierten Durchlaufes

Aufgrund der schlechten Ergebnisse der beschriebenen Durchläufe wird das Arima-Modell nicht weiter verfolgt. Neben den schlechten Ergebnissen konnten einige Arima-Konfigurationen auf dem SAP HANA System aufgrund nicht identifizierbarer Fehler durchgeführt werden. Die fehlgeschlagenen Konfigurationen und der Grund hierfür sind in Tabelle 67 zu finden.

Arima-Parameter	**Grund**
$P = 0, Q = 2, D = 1$	Interner SAP HANA Fehler
$P = 0, Q = 2, D = 1$	Interner SAP HANA Fehler
$P = 0, Q = 2, D = 1$	Interner SAP HANA Fehler
$P = 0, Q = 0, D = 1$	Keine Vorhersagewerte erstellt
$P = 0, Q = 1, D = 1$	Keine Vorhersagewerte erstellt

Tabelle 67: Nicht durchführbare Arima-Konfigurationen

Ein weiterer Grund das Arima-Modell nicht weiter zu verfolgen, ist die Tatsache, dass der Algorithmus innerhalb von SAP HANA lediglich eine Variable zur Berechnung von Prognosen zulässt. Das bedeutet: Die Berechnung des zukünftigen Energieverbrauches ist ausschließlich anhand des historischen Energieverbrauches möglich. Kommen jedoch weitere Variablen hinzu, zum Beispiel Temperatur- oder Winddaten, so ist dieser Algorithmus für diese Berechnungen nicht geeignet.

Single Exponential Smoothing Mit Hilfe des Single Exponential Smoothing-Algorithmus konnte keine adäquate Vorhersage der Verbrauchsdaten generiert werden. Grund hierfür

ist die Eigenschaft dieser Algorithmengruppe, die neueren Daten eine höhere Priorität zurechnet als älteren Daten. Daraus folgt, dass Entwicklungen zum Ende einer Betrachtungsperiode bis in ein positives oder negatives Extrem vorgesetzt werden (vgl. Kapitel 11). Ein unendlich, positiver Energieverbrauch ist jedoch genauso wenig wie ein unendlich negativer Energieverbrauch zu erwarten. Die Validität eines negativen Energieverbrauchs ist gänzlich in Frage zu stellen. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/SingleExpSmoothing` zu finden.

Double Exponential Smoothing Die zuvor beschriebene Problematik des Single Exponential Smoothing trifft auch beim Double Exponential Smoothing zu. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/DoubleExpSmoothing` zu finden.

Triple Exponential Smoothing Die Problematik des Single Exponential Smoothing und Double Exponential Smoothing trifft ebenfalls beim Double Triple Smoothing zu. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/TripExpSmoothing` zu finden.

Forecast Smoothing Forecast Smoothing wird von der SAP-PAL-Bibliothek angeboten, um die optimalen Parameter für die Algorithmen Double Exponential Smoothing, Triple Exponential Smoothing und Forecast Smoothing automatisch zu errechnen [SAP14b]. Aufgrund der berechneten Parameter wird anschließend der am besten passende Algorithmus für die zugrunde liegende Datenmenge automatisch ausgewählt und das Modell anschließend errechnet. Trotz dieser Vorteile kann mit diesem Algorithmus dennoch keine adäquate Vorhersage der Verbrauchsdaten errechnet werden. Die Eigenschaften der zuvor beschriebenen Algorithmen (Single-, Double-, und Triple Exponential Smoothing) treffen auch auf diesen Algorithmus zu. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/ForecastSmoothing` zu finden.

BI Variate Natural Logarithmic Regression Die Vorhersage des Stromverbrauches gibt für den gesamten Betrachtungszeitraum (Januar 2014) konstante Werte aus. Wenn die Vorhersage für einen größeren Zeitraum (z.B bis 2017) durchgeführt wird, vergrößert sich die Vorhersage für den Stromverbrauch allerdings ist die Steigung hier sehr klein. Fazit: Der Algorithmus liefert eine bis unendliche Vorhersage, wobei die Steigerung so klein ist, dass die Vorhersage bis für Januar 2014 konstant bleibt. Der Algorithmus liefert daher keine zuverlässige Vorhersagen. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/BI-Variate_NaturalLog` zu finden.

BI Variate Geometric Regression Die Vorhersage des Stromverbrauches gibt für den gesamten Betrachtungszeitraum (Januar 2014) konstante Werte aus. Wenn die Vorhersage für einen größeren Zeitraum (z.B bis 2017) durchgeführt wird, vergrößert sich die Vorhersage für den Stromverbrauch, allerdings ist diese Steigung sehr klein. Fazit: Der Algorithmus liefert eine bis unendliche Vorhersage, wobei die Steigerung so klein ist, dass die Vorhersage für Januar 2014 konstant Werte liefert. Der Algorithmus liefert daher keine zuverlässige Vorhersagen. Die Ergebnisse zu den verschiedenen Durchläufen sind auf der beigefügten CD im Ordner `Forecast/Basis/BI-Variate-GeometrischeRegression` zu finden.

Multiple Lineare Regression Mit Hilfe von Multiple Regression wurde der Stromverbrauch für den Zeitraum Januar 2014 vorhergesagt. Hierzu wird zunächst die `ID` als Prädiktor verwendet. Das heißt, die Daten enthalten nur Verbrauchdaten über die Zeit und keine weiteren Prädikatoren, die den Verbrauch beeinflussen könnten. Auch hier hat das Modell keine adäquaten Vorhersagen generiert. Eine multiple lineare Regression ist nur dann sinnvoll, wenn die Prädiktoren Xn mit dem Kriterium Y korreliert sind. Die Ergebnisse zu den verschiedenen Durchläufen sind unter dem Ordner „Forecast/Basis/MultipleLineareRegression“ zu finden.

Polynomiale Regression Mit Hilfe der polynomialen Regression werden die Daten mit einer Polynomfunktion modelliert. Die unbekannten Modellparameter werden dabei aus den Daten geschätzt [SAP14b]. Dieses Modell hat ebenfalls kein adäquates Ergebnis erzeugt. Wie bei der multiplen linearen Regression ist eine polynomiale Regression nur dann sinnvoll, wenn die Prädikatoren Xn mit dem Kriterium Y korreliert sind. Die Ergebnisse zu den verschiedenen Durchläufen sind im Ordner „Forecast/Basis/Polynomialregression“ zu finden.

Exponentielle Regression Mit Hilfe der exponentiellen Regression wird die Beziehung zwischen Verbrauch und ID dargestellt. In diesen Versuch werden die Daten mit der Exponentialfunktionen modelliert. Das Ergebnis, das dieses Modell erzeugt ist ebenfalls nicht signifikant. Die zuvor beschriebene Problematik liegt auch bei diesen Algorithmus vor. Die Ergebnisse zu den verschiedenen Durchläufen sind unter dem Ordner `Forecast/Basis/Exponentialregression` zu finden.

Lineare Regression mit gedämpftem Trend und saisonaler Anpassung Seit dem SAP HANA Service-Pack 09 steht der Algorithmus Lineare Regression mit gedämpftem Trend und saisonaler Anpassung zur Verfügung (Siehe Abschnitt 11.1.1). Der Algorithmus berechnet ähnlich der Arima-Zeitreihenanalyse Prognosen anhand einer Variable. Dies ist der

historische Stromverbrauch. Um mit diesen Algorithmus Prognosen zu errechnen werden die Parameter des Algorithmus entsprechend Tabelle 68 wie festgelegt:

Parameter	**Einstellung**	**Erläuterung**
FORECAST_LENGTH	745	Die Anzahl der Vorhersagen.
TREND	0.9	Einfluss des gedämpften Trends.
SEASONALITY	1	Legt fest, dass die historischen Daten saisonelle Einflüsse haben.
PERIODS	8760	Legt die Periodizität fest (Hier: Jede Stunde ist eine Periode (24x365)).

Tabelle 68: Parametereinstellungen der linearen Regression mit gedämpftem Trend

Die View `VIEW_CONSUMPTION_Training` enthält die Trainingsdaten, mit denen das Modell gebildet wird. Um das Trainingsset in die Modellbildung des Algorithmus einzubinden wird das in Listing 32 abgebildete SQL-Statement verwendet.

```
1 CREATE COLUMN TABLE PAL_FORECASTSLR_DATA_TBL LIKE PAL_FORECASTSLR_DATA_T;
2 INSERT INTO PAL_FORECASTSLR_DATA_TBL select "ID", "CONSUMPTION" FROM "PAL".
      "CONSUMPTION_TRAINING";
```

Abbildung 32: SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung

Die Tabelle 69 zeigt die produzierten Fehlerkennzahlen des Algorithmus. Dabei bezieht sich der Durchlauf „Jan2014“ auf die Verwendung des gesamten Trainingsdatensatzes. Der Durchlauf „Wo2014“ bezieht lediglich die Arbeitstage der Trainingsdaten in die Modellbildung mit ein und berechnet entsprechend nur die Arbeitstage des Zeitraums Januar 2014. Der Durchlauf „WoE2014“ bezieht lediglich die Wochenenden in die Modellbildung mit ein und prognostiziert dementsprechend nur die Wochenenden für den Zeitraum Januar 2014. Im Durchlauf „Zusammen“ werden die beiden Durchläufe „Wo2014“ und „WoE2014“ zusammengefügt. Die hierzu verwendeten SQL-Codes, die das Traingsdatenset einschränken sind in Listing 33 und 34 zu finden. Prinzipiell sind die beiden Statements identisch zu Listing 32, jedoch werden die Daten aus einer zuvor vorbereiteten View bezogen, die jeweils nur Trainingsdaten für die Arbeitstage beziehungsweise Wochenenden der Trainingsdaten von 2009 bis 2013 enthält.

```
1 CREATE COLUMN TABLE PAL_FORECASTSLR_DATA_TBL LIKE PAL_FORECASTSLR_DATA_T;
2 INSERT INTO PAL_FORECASTSLR_DATA_TBL select "ID", "CONSUMPTION" FROM "PAL".
      "CONSUMPTION_TRAINING_WEEK";
```

Abbildung 33: SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung für Wochenarbeitstage

Die Abbildung 35 zeigt die grafische Darstellung der Vorhersagen für den Durchlauf

```
1 CREATE COLUMN TABLE PAL_FORECASTSLR_DATA_TBL LIKE PAL_FORECASTSLR_DATA_T;
2 INSERT INTO PAL_FORECASTSLR_DATA_TBL select "ID", "CONSUMPTION" FROM "PAL".
      "CONSUMPTION_TRAINING_WEEKEND";
```

Abbildung 34: SQL-Statement für die Modellbildung der linearen Regression mit gedämpftem Trend und saisonaler Anpassung für Wochenenden

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Jan2014	0,46	7911,11	9184,79	0,12	0,14
Wo2014	0,78	6343,98	7182,53	0,09	0,11
WoE2014	0,9	6579,23	6846,36	0,12	0,12
Zusammen	0,82	6404,69	7097,30	0,10	0,11

Tabelle 69: Fehlerkennzahlen der Anwendung des Modells

„Jan2014“. Die Abbildung 36 zeigt die grafische Darstellung des Durchlaufes „Zusammen“.

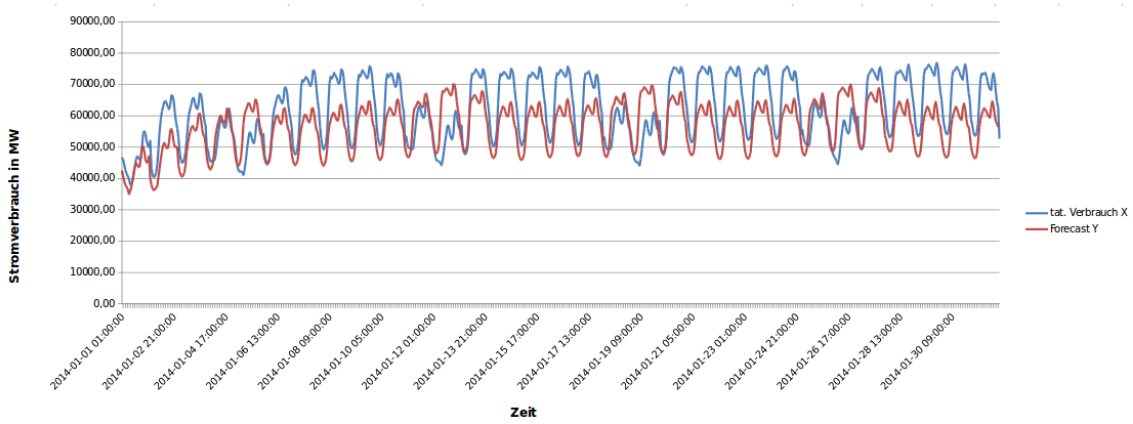


Abbildung 35: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Jan2014“

Betrachtet man die grafische Darstellung des Durchlaufes „Jan2014“, wird deutlich, dass der Verlauf der einzelnen Tage mehr oder weniger gut bis mittelmäßig ersichtlich wird. Weiterhin fällt auf, dass die Stromverbrauchsschwankungen innerhalb eines Tages dargestellt werden. Insbesondere am vierten Prognosetag entsteht eine an den tatsächlichen Verbrauch sehr angepasste Vorhersage. Es fällt jedoch auf, dass an den anderen Tagen ein „Sockelbetrag“ fehlt, welcher die Vorhersagewerte anheben und damit die Prognoseergebnisse weiter an den tatsächlichen Verbrauch anpassen könnte. Dieser Trend verstärkt sich, je länger der Prognosehorizont in der Zukunft liegt. Ebenfalls fällt auf, dass der Algorithmus die Auswirkungen des Wochenendes offenbar nicht stark genug berücksichtigt. Hier sind - im Vergleich zum tatsächlichen Verbrauch an den Wochenenden - starke Abweichungen in der Grafik ersichtlich. Die Beobachtungen werden auch an den Fehlerkennzahlen des Algorithmus deutlich. Der R-Squared liegt hier bei 0,46, was einer schlech-

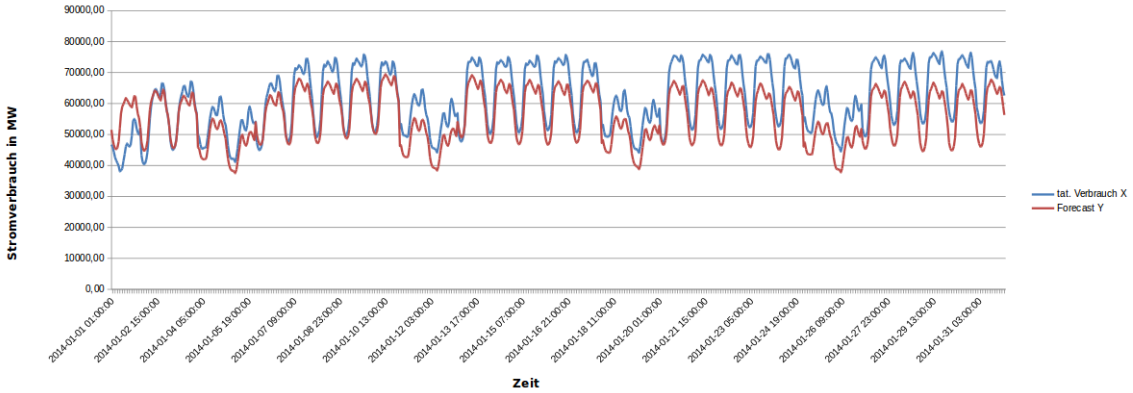


Abbildung 36: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Zusammen“

ten bis mittelmäßigen Anpassung des Algorithmus an die tatsächlichen Daten entspricht. Der MAE dieses Durchlaufes liegt bei 7911,11 und der RMSE liegt bei 9184,79 Punkten. Die hierzu errechneten relativen Fehlerkennzahlen sind ebenfalls in der Tabelle 69 ersichtlich. Um die Vorhersagen zu verbessern, wurden zusätzlich die Durchläufe „Wo2014“ und „WoE2014“ durchgeführt. Der Durchlauf „Wo2014“ bildet ein Modell auf Basis der Wochentage Montag bis Freitag der Trainingsdaten und errechnet entsprechend Prognosen für diese Tage des Monats Januar 2014. Analog hierzu bildet der Durchlauf „WoE2014“ ein Modell auf Basis der Tage Samstag und Sonntag der Trainingsdaten und berechnet entsprechend Vorhersagewerte für diese Tage im Zeitraum Januar 2014. Im Durchlaufes „Zusammen“ werden die Ergebnisse der beiden Durchläufe zusammengefügt. Betrachtet man zunächst die Fehlerkennzahlen der Durchläufe „Wo2014“ und „WoE2014“ wird deutlich, dass mit einer getrennten Betrachtung von Wochenenden und Arbeitswochen eine Verbesserung der Prognoseergebnisse erzielt werden kann. Hier liegt der R-Squared bei 0,78 (Durchlauf „Wo2014“) und 0,9 (Durchlauf „WoE2014“), was einer guten Anpassung der Prognoseergebnisse an die tatsächlichen Daten entspricht. Dieses Verhalten wird auch an den Fehlerkennzahlen MAE und RMSE ersichtlich: Die beiden Durchläufe „Wo2014“ und „WoE2014“ erzielen bei beiden Fehlerkennzahlen niedrigere Werte als der Durchlauf „Jan2014“. Im Durchlauf „Zusammen“ werden die beiden Ergebnisse der vorigen Durchläufe zusammengefügt. Vergleicht man die grafische Abbildung 36 zu diesem Durchlauf mit der Abbildung 69 wird deutlich, dass nun eine bessere Anpassung des Algorithmus an die tatsächlichen Daten erfolgt. Insbesondere werden die Wochenenden nun besser erkannt und prognostiziert. Insgesamt fehlt jedoch immer noch ein „Sockelbetrag“ von etwa 5000 - 6000 Megawatt (MW), welcher die gesamte Vorhersage nochmals verbessern könnte. Die Fehlerkennzahlen des Durchlaufes „Zusammen“ sind hier wie folgt: der R-Squared liegt bei 0,82, was einer guten Anpassung der von dem Algorithmus produzierten Daten an die tatsächlichen Daten entspricht. Ebenso liegt der MAE mit 6404,69 und der RMSE mit 7097,30 Punkten um 1506,42 (MAE) und 2087,49 (RMSE) niedriger,

verglichen mit den Durchlauf „Jan2014“. Verglichen mit den Fehlerkennzahlen des Durchlaufes „Jan2014“ konnte hier also eine signifikante Verbesserung des Algorithmus durch die getrennte Betrachtung von Arbeits- und Wochenendtagen erzielt werden.

Zusammenfassung der Ergebnisse Die durchweg schlechten Ergebnisse der Algorithmen[10] auf die Views in Listing 26 und 27 haben gezeigt, dass die Vorhersage des Stromverbrauches anhand historischer Stromverbrauchsdaten und einer fortlaufenden ID keine adäquaten Ergebnisse für die betrachteten Algorithmen liefert. In diesem Kontext sind insbesondere die Ergebnisse der Arima-Zeitreihenanalyse enttäuschend: Dieser Algorithmus akzeptiert neben der ID und dem Stromverbrauch keine weiteren Eingabeprädikatoren. In diesem Fall kann lediglich eine Justierung der Parameter des Modells Abhilfe schaffen, was jedoch aufgrund mehrerer SAP HANA Studio-Fehler - wie oben beschrieben - zu keine weiteren Ergebnisse führt. Ebenfalls führen die Testläufe der Algorithmen lineare, exponentielle- oder polynomiale Regression zu keine adäquaten Ergebnisse. Wahrscheinlich sorgt in diesem Fall die fortlaufende ID der Datensätze für eine zu starke Glättung der Prediktionswerte. In Abbildung 37 wird ein solcher Kurvenverlauf für die polynomiale Regression stellvertretend für alle weiteren Algorithmen für die zu starke Glättung gezeigt.

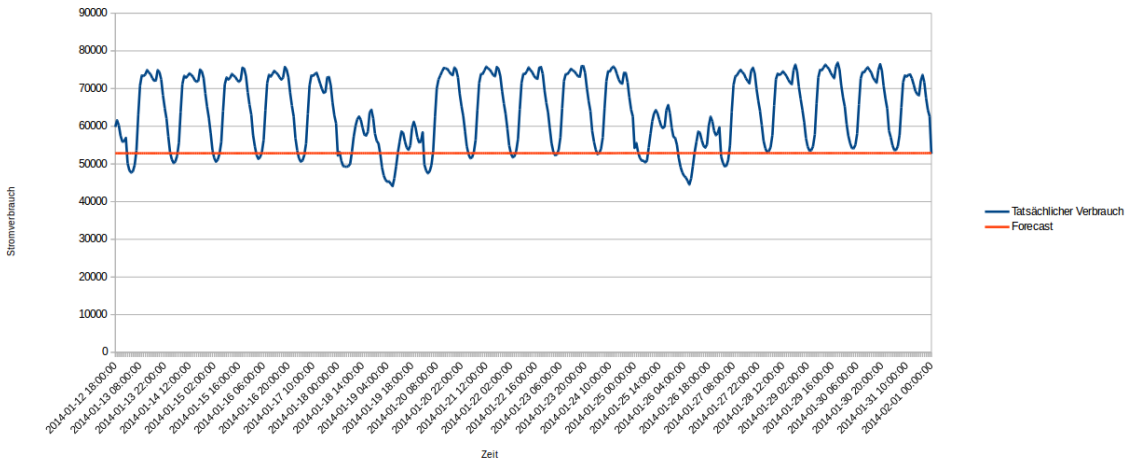


Abbildung 37: Grafische Darstellung der Vorhersage mit der polynomialen Regression

Der Kurvenverlauf in der Abbildung zeigt die zu starke Glättung des polynomialen Regressionsmodells. Die Vorhersage produziert zwar verschiedene Werte, jedoch unterscheiden sich diese untereinander kaum. Der R-Squared beträgt für dieses Modell lediglich 0,11, was einer sehr schlechten Anpassung der prognostizierten Daten an die tatsächlichen Daten entspricht. In diesem Durchlauf kann lediglich die lineare Regression mit gedämpftem Trend und saisonaler Anpassung überzeugen. Hier erzielt der Durchlauf „Zusammen“, in dem Arbeitstage und Wochenenden bei der Modellerzeugung getrennt betrachtet werden,

[10] Bis auf die lineare Regression mit gedämpftem Trend und saisonaler Anpassung.

die besten Ergebnisse. Dieser Durchlauf erzielt folgende Fehlerkennzahlen: Der R-Squared beträgt 0,82, der MAE liegt bei 6404,69 und der RMSE liegt bei 7097,30 Punkten, was einer relativen Abweichung zu den tatsächlichen Daten von 11 % entspricht. Dieser Algorithmus erlaubt jedoch nur die Berechnung von Vorhersagen anhand einer Variable, weshalb er in den folgenden Ausführungen nicht weiter betrachtet werden kann. Nach Abschluss dieses ersten Durchlaufes haben sich die Mitglieder der Projektgruppe Gedanken darüber gemacht, wie die Qualität der Vorhersagen verbessert werden kann. Hierbei gilt zu beachten, dass für die Datenbasis weiterhin ausschließlich der Stromverbrauch anhand des Stromverbrauches sowie des zeitlichen Verlaufes vorhergesagt werden soll. Daraus folgt, dass der zeitliche Bezug der Stromverbrauchsdaten differenzierter dargestellt werden muss. Die Realisierung dieses Ansatzes wird im folgenden Abschnitt beschrieben.

12.3.2 Zweiter Versuch der Datenbasis

Zentrale Feststellung und Schlussfolgerung aus den ersten Versuch ist, dass die verwendeten Algorithmen[11] keine adäquaten Prädiktionen des Stromverbrauches nur auf Basis einer fortlaufenden ID sowie des historischen Stromverbrauches selbst errechnen können. In diesem zweiten Versuch wird dargestellt, wie der Verlauf der Zeit differenzierter dargestellt werden kann. Hierzu wird wie folgt vorgegangen: Neben der fortlaufenden ID (Voraussetzung für jeden SAP-PAL-Algorithmus) werden weitere Spalten hinzugefügt, welche die Stromverbrauchsdaten mit folgenden weiteren Daten anreichern:

1. die fortlaufende Nummerierung der Stunden über den gesamten Datensatz
2. die jeweilige Stunde des Tages
3. den jeweiligen Tag der Woche
4. den jeweiligen Tag des Monats
5. der aktuelle Monat
6. das jeweilige Quartal
7. das jeweilige Jahr

Mit diesen zusätzlichen Informationen sollen nun neue Vorhersagen getroffen werden. Die oben genannten zusätzlichen Informationen werden in separaten Spalten gespeichert. Dadurch ergibt sich, das fortan nur noch Algorithmen zur Auswahl stehen, die N unabhängige Variablen als Input zulassen. Damit beschränkt sich die Auswahl der Algorithmen für alle weiteren Versuche auf die Folgenden:

- Multiple Lineare Regression
- Exponentielle Regression
- Support Vector Machine

[11] Bis auf die lineare Regression mit gedämpftem Trend und saisonaler Anpassung.

Datenfluss Dieser Abschnitt beschreibt den Datenfluss für den zweiten Versuch. Die Datenbasis entspricht dabei der Basis des ersten Versuches. Das heißt, die Datentabelle `OLIMP.ENTOSE_POWER_CONSUMPTION` (siehe Tabelle 60 sowie die auf dieser Tabelle ausgeführte Prozedur aus Listing 25 gilt ebenfalls für diesen Versuch. Auf dieser Datenbasis aufbauend wird anschließend jeweils eine View erstellt, welche die Trainingsdaten (Zeitraum 01.01.2009 00:00:00 Uhr bis 31.12.2013 23:00:00 Uhr) sowie die Testdaten (Zeitraum 01.01.2014 00:00:00 Uhr bis 31.01.2014 00:00:00 Uhr) beinhalten. Die SQL-Codes hierzu sind in Listing 38 und 39 zu finden.

```
1  CREATE VIEW "PAL"."CONSUMPTION_TRAINING" ( "ID",
2              "CONSUMPTION",
3              "HOUR_COUNT",
4              "HOUR_OF_DAY",
5              "DAY_OF_WEEK",
6              "DAY_OF_MONTH",
7              "MONTH",
8              "QUARTER",
9              "YEAR" ) AS SELECT
10             c.ID-1 as ID,
11             c.CONSUMPTION as CONSUMPTION,
12             c.ID as HOUR_COUNT,
13             t.HOUR_INT as HOUR_OF_DAY,
14             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
15             t.DAY_INT as DAY_OF_MONTH,
16             t.MONTH_INT as "MONTH",
17             t.QUARTER_INT as "QUARTER",
18             t.YEAR_INT as "YEAR"
19 FROM "PAL"."CONSUMPTION_ALL_CLEAN" as c JOIN "_SYS_BI"."M_TIME_DIMENSION" as
       t ON c.ID+8784 = t."HOUR_COUNT"
20 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 38: View für die Trainingsdaten der Stromverbräuche mit zusätzlichen Zeitangaben

Wie bereits im ersten Versuch beschrieben, wird mit der `Timestamp`-Repräsentation der Zeit in der Tabelle `OLIMP.ENTOSE_POWER_CONSUMPTION` und der SAP HANA Tabelle `_SYS_BI.M_TIME_DIMENSION` die Repräsentation der Zeit in eine fortlaufende ID umgerechnet. Zusätzlich werden diese Views mit den am Anfang dieses Abschnittes erläuterten zusätzlichen Informationen zur Zeit angereichert. Diese Angaben stammen ebenfalls aus der SAP HANA Tabelle `_SYS_BI.M_TIME_DIMENSION`. Tabelle 70 und 71 zeigen hierzu einige Beispieldaten.

Tabelle 70 zeigt die ersten drei Datensätze des Trainingssets und Tabelle 71 zeigt die ersten drei Datensätze des Testsets. Die Spaltennamen sind dabei identisch: Spalte `ID` ist die fortlaufende ID der Datensätze. Spalte `CONSUMPTION` [12] beinhaltet den stündlichen Stromverbrauch. Spalte `HOUR_COUNT` ist die fortlaufende Nummerierung der Stunden über den gesamten Datensatz. Die Spalte `HOUR_OF_DAY` beinhaltet die jeweilige Stunden des Daten-

[12] nur im Trainingsset vorhanden, da genau diese Werte im Testset vorhergesagt werden.

```
1  CREATE VIEW "PAL"."CONSUMPTION_FORECAST" ( "ID",
2              "HOUR_COUNT",
3              "HOUR_OF_DAY",
4              "DAY_OF_WEEK",
5              "DAY_OF_MONTH",
6              "MONTH",
7              "QUARTER",
8              "YEAR" ) AS SELECT
9              t.HOUR_COUNT-8783-43824-1 as ID,
10             t.HOUR_COUNT-8783 as HOUR_COUNT,
11             t.HOUR_INT as HOUR_OF_DAY,
12             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
13             t.DAY_INT as DAY_OF_MONTH,
14             t.MONTH_INT as "MONTH",
15             t.QUARTER_INT as "QUARTER",
16             t.YEAR_INT as "YEAR"
17 FROM "_SYS_BI"."M_TIME_DIMENSION" as t
18 WHERE t.DATETIMESTAMP >= '2014-01-01 00:00:00'
19 AND t.DATETIMESTAMP <= '2014-01-31 24:00:00'
20 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 39: View für die Testdaten der Stromverbräuche mit zusätzlichen Zeitangaben

satzes. Der Wertebereich dieser Spalte liegt zwischen 0 (00:00 Uhr) und 23 (23:00 Uhr). Die Spalte `DAY_OF_WEEK` beinhaltet den aktuellen Tag der Woche als Integer-Representation. Der Wertebereich liegt hier zwischen 0 und 6, wobei der Wert 0 für Montag steht und der Wert 6 für Sonntag steht. Die Spalte `DAY_OF_MONTH` zeigt den aktuellen Tag des Monats an, hier sind also Werte zwischen 1 und 31 möglich. Die Spalte `MONTH` zeigt den jeweiligen Monat des Datensatzes. Hier steht der Wert 1 für Januar und der Wert 12 für Dezember. Die Spalte `QUARTER` zeigt das Quartal des Datensatzes an. Der Wertebereich dieser Spalte liegt zwischen 1 und 4. Die Spalte `YEAR` zeigt das Jahr des Datensatzes an. Die Werte in dieser Spalte liegen zwischen 2009 und 2013 (VIEW_CONSUMPTION_TRAINING) und 2014 (VIEW_CONSUMPTION_FORECAST).

ID	CONSUMPTION	HOUR_COUNT	HOUR_OF_DAY	DAY_OF_WEEK	DAY_OF_MONTH	MONTH	QUARTER	YEAR
0	44.385	1	1	3	1	1	1	2009
1	46.380	2	2	3	1	1	1	2009
2	44.768	3	3	3	1	1	1	2009

Tabelle 70: View `PAL.CONSUMPTION.TRAINING`

ID	HOUR_COUNT	HOUR_OF_DAY	DAY_OF_WEEK	DAY_OF_MONTH	MONTH	QUARTER	YEAR
0	43825	0	2	1	1	1	2014
1	43826	1	2	1	1	1	2014
2	43827	2	2	1	1	1	2014

Tabelle 71: View `PAL.CONSUMPTION.FORECAST`

Multiple Lineare Regression Der Versuch der Vorhersage mit Hilfe der Linearen Regression teilt sich in zwei Teilversuche auf. Mit Hilfe der View `CONSUMPTION_TRAINING` (siehe Listing 38) werden zwei Modelle gebildet. Der erste Teilversuch bezieht sich dabei auf die gesamten Trainingsdaten. Der zweite Teilversuch bezieht sich auf das Jahr 2013 als Trainingsdaten. Die Scripte hierzu können auf der beigelegten CD im Ordner `svn/Forecast/MultipleLineareRegressionSQL/` gefunden werden. Im Ordner `2.Versuch` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz. Im Ordner `3.Versuch` befinden sich die Daten zur Modell- und Prädiktionsbildung für das Jahr 2013. In diesen beiden Ordnern sind folgende Dateien vorhanden:

Datei	Inhalt
build_modell_mlr.sql	Script zum Erstellen des Modells.
build_forecast_mlr.sql	Script zum Erstellen der Vorhersage für Januar 2014.
fehlerkennzahlen.csv	Vom Algorithmus erstellen Fehlerkennzahlen, bezogen auf die Trainingsdaten.
forecast.csv	Vorhersagewerte für Januar 2014.
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014.

Tabelle 72: Relevante Dateien für die Multiple Lineare Regression

Um das Trainingsset für die jeweiligen Durchläufe einzuschränken, wird in der Datei `build_modell_lmr.sql` das folgende SQL-Statement verwendet:

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT * FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 40: SQL-Statement für den gesamten Trainingsdatensatz (2009-2013)

Listing 40 zeigt, wie der gesamte Trainingsdatensatz in die Modellbildung eingebunden wird. Listing 41 zeigt, wie lediglich das Jahr 2013 des Trainingsdatensatzes in die Modell-

bildung eingebunden wird. Dies wird über die Spalte `ID` realisiert. Hierbei entspricht die ID mit dem Wert 35064 das Datum 01.01.2013, 00:00:00 Uhr.

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT * FROM "PAL"."CONSUMPTION_TRAINING"
    where "ID" >= 35064;
```

Abbildung 41: SQL-Statement für das Jahr 2013 als Trainingsdatensatz

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden muss folgendermaßen vorgegangen werden: Zunächst muss das Script `build_modell_mlr.sql` in SAP HANA ausgeführt werden. Anschließend wird der SQL-Code `build_forecast_mlr.sql` in SAP HANA ausgeführt. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und Übersichtsgründen wurden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- beziehungsweise Microsoft-Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Vorhersage. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten. Die Parametereinstellungen werden für die folgenden Durchläufe wie folgt gewählt:

Parameter	**Einstellung**	**Erläuterung**
THREAD_NUMBER	8	Modellbildung wird mit 8 Threads durchgeführt.
PMML_EXPORT	1	Gibt an, dass das Modell im PMML-Format vorliegt.
ADJUSTED_R2	1	R-Squared und R-Squared-Adjusted werden berechnet.
VARIABLE_SELECTION	0	Alle in der View vorhandenen Variablen werden zur Modellbildung einbezogen.

Tabelle 73: Parametereinstellungen der multiplen linearen Regression

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/MultipleLineareRegressionSQL/2.Versuch` und `svn/Forecast/MultipleLineareRegressionSQL/3.Versuch`. Wie bereits weiter oben erwähnt finden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Die Fehlerkennzahlen zu beiden Vorhersagen sind in Tabelle 74 zusammengefasst dargestellt. Dabei bezieht sich der Durchlauf „2009_2013“ auf die Verwendung der kompletten Trainingsdaten (2009-2013) für die Modellbildung. Der Durchlauf „2013“ bezieht sich auf die Verwendung der Trainingsdaten für das Jahr 2013.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
2009_2013	0,39	8229,84	10049,13	0,13	0,16
2013	0,40	11003,34	12877,54	0,17	0,20

Tabelle 74: Fehlerkennzahlen der Multiplen Linearen Regression

Abbildung 42 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten.

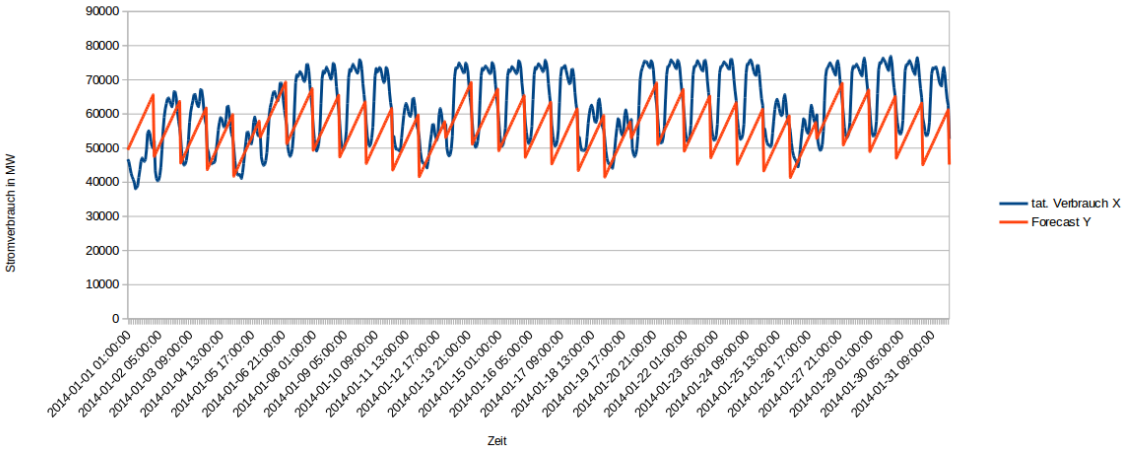


Abbildung 42: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2009_2013“

Abbildung 43 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten.

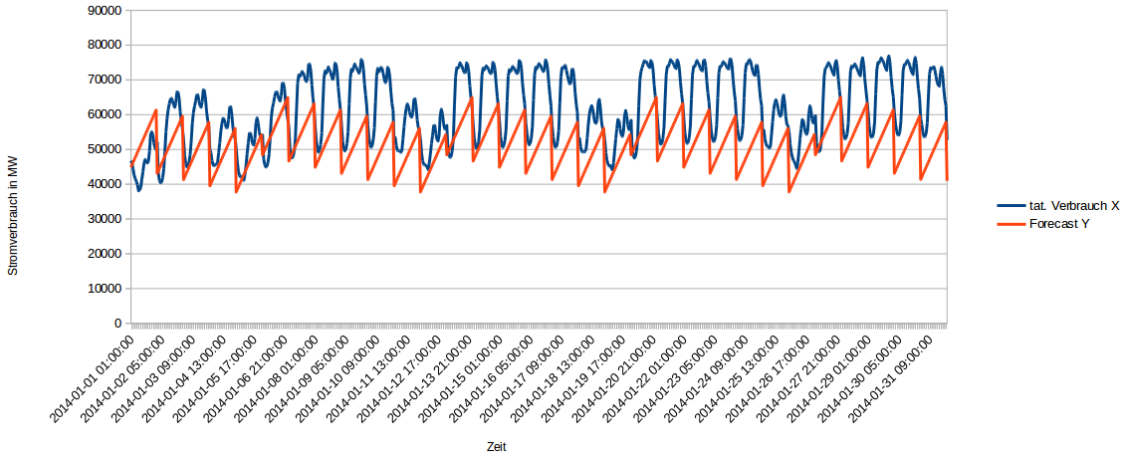


Abbildung 43: Vergleich von tatsächlichen Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2013“

Vergleicht man die grafischen Ergebnisse dieses Durchlaufes mit denen des ersten Versuches der Datenbasis, fällt auf, dass nun in beiden Durchläufen ein „Muster“ im Verlauf der Prädiktionskurve gebildet wird, an denen zumindest die einzelnen Tage grob erkennbar sind. Die „differenziertere“ Darstellung der Zeit hat hier dazu geführt, das im Vergleich zum ersten Versuch der Datenbasis bessere - jedoch nicht signifikante - Ergebnisse produziert werden. Dies wird dementsprechend auch in den Grafiken in Abbildung 42 und 43 sichtbar: Beiden Prädiktionen fehlt ein „Sockelbetrag“ welcher den Vorhersagen zu einer höheren Qualität verhelfen könnte. Dies ist im Durchlauf „2013“ besonders deutlich sichtbar. Ebenso fehlt beiden Prädiktionen die korrekte Modellierung des Stromverbrauches im Verlauf eines Tages (hier existiert lediglich eine Spitze, ohne das die Schwankungen im Verlaufe eines Tages weiter beachtet werden). Dementsprechend fallen die Fehlerkennzahlen der beiden Durchläufe aus: Bei beiden Durchläufen liegt der R-Squared mit 0,39 (Durchlauf „2009_2013“) und 0,40 (Durchlauf „2013“) etwa gleichauf, jedoch liegt der MAE des Durchlaufes „2009_2013“ mit 8229,84 Punkten im Vergleich zum Durchlauf „2013“ um 2773,5 Punkte niedriger. Dies ist auch beim RMSE zu beobachten. Beim Durchlauf „2009_2013“ liegt dieser bei 10049,13 Punkten und ist damit um 2828,41 Punkte niedriger als der Durchlauf „2013“. Hieraus lässt sich also schlussfolgern, dass eine höhere Menge von Trainingsdaten zu besseren Prädiktionsergebnissen führt. Dennoch reicht die differenzierte Darstellung des zeitlichen Verlaufes in den Trainingsdaten offenbar nicht aus, um adäquate, passende Prädiktionsergebnisse zu erzielen.

Exponentielle Regression Der Versuch der Vorhersage mit Hilfe der exponentiellen Regression teilt sich in drei Teilversuche auf. Mit Hilfe der View `CONSUMPTION_TRAINING` (siehe Listing 70) werden drei Modelle gebildet. Der erste Teilversuch bezieht sich dabei auf die gesamten Trainingsdaten. Der zweite Teilversuch bezieht sich auf den Zeitraum Juni bis Dezember 2013 als Trainingsdaten. Der dritte Teilversuch bezieht sich auf das Jahr 2013 als Trainingsdaten. Die Scripte hierzu können auf der beigelegten CD im Ordner `svn/Forecast/Basis/Exponentialregression/` gefunden werden. Im Ordner `1.Versuch` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz. Im Ordner `2.Versuch` befinden sich die Daten zur Modell- und Prädiktionsbildung über Juni bis Dezember 2013. Im Ordner `3.Versuch` befinden sich die Daten zur Modell- und Prädiktionsbildung für das Jahr 2013. In diesen drei Ordnern sind folgende Dateien vorhanden:

Um das Trainingsset für die jeweiligen Durchläufe einzuschränken, wird in der Datei `Exponentialregression.sql` das folgende SQL-Statement verwendet:

Listing 44 zeigt, wie der gesamte Trainingsdatensatz in die Modellbildung eingebunden wird. Listing 46 zeigt, wie lediglich das Jahr 2013 des Trainingsdatensatzes in die Modellbildung eingebunden wird. Dies wird über die Spalte `ID` realisiert. Hierbei entspricht die ID mit den Wert 35064 das Datum 01.01.2013, 00:00:00 Uhr. Listing 45 zeigt, wie lediglich

Datei	Inhalt
Exponentialregression.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014.
fehlerkennzahlen.csv	Vom Algorithmus erstellen Fehlerkennzahlen, bezogen auf die Trainingsdaten.
forecast.csv	Vorhersagewerte für Januar 2014.
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014.
Diagramm.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm.

Tabelle 75: Relevaten Dateien für die Exponentielle Regression

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER" FROM "
    PAL"."CONSUMPTION_TRAINING";
```

Abbildung 44: SQL-Statement für den gesamten Trainingsdatensatz

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER" FROM "
    PAL"."CONSUMPTION_TRAINING" WHERE "ID" > 38785;
```

Abbildung 45: SQL-Statement für Juni bis Dezember 2013 als Trainingsdatensatz

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER" FROM "
    PAL"."CONSUMPTION_TRAINING" WHERE "ID" >= 35064;
```

Abbildung 46: SQL-Statement für das Jahr 2013 als Trainingsdatensatz

die Monate Juni bis Dezember 2013 des Trainingsdatensatzes in die Modellbildung eingebunden wird. Dies wird über die Spalte `ID` realisiert. Hierbei entspricht die ID mit dem Wert 38785 das Datum 01.06.2013, 00:00:00 Uhr. Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Zunächst muss das Script `Exponentialregression.sql` in SAP HANA ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und Übersichtsgründen wurden diese Tabellen jeweils exportiert und liegen in den entsprechenden Ordnern im CSV- beziehungsweise Microsoft-Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der generierten Vorhersagen. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten.

Aus Tabelle 76 sind die gewählten Parametereinstellungen zu entnehmen.

Parameter	Einstellung	Erläuterung
THREAD_NUMBER	64	Modellbildung wird mit 64 Threads durchgeführt.
PMML_EXPORT	2	Gibt an, dass das Modell im PMML-Format exportiert wird.

Tabelle 76: Parametereinstellungen der exponentiellen Regression (für alle Durchläufe)

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/Basis/Exponentialregression/`. Wie bereits weiter oben erwähnt, finden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Die Fehlerkennzahlen zu den drei Vorhersagen sind in Tabelle 77 zusammengefasst dargestellt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
2009-2013	0,38	9641,77	11565,79	0,15	0,18
Jun2013	0,35	9335,74	11272,23	0,15	0,18
2013	0,38	24438,78	25975,80	0,39	0,42

Tabelle 77: Fehlerkennzahlen der Prognose mit der exponentiellen Regression

Dabei bezieht sich der Durchlauf „2009-2013“ auf die Verwendung der gesamten Trainingsdaten für den Zeitraum von 2009 - 2013. Der Durchlauf „Jun2013“ bezieht sich auf die Verwendung der Trainingsdaten für den Zeitraum Juni 2013 bis Dezember 2013. Der Durchlauf „2013“ bezieht sich auf die Verwendung der Trainingsdaten für den Zeitraum 2013 komplett. Abbildung 47 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zum prognostizierten Verbrauch für den Durchlauf „2009-2013“.

Abbildung 48 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zum prognostizierten Stromverbrauch für den Durchlauf „Jun2013“.

Abbildung 49 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktonswerten.

Vergleicht man die Ergebnisse dieses Durchlaufes mit denen des ersten Versuches der Datenbasis, fällt auf, dass nun in allen drei Durchläufen ein „Muster“ im Verlauf der Prädiktionskurve gebildet wird, an denen zumindest die einzelnen Tage grob erkennbar sind. Die „differenziertere“ Darstellung der Zeit hat hier dazu geführt, dass im Vergleich zum ersten Versuch der Datenbasis bessere - jedoch nicht signifikante - Ergebnisse produziert werden. Im Vergleich zum ersten Versuch der Datenbasis fallen hier besonders die beiden Durchläufe „2009-2013“ und „Jun2013“ positiv auf. Der Durchlauf „2023“ produziert - wie aus der entsprechenden grafischen Abbildung 49 ersichtlich - kontinuierlich zu hohe Werte. Die beiden Durchläufe „2009-2013“ und „Jun2013“ produzieren jedoch ebenfalls keine wirklich adäquaten Ergebnisse. Insbesondere werden die Stromverbrauchsschwankungen innerhalb eines Tages nicht dargestellt. Ebenfalls werden Arbeitstage und

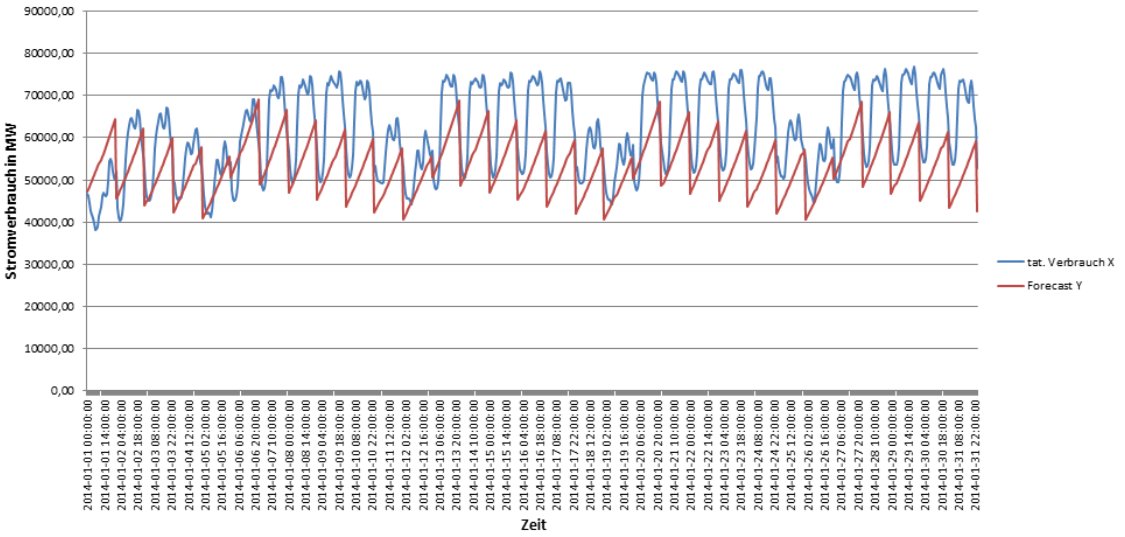


Abbildung 47: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „2009-2013“

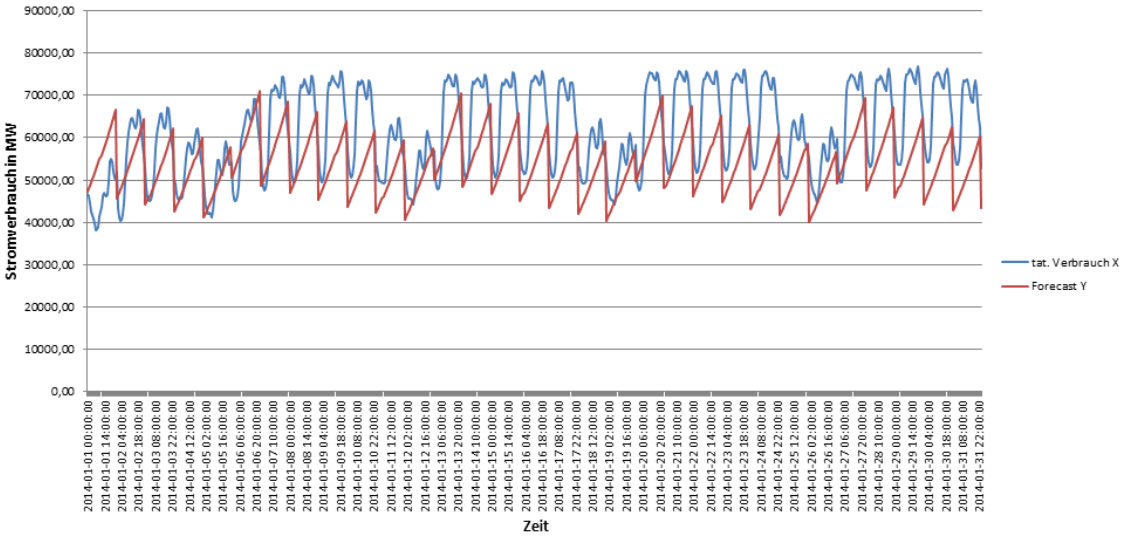


Abbildung 48: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Jun2013“

Wochenenden nur mangelhaft erkannt und die Stromverbrauchshöhen werden von beiden Durchläufen zu niedrig prognostiziert. Betrachtet man alle drei Grafiken, wird ersichtlich, dass kein Durchlauf signifikante Ergebnisse produziert. Dies wird auch an den Fehlerkennzahlen (siehe Tabelle 77) ersichtlich: Der Durchlauf „Jun2013“ erreicht hier einen MAE von 9335,74 und einen RMSE von 11272,23 Punkten. Gemessen an diesen Fehlerkennzahlen erreicht dieser Durchlauf die besten Ergebnisse dieses Versuches. Knapp dahinter erzielt der Durchlauf „2009-2013“ mit einen MAE von 9641,77 und einen RMSE von 11656,79 die zweitbesten Ergebnisse. Der R-Squared dieses Durchlaufes liegt bei 0,38. Verglichen

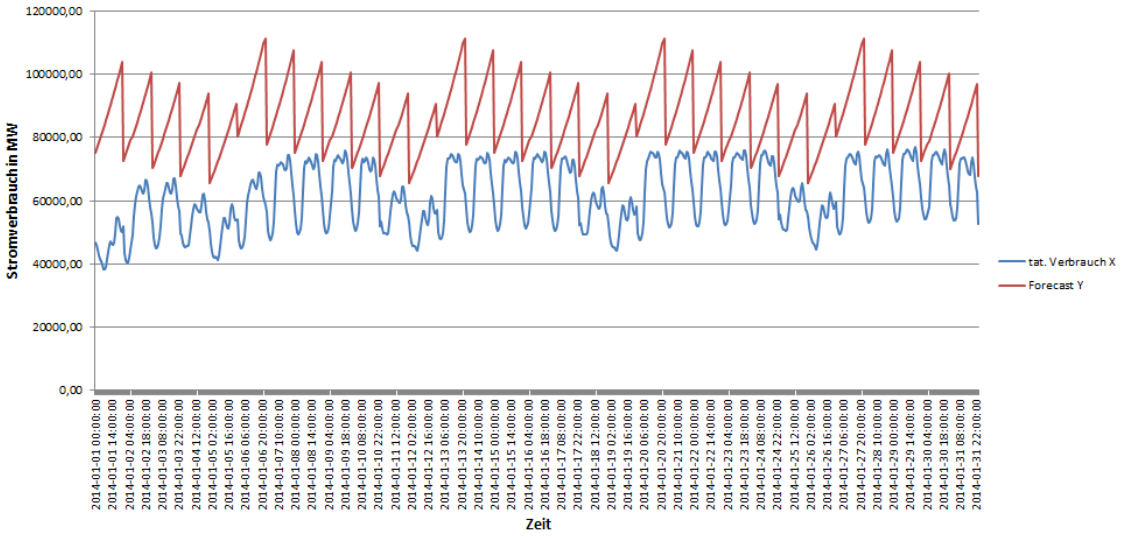


Abbildung 49: Diagramm Vergleich Ist und Forecast

mit dem R-Squared des Durchlaufes „Jun2013“ mit 0,35 Punkten, ist die Anpassung des Algorithmus an die tatsächlichen Daten jedoch etwas besser als beim besten Durchlauf „Jun2013“. Die Fehlerkennzahlen des Durchlaufes „2013“ weisen hier die schlechtesten Ergebnisse auf. Dies wird auch in der grafischen Abbildung 49 sichtbar. Hier liegt der MAE bei 24438,78 und der RMSE bei 25975,8 Punkten. Der R-Squared erzielt hier einen Wert von 0,38 Punkten, was zwar einer Anpassung des Algorithmus an die tatsächlichen Daten analog zum Durchlauf „2009-2013“ entspricht, jedoch werden die Prognosen zu hoch berechnet.

Support Vector Machine Im Zuge dieses Versuchs mit der Support Vector Machine wird ein radialer Kern genutzt, um von einer geringer benötigten Rechenleistung im Vergleich zum polynomialen Kernel zu profitieren. Zuvor durchgeführte Testläufe mit dem Polynomialen Kernel zeigen auf, dass Prognosen aufgrund der für diese Hypothese relevanten Daten Laufzeiten von bis zu mehreren Stunden aufweisen. Deshalb wird der polynomiale Kernel nicht weiter betrachtet. Aufgrund den Erfahrungen, die bereits durch die vorherigen Algorithmen gesammelt worden sind, werden die gesamten Daten für den Zeitraum 2009 bis 2013 für die Modellbildung miteinbezogen. Denn es hat sich ergeben, dass kein signifikanter Unterschied zwischen der Nutzung der gesamten und einer selektiven Datenmenge zu erkennen ist. Diese Voraussetzungen werden auch für alle weiteren Hypothesen übernommen.

In diesem Fall wird die View *CONSUMPTION_TRAINING* für den Versuch ausgewählt. Weiterhin kann das Script hierzu auf der beigelegten CD im Ordner `svn/Forecast/Basis SVM/` gefunden werden. In diesen Ordnern sind folgende Dateien vorhanden:

Um das Trainingsset für den Durchlauf einzuschränken, wird in der Datei `build_and`

Datei	Inhalt
build_and_forecast_svm_yr_qt_mth_dom_dow_hod.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014.
g0_001#c1000.xlsx	Mehrere Dateien nach dem Muster g(Gamma-Wert mit Unterstrich als Komma)#c(C-Wert mit Unterstrich als Komma).xlsx; Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 mitsamt Fehlerkennzahlen.
Diagramm.xlsx	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 78: Relevante Dateien für die Support Vector Machine

`_forecast_svm_yr_qt_mth_dom_dow_hod.sql` das folgende SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6 FROM "PAL"."
      CONSUMPTION_TRAINING";
```

Abbildung 50: SQL-Statement für den gesamten Trainingsdatensatz bei der SVM

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden muss folgendermaßen vorgegangen werden: Zunächst muss das Script `build_and_forecast_svm_yr_qt_mth_dom_dow_hod.sql` in SAP HANA ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Diese Daten müssen anschließend exportiert und in die entsprechende Spalte der Datei `Januar2014.xlsx` eingetragen werden. Die Datei `Januar2014.xlsx` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Vorhersage. Des Weiteren sind hier auch die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten.

In der folgenden Tabelle 79 sollen die verschiedenen Parameterpaare den bei den Durchläufen entstehenden Fehlerkennzahlen gegenübergestellt werden.

Jetzt erfolgt die Beschreibung einiger prägnanter Ergebnisse. Abbildung 51 zeigt das Ergebnis mit den Parametern $\gamma = 1$ und $C = 1$. Es wird deutlich, dass die Prognose sich sehr schlecht an den tatsächlichen Verbrauch anpasst. Stattdessen wird eine Art Durchschnittsverlauf deutlich.

Eine weitere Erhöhung von γ auf 1000 führt zu keinem sichtbaren Erfolg. Die Prognose erfolgt weiterhin linear ohne Steigung als Durchschnittsverlauf. Durch eine Reduzierung

Lauf	Parameter		Fehlerkennzahlen				
γ	C	R^2	MAE	RMSE	CV_MAE	CV_RMSE	CV_RMSE
1	0,001	1	0,39	10371,75	12648,42	0,17	0,21
2	1	1	0,65	10349,01	12615,72	0,17	0,20
3	1000	1	0,002	10397,79	12675,47	0,17	0,21
4	0,001	0,1	0,39	10394,57	12671,9	0,17	0,21
5	0,001	100	0,38	9253,88	11338,56	0,15	0,18
6	1	100	0,8	75097,39	75478,3	1,22	1,23
7	0,01	100	0,42	9061,93	10829,23	0,15	0,18
8	0,1	100	0,68	79567,49	79820,51	1,29	1,3
9	0,05	100	0,62	9281,22	10693,67	0,15	0,17
10	0,05	1000	0,7	77644,5	77866,35	1,26	1,26
11	0,01	1000	0,57	8814,5	10230,36	0,14	0,17
12	0,005	1000	0,46	8989,79	10607,73	0,15	0,17
13	0,015	1000	0,61	8995,55	10316,82	0,15	0,17
14	0,015	1000	0,65	77504,91	77757,49	1,26	1,26
15	0,01	10000	0,63	77465,2	77728,26	1,26	1,26
16	0,001	10000	0,42	75332,91	75723,82	1,22	1,23

Tabelle 79: Parametereinstellungen der Support Vector Machine mit radialem Kernel

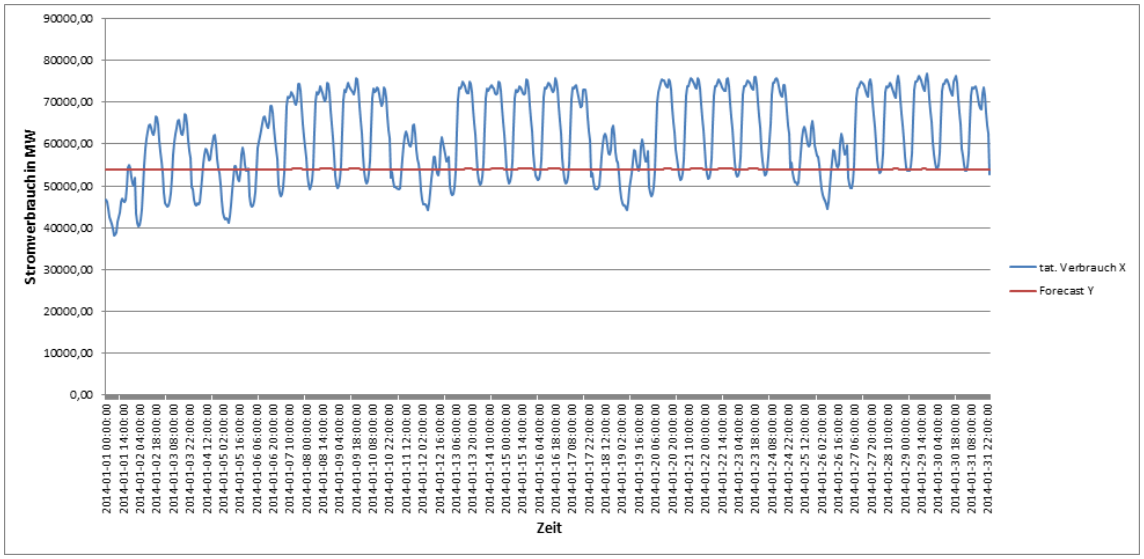


Abbildung 51: Support Vector Maschine mit $\gamma = 1$ und $C = 1$

von γ auf 0.01 und Erhöhung von C auf 100 wird eine deutlich bessere Prognose ermöglicht. Abbildung 52 zeigt den Verlauf.

Jetzt wird C weiter auf 1000 erhöht. Abbildung 53 zeigt den Verlauf mit den Werten $\gamma = 0.01$ und $C = 1000$. Unter Beachtung der Fehlerkennzahlen in Tabelle 79 wird deutlich, dass diese Anpassung des Modells zu der bisher besten Prognose in Verbindung mit der Support Vector Maschine führt.

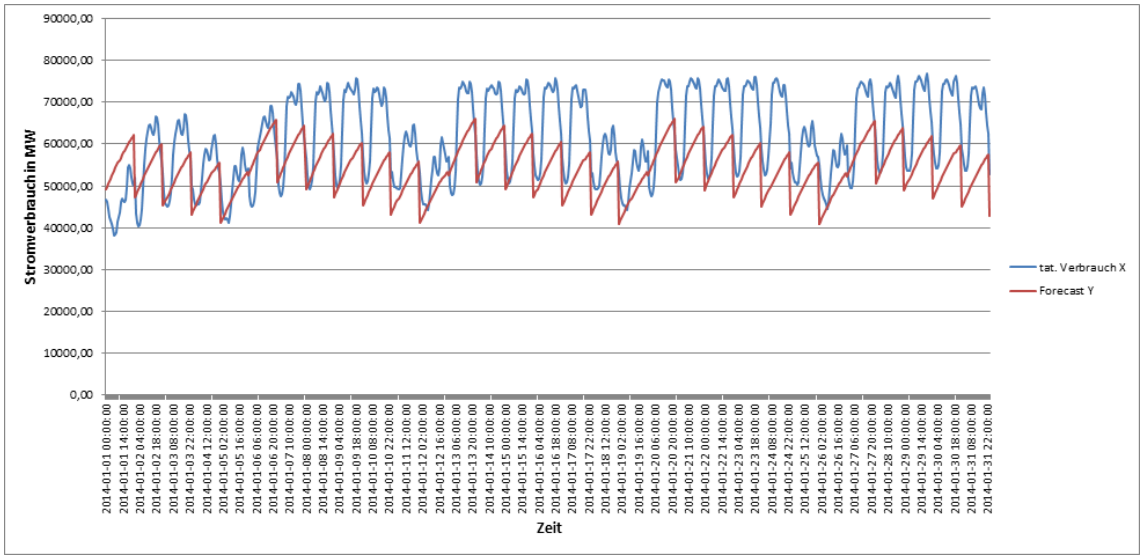


Abbildung 52: Support Vector Maschine mit $\gamma = 0.01$ und $C = 100$

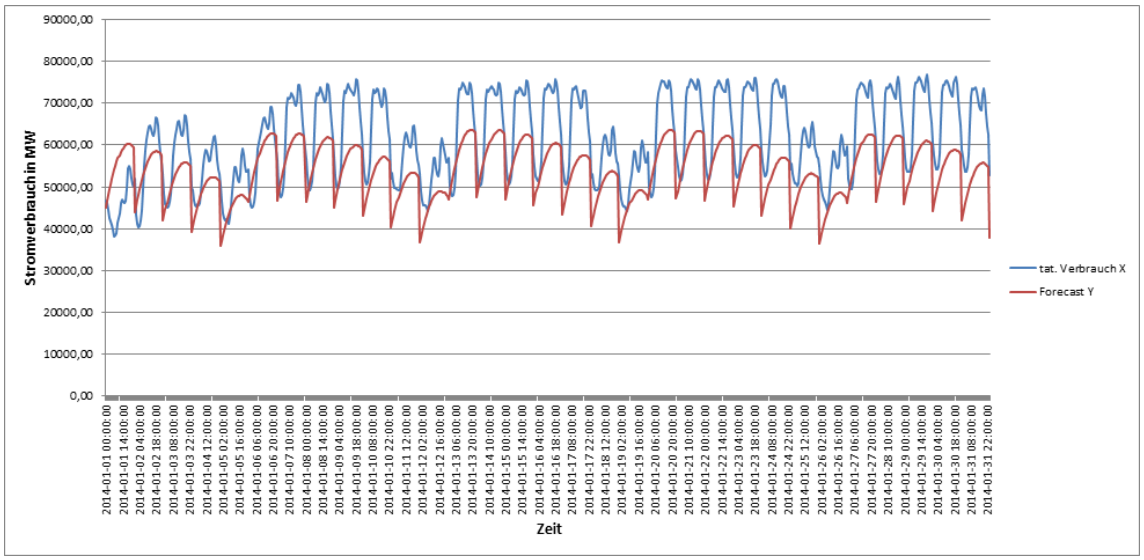


Abbildung 53: Support Vector Maschine mit $\gamma = 0.01$ und $C = 1000$

Zusammenfassung der Ergebnisse In diesem Versuch fließt der Verlauf der Zeit im Vergleich zum ersten Versuch in differenzierter Art und Weise in die Modell- und Prognosebildung ein. Dies führt dazu, dass die Algorithmen multiple lineare Regression und exponentielle Regression im Vergleich zum ersten Versuch der Datenbasis bessere Ergebnisse produzieren. Insbesondere ist bei allen Prognosen der beiden Algorithmen nun ein Muster in den Vorhersagen ersichtlich, aus denen teilweise der Verlauf einzelner Tage ersichtlich wird. Dies ist im ersten Versuch der Datenbasis bei beiden Algorithmen nicht der Fall. Dadurch wird die Haupthypothese, dass durch eine höhere Anzahl von Daten bessere Prognoseergebnisse generiert werden, unterstützt. In diesem Versuch wird ebenfalls die

Support Vector Machine eingeführt, da sie neben der exponentiellen und der multiplen linearen Regression ebenfalls N Variablen für die Modell- und Prognoseberechnung erlaubt. Es folgt ein Vergleich der besten Ergebnisse der in Konkurrenz zueinander stehenden Algorithmen. Das sind die multiple lineare Regression, die exponentielle Regression sowie die Support Vector Machine. Die folgenden Durchläufe beziehen sich dabei auf den besten Durchlauf des jeweiligen Algorithmus in diesem Versuch.

Die beste Prognose der multiplen linearen Regression erzielt der Durchlauf „2009_2013“. Die Fehlerkennzahlen liegen hier bei: R-Squared= 0,39, MAE= 8229,84, RMSE= 10049,13. Die beste Prognose der exponentiellen Regression erzielt der Durchlauf „Jun2013“ mit einem R-Squared von 0,35, einem MAE von 9335,74 und einem RMSE von 11272,23. Fraglich ist bei diesem Durchlauf jedoch die Tatsache, dass eine Modellbildung auf Basis von Trainingsdaten Juni 2013 bis Dezember 2013 erfolgt, die Prognose jedoch für den Januar 2014 getätigt wird. Es wird also ein Zeitraum prognostiziert, dessen historische Daten nicht in den Trainingsdaten enthalten ist. Aus diesem Grund soll hier noch einmal der zweitbeste Durchlauf der exponentiellen Regression erwähnt werden. Dies ist der Durchlauf „2009-2013“, bei dem das gesamte Trainingsdatenset in die Modellbildung einbezogen wird. Bei diesen Durchlauf liegt der R-Squared bei 0,38, der MAE bei 9641,77 und der RMSE bei 11656,79. Die beste Prognose der Support Vector Machine wird in diesen Durchlauf mit den Anpassungsfaktoren $C = 1000$ und $\gamma = 0,01$ erzielt. Hier liegen die Fehlerkennzahlen bei R-Squared = 0,57, MAE = 8814,08 und RMSE= 10230,36. Im Vergleich dieser Ergebnisse wird deutlich, dass der Durchlauf „2009_2013“ der multiplen linearen Regression das beste Ergebnis aller drei Algorithmen erzielt.

12.4 1. Hypothese

Die erste Hypothese lautet, dass die Prädiktionen für den Stromverbrauch, angereichert mit Informationen zu Samstagen, Sonntagen und Feiertagen, genauere Ergebnisse liefern als die Prädiktionen der Datenbasis (siehe Kapitel 12.3.1). Die Daten für diese Hypothese bestehen also aus den historischen Energieverbrauchsdaten der Entso-E vom 01.01.2009 00:00:00 bis zum 31.12.2013 23:00:00 Uhr. Zusätzlich werden die Daten mit Informationen zu Samstagen, Sonntagen und Feiertagen der Jahre angereichert. Ebenso sind die Informationen zur differenzierten Darstellung der Zeit aus dem zweiten Versuch der Datenbasis enthalten (siehe Kapitel 12.3.2). Im folgenden Abschnitt wird zunächst der Datenfluss beschrieben. Die erzielten Ergebnisse der Algorithmen werden im Anschluss erläutert.

Datenfluss Die Informationen der Sonn- und Feiertage stammen aus der Tabelle `OLIMP.DWD_Astro`. Diese Tabelle enthält für jeden Tag die Information, ob ein Feiertag vorliegt (zum Beispiel die Osterfeiertage sowie die Weihnachtsfeiertage). In Tabelle 80 sind hierzu einige Beispieldaten notiert. Die Spalte `ArbZeitFaktor` definiert hier binär, ob ein Feiertag (0) oder kein Feiertag (1) vorliegt.

Datum	**...**	**...**	**ArbZeitFaktor**
01.01.2009	...	...	0
02.01.2009	...	...	1
...	...	...	...

Tabelle 80: Tabelle `OLIMP.DWD_ASTRO`

Mittels des in Listing 54 angegebenen SQL-Codes werden diese Informationen in die Tabelle `_SYS_BI.M_TIME_DIMENSION` übertragen. Hierzu wird zunächst manuell eine neue Spalte in der Tabelle `_SYS_BI.M_TIME_DIMENSION` mit dem Namen `WORKDAY` angelegt. Anschließend werden die Feiertagsinformationen über den Vergleich der Felder `Datum` der Tabelle `DWD_ASTRO` und `DATE_SQL` der Tabelle `_SYS_BI.M_TIME_DIMENSION` in die Tabelle übertragen.

```
1 UPDATE "_SYS_BI"."M_TIME_DIMENSION" SET "WORKDAY" = a."ArbZeitFaktor"
2 FROM (select "Datum", "ArbZeitFaktor" FROM "OLIMP"."DWD_Astro") a
3 WHERE a."Datum"  = "_SYS_BI"."M_TIME_DIMENSION"."DATE_SQL"
```

Abbildung 54: SQL-Code zum Übertragen der Feiertagsinformationen

Es können die Informationen in die View für die Algorithmen eingefügt werden. Hierzu wird die vorhandene View aus 38 erweitert. Der SQL-Code hierzu ist in Listing 55 zu finden. Für die Evaluation dieser Hypothese werden der View zwei Spalten hinzugefügt.

```
1  CREATE VIEW "PAL"."CONSUMPTION_TRAINING" ( "ID",
2              "CONSUMPTION",
3              "HOUR_COUNT",
4              "HOUR_OF_DAY",
5              "DAY_OF_WEEK",
6              "DAY_OF_MONTH",
7              "MONTH",
8              "QUARTER",
9              "YEAR",
10             "WORKDAY",
11             "WEEKEND" ) AS SELECT
12             c.ID-1 as ID,
13             c.CONSUMPTION as CONSUMPTION,
14             c.ID as HOUR_COUNT,
15             t.HOUR_INT as HOUR_OF_DAY,
16             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
17             t.DAY_INT as DAY_OF_MONTH,
18             t.MONTH_INT as "MONTH",
19             t.QUARTER_INT as "QUARTER",
20             t.YEAR_INT as "YEAR" ,
21             t.WORKDAY as "WORKDAY",
22             CASE WHEN t.DAY_OF_WEEK_INT = 5
23  or t.DAY_OF_WEEK_INT = 6
24  THEN 1
25  ELSE 0
26  END as "WEEKEND"
27  FROM "PRE"."CONSUMPTION_ALL_CLEAN" as c JOIN "_SYS_BI"."M_TIME_DIMENSION" as
        t ON c.ID+8783 = t."HOUR_COUNT"
28  ORDER BY ID ASC WITH READ ONLY
```

Abbildung 55: View für die Trainingsdaten der Stromverbräuche mit zusätzlichen Zeitangaben sowie Feiertage und Wochenende

Diese sind `WORKDAY` und `WEEKEND`. Die Spalte `WORKDAY` bezeichnet dabei die Angabe, ob es sich bei dem aktuellen Tag um einen Feiertag handelt und bezieht ihre Informationen aus der Tabelle `_SYS_BI.M_TIME_DIMENSION`. Die Spalte `WEEKEND` gibt an, ob es sich bei dem aktuellen Tag um einen Samstag oder Sonntag handelt. Dies wird über eine Abfrage im SQL-Code realisiert: Wenn die Spalte `DAY_OF_WEEK` den Wert 5 oder 6 hat (siehe Kapitel 39), so nimmt die Spalte `WEEKEND` den Wert 1 an. Dies bedeutet, dass es sich bei dem aktuellen Datensatz um ein Wochenende handelt. Analog hierzu ist die View für die Testdaten (Zeitraum 01.01.2014 00:00:00 bis 31.01.2014 23:00:00 Uhr) angepasst. Der SQL-Code hierzu ist in Listing 56 zu finden. Analog zu 39 enthält die View alle vorherigen Daten ohne den Stromverbrauch selbst sowie die im obigen Abschnitt erläuterten Informationen zu Feiertagen und Wochenenden.

Versuchsdurchführung Die Vorhersagen für die Evaluation dieser Hypothese teilen sich in drei Versuche auf. Mit Hilfe der View `CONSUMPTION_TRAINING` (siehe Kapitel 55) werden Modelle mit den entsprechenden Algorithmen gebildet. Dabei werden folgende Versuche betrachtet:

```
1  CREATE VIEW "PAL"."CONSUMPTION_FORECAST" ( "ID",
2              "HOUR_COUNT",
3              "HOUR_OF_DAY",
4              "DAY_OF_WEEK",
5              "DAY_OF_MONTH",
6              "MONTH",
7              "QUARTER",
8              "YEAR",
9              "WORKDAY",
10             "AIRTEMP",
11             "WEEKEND" ) AS SELECT
12             t.HOUR_COUNT-8783-43824-1 as ID,
13             t.HOUR_COUNT-8783 as HOUR_COUNT,
14             t.HOUR_INT as HOUR_OF_DAY,
15             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
16             t.DAY_INT as DAY_OF_MONTH,
17             t.MONTH_INT as "MONTH",
18             t.QUARTER_INT as "QUARTER",
19             t.YEAR_INT as "YEAR",
20             t.WORKDAY as "WORKDAY",
21             CASE WHEN t.DAY_OF_WEEK_INT = 5
22 or t.DAY_OF_WEEK_INT = 6
23 THEN 1
24 ELSE 0
25 END as "WEEKEND"
26 FROM "_SYS_BI"."M_TIME_DIMENSION" as t
27 WHERE t.DATETIMESTAMP >= '2014-01-01 00:00:00'
28 AND t.DATETIMESTAMP <= '2014-01-31 24:00:00'
29 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 56: View für die Testdaten der Stromverbräuche mit zusätzlichen Zeitangaben sowie Feiertage und Wochenende

1. Prädiktion mit Angaben zu Feiertagen
2. Prädiktion mit Angaben zu Wochenenden
3. Prädiktion mit Angaben zu Feiertagen und Wochenenden

Multiple lineare Regression Alle Versuche der multiplen linearen Regression können auf der beigelegten CD im Ordner `svn/Forecast/1.Hypothese/MultipleLineareRegression/` gefunden werden. Im Ordner `Feiertage` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Angaben zu Feiertagsdaten. Im Ordner `Feiertage_und_Wochenende` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Angaben zu Feiertagen und Wochenenden. Im Ordner `Wochenende` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Angaben zum Wochenende. Tabelle 81 zeigt, welche Dateien in dem besagten Ordner vorhanden sind.

Um das Trainingsset für die Durchläufe einzuschränken, werden in der jeweiligen Modellbildung SQL-Codes verwendet. Listing 57 zeigt, wie der gesamte Trainingsdatensatz mit

Datei	Inhalt
build_modell_mlr.sql	Script zum Erstellen des Modells
build_forecast_mlr.sql	Script zum Erstellen der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellte Fehlerkennzahlen, bezogen auf die Trainingsdaten
data.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014
Diagramm.xlxs	Grafische Darstellung von tatsächlichem und vorhergesagtem Stromverbrauch

Tabelle 81: Relevante Dateien für die multiple lineare Regression

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WORKDAY
    " FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 57: SQL-Statement für den gesamten Trainingsdatensatz mit Feiertagen

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    " FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 58: SQL-Statement für den gesamten Trainingsdatensatz mit Wochenenden

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ", "WORKDAY" FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 59: SQL-Statement für den gesamten Trainingsdatensatz mit Wochenenden und Feiertagen

den Feiertagen in die Modellbildung eingebunden wird. Listing 58 zeigt, wie der gesamte Trainingsdatensatz mit den Wochenenden in die Modellbildung eingebunden wird. Listing 59 zeigt, wie der gesamte Trainingsdatensatz, sowie die Feiertage und die Wochenenden in die Modellbildung eingebunden werden. Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Zunächst wird das Script `build_modell_mlr.sql` in SAP HANA ausgeführt, welches das mathematische Modell des Algorithmus auf Basis der Trainingsdaten anlegt. Anschließend muss das Script `build_forecast_mlr.sql` ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und aus Übersichtsgründen werden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die vom SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `data.csv` beinhaltet die von

dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset, enthalten. Die Datei `Diagramm.xlsx` zeigt die grafische Darstellung von tatsächlichem und vorhergesagtem Stromverbrauch.

Die Parametereinstellungen des Algorithmus sind aus Tabelle 82 zu entnehmen.

Parameter	**Einstellung**	**Erläuterung**
THREAD_NUMBER	8	Modellbildung wird mit 8 Threads durchgeführt.
PMML_EXPORT	1	Gibt an, dass das Modell im PMML-Format vorliegt.
ADJUSTED_R2	1	R-Squared und R-Squared-Adjusted werden berechnet.
VARIABLE_SELECTION	0	Alle in der View vorhandenen Variablen werden zur Modellbildung einbezogen.

Tabelle 82: Parametereinstellungen der multiplen linearen Regression (für alle Durchläufe)

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/1.Hypothese/MultipleLineareRegression`. Wie bereits weiter oben erwähnt, finden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Die Fehlerkennzahlen zu beiden Prädiktionen sind in Tabelle 83 zusammengefasst dargestellt.

Durchlauf	**R-Squared**	**MAE**	**RMSE**	**CV_MAE**	**CV_RMSE**
Fe	0,48	7891,41	9740,15	0,13	0,16
Wo	0,45	8006,52	9587,26	0,13	0,16
Fe_Wo	0,54	7585,87	9165,80	0,12	0,14

Tabelle 83: Ergebnisse der Hypothese

Dabei bezieht sich der Durchlauf „Fe“ nur auf die Verwendung der Feiertagsangaben. Der Durchlauf „Wo“ bezieht sich auf die Verwendung der Angaben zum Wochenende. Der Durchlauf „Fe_Wo“ bezieht sich auf die Verwendung von Wochenends- und Feiertagsangaben.

Abbildung 60 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe“.

Abbildung 61 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Wo“.

Abbildung 62 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe_Wo“.

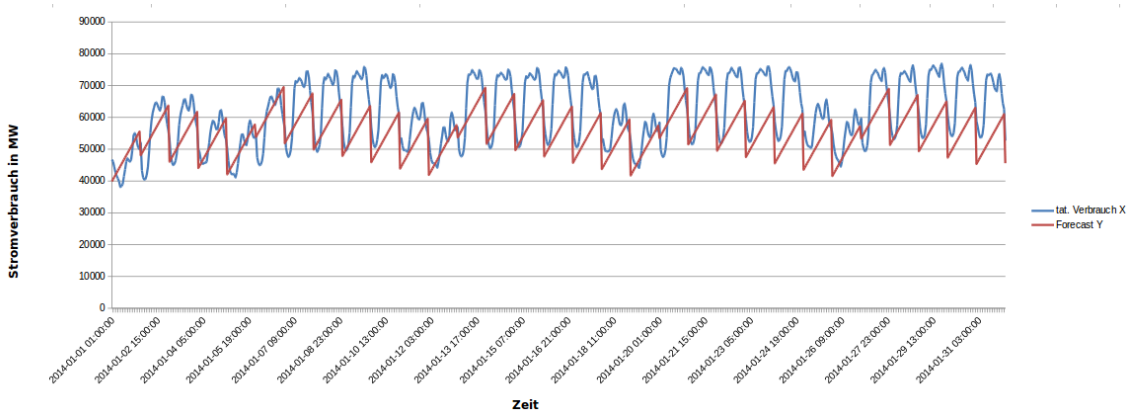


Abbildung 60: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe“

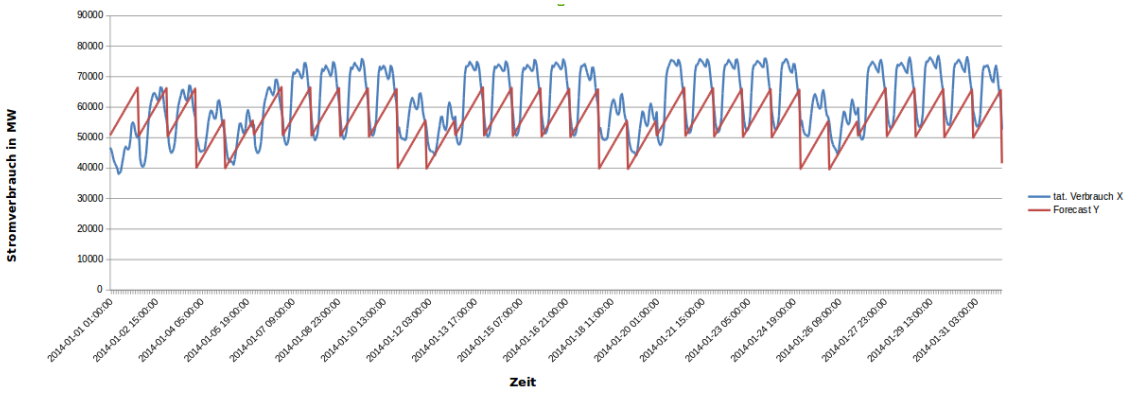


Abbildung 61: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Wo“

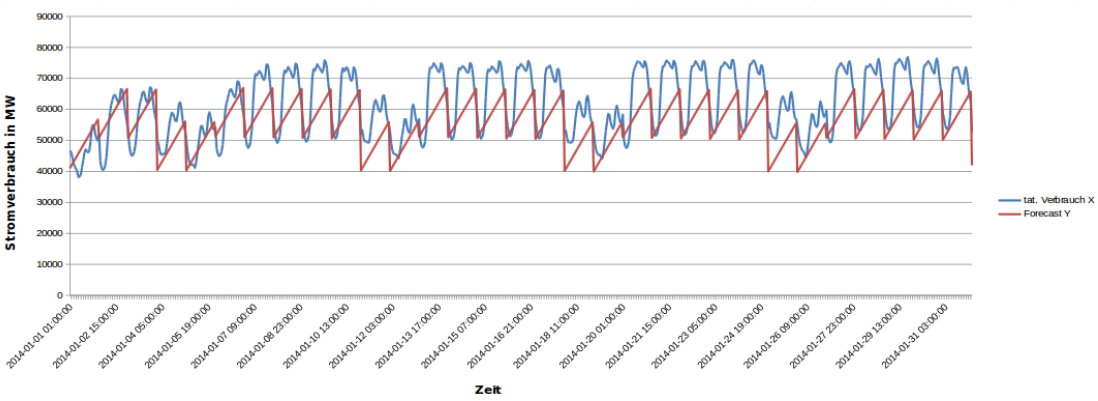


Abbildung 62: Vergleich von tatsächlichen Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe_Wo“

Wie im zweiten Versuch der Datenbasis bereits festgestellt, zeigen die grafischen Darstellungen der Prädiktionsergebnisse in Abbildung 60, 61 und 62 nun einen „Musterverlauf“

in dem die einzelnen Tage bereits deutlich erkennbar sind. Hier zeigen insbesondere die beiden Durchläufe „Fe“ sowie „Fe_Wo“ in den ersten Tagen des Prädiktionszeitraumes bessere Ergebnisse verglichen mit dem Durchlauf „Wo“. Der Durchlauf „Wo“ modelliert zwar sehr eindeutig Arbeitstage und Wochenende, sowie die Stromverbrauchstiefen (die Punkte, an denen der Stromverbrauch am niedrigsten ist), jedoch haben alle Grafiken jedoch gemeinsam, dass die Stromverbrauchshöhen zu niedrig vorhergesagt werden. Ebenso wird der Verlauf des Stromverlaufes eines Tages in allen Versuchen nicht korrekt modelliert. Hier existiert eine Spitze, ohne dass die Schwankungen im Verlaufe eines Tages weiter beachtet werden. Dementsprechend sind alle drei Ergebnisse nicht befriedigend, was sich auch in den Fehlerkennzahlen der Algorithmen in Tabelle 83 widerspiegelt. Grafisch - wie auch in den Fehlerkennzahlen ersichtlich - erreicht hier der Durchlauf „Fe_Wo“ die besten Ergebnisse. Hier ist nicht nur der R-Squared mit 0,54 um 0,6 Punkte höher als das zweitbeste Ergebnis (Durchlauf „Fe“). Der Durchlauf „Wo“ liefert mit einem R-Squared von 0,45 das schlechteste Ergebnis. Dieses Verhalten ist auch an den weiteren Fehlerkennzahlen erkennbar. Der Durchlauf „Fe_Wo“ erreicht für diesen Versuch den niedrigsten MAE von 7585,87 Punkten. Wie auch beim R-Squared erreicht der Durchlauf „Fe“ hier das zweitbeste Ergebnis mit 7891,41 Punkten. Auch bei dieser Fehlerkennzahl erreicht der Durchlauf „Wo“ mit 8006,52 Punkten das schlechteste Ergebnis für diesen Durchlauf. Gemessen an der Fehlerkennzahl RMSE erreicht auch hier der Durchlauf „Fe_Wo“ das beste Ergebnis mit 9165,80 Punkten. Das zweitbeste Ergebnis erreicht hier jedoch der Durchlauf „Wo“ mit 9587,26 Punkten. Der Durchlauf „Fe“ erreicht hier lediglich 9740,15 Punkte. In der Tabelle sind ebenfalls die für die Fehlerkennzahlen MAE und RMSE errechneten relativen Kennzahlen ersichtlich. Zusammenfassend lässt sich - bezogen auf die multiple lineare Regression - für diesen Versuch sagen, dass der Durchlauf „Fe_Wo“ das beste Ergebnis erzielt. Dieser Durchlauf enthält für diese Hypothese auch die meisten Daten, da für die Prädiktion die Informationen zu Feiertagen sowie zum Wochenende genutzt werden. Dennoch besteht - hinsichtlich der Prädiktionsqualität - durchaus Steigerungsbedarf.

Exponentielle Regression Die exponentielle Regression wird in zwei weiteren Teilversuchen durchgeführt. Im ersten Teilversuch werden alle Trainingsdaten von 2009 bis 2013 für die Modellbildung miteinbezogen. Im zweiten Teilversuch erfolgt lediglich der Miteinbezug der Trainingsdaten von 2013. Die Versuche zur exponentiellen Regression können auf der beigelegten CD im Ordner `svn/Forecast/1.Hypothese/Exponentialregression/` gefunden werden. Im Ordner `Feiertag` befinden sich die Daten zur Modell- und Prädiktionsbildung für den Durchlauf mit mit den Angaben zu Feiertagen. Im Ordner `Wochenende` befinden sich die Daten zur Modell- und Prädiktionsbildung für den Durchlauf mit Angaben zu Wochenenden. Im Ordner `FeiertagWochenende` befinden sich die Daten zur Modell- und Prädiktionsbildung für den Durchlauf mit den Angaben zu Wochenenden und Feiertagen. Es existieren innerhalb der Ordner für die Durchläufe außerdem Unterordner für die Unterscheidung der Versuche. In jedem dieser Unterordner sind folgende

Dateien vorhanden:

Datei	Inhalt
Exponentialregression.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellte Fehlerkennzahlen, bezogen auf die Trainingsdaten
forecast.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014
Diagramm.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 84: Relevante Dateien für die exponentiale Regression

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Zunächst muss das Script `Exponentialregression.sql` in SAP HANA ausgeführt werden. Hiernach liegen die Vorhersagen für Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und Übersichtsgründen werden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die vom SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset, enthalten. Die Parametereinstellungen des Algorithmus sind aus Tabelle 85 zu entnehmen.

Parameter	Einstellung	Erläuterung
THREAD_NUMBER	64	Modellbildung wird mit 64 Threads durchgeführt.
PMML_EXPORT	2	Gibt an, dass das Modell im PMML-Format exportiert wird.

Tabelle 85: Parametereinstellungen der exponentialen Regression (für alle Durchläufe)

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/1.Hypothese/Exponentialregression`. Wie bereits weiter oben erwähnt, befinden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Dabei bezieht sich jeweils der Durchlauf „Fe“ nur auf die Verwendung der Feiertagsangaben. Der Durchlauf „Wo“ bezieht sich auf die Verwendung der Angaben zum Wochenende. Der Durchlauf „Fe_Wo“ bezieht sich auf die Verwendung von Wochenends- und Feiertagsangaben.

Die Fehlerkennzahlen zu den Prädiktionen des ersten Teilversuchs (Trainingsdaten von

2009-2013) sind in Tabelle 86 zusammengefasst dargestellt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Fe	0,46	9184,28	11181,91	0,15	0,18
Fe_Wo	0,54	8810,64	10540,96	0,14	0,17
Wo	0,44	9308,79	10968,18	0,15	0,18

Tabelle 86: Fehlerkennzahlen der Prognose des ersten Teilversuchs (2009-2013)

Abbildung 63 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe“.

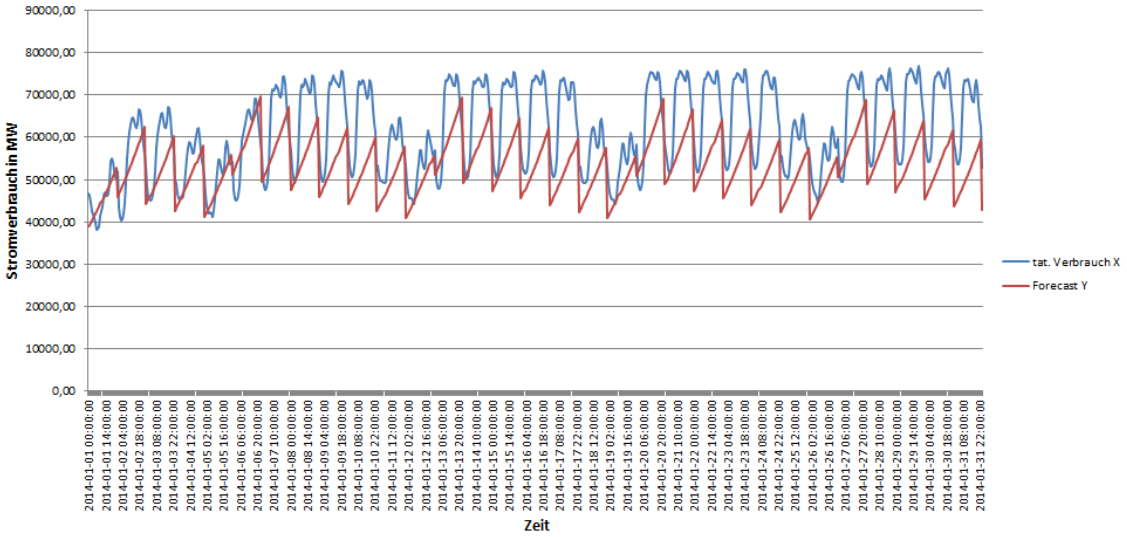


Abbildung 63: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe“

Abbildung 64 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Wo“.

Abbildung 65 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe_Wo“.

Die Fehlerkennzahlen zu den Prädiktionen des zweiten Teilversuchs (Trainingsdaten nur für 2009) sind in Tabelle 87 zusammengefasst dargestellt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Fe	0,47	9762,66	11764,89	0,16	0,19
Wo	0,45	9974,55	11656,78	0,16	0,19
Fe_Wo	0,55	9473,38	11216,39	0,15	0,18

Tabelle 87: Fehlerkennzahlen der Prognose des zweiten Teilversuchs (2013)

Abbildung 66 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe“.

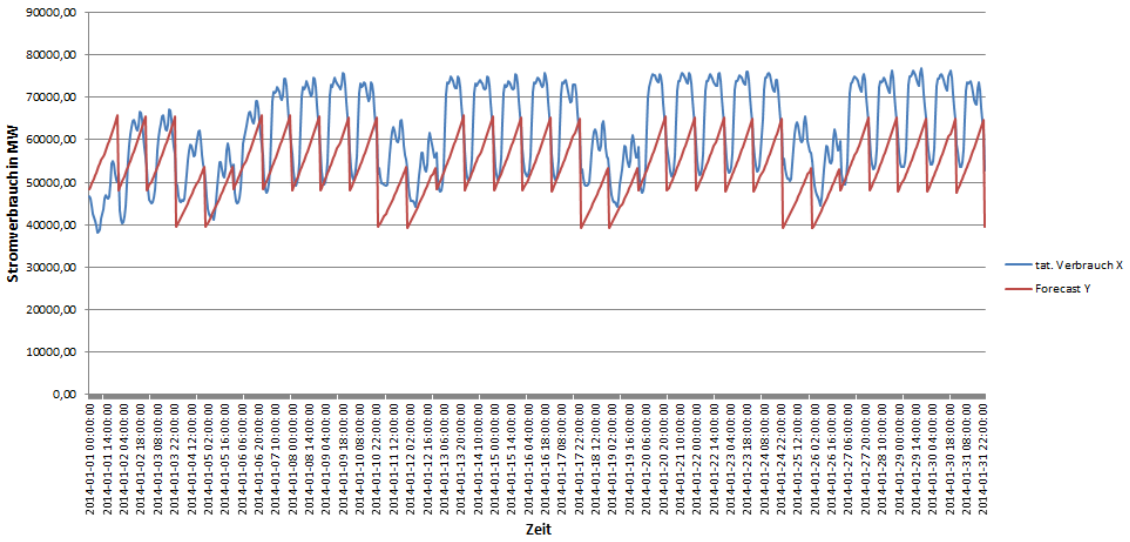


Abbildung 64: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Wo“

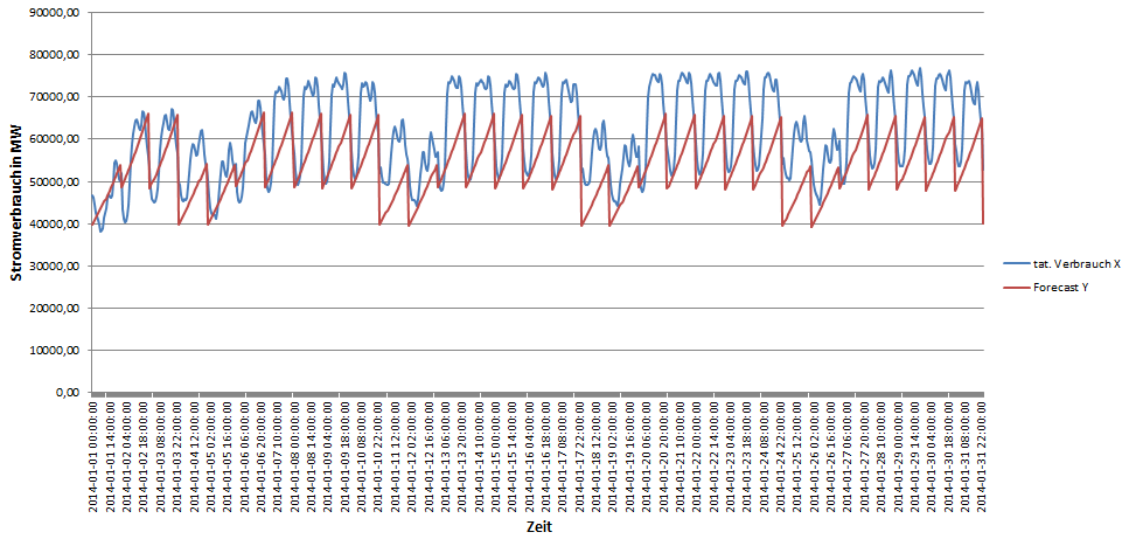


Abbildung 65: Vergleich von tatsächlichem Stromverbrauch zum vorhergesagten Verbrauch. Durchlauf: „Fe_Wo“

Abbildung 67 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Wo“.

Abbildung 68 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf „Fe_Wo“.

Interpretation des ersten Teilversuchs Der erste Teilversuch (siehe auch Tabelle 86) enthält die Datensätze von 2009-2013 als Trainingsdaten. In den ersten Tagen des Prädik-

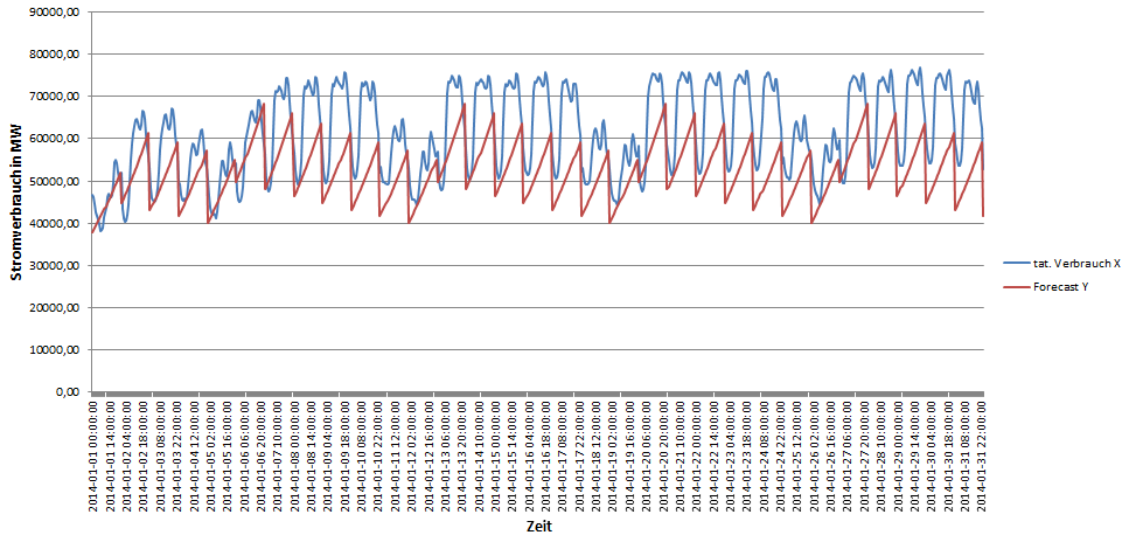


Abbildung 66: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe“

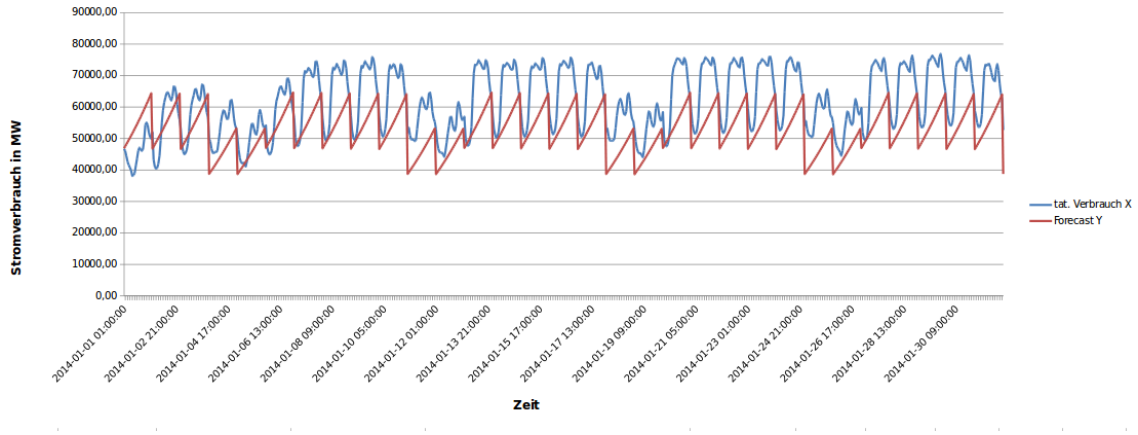


Abbildung 67: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Wo“

tionszeitraums haben die Durchläufe „Fe“ und „Fe_Wo“ ein ähnliches grafisches Bild wie „Wo“. Am Anfang nähern sich bei den beiden Durchläufen „Fe“ und „Fe_Wo“ die Prädiktionswerte mehr dem tatsächlichen Verbrauch an als bei „Wo“. Danach zeigt der Durchlauf „Fe“ eine stärkere Abweichung im Vergleich zu Durchlauf „Fe_Wo“.

Wie Tabelle 86 zeigt, ergibt der Durchlauf „Fe_Wo“ bessere Ergebnisse als die Durchläufe „Fe“ und „Wo“. Die Werte von R-Squared, CV MAE, CV RMSE, MAE und RMSE betragen jeweils 0.54, 0,14, 0,17, 8810,64 und 10540,96. Dementsprechend sind alle drei Ergebnisse nicht befriedigend.

Grafisch - wie auch in den Fehlerkennzahlen ersichtlich - erreicht hier der Durchlauf „Fe _Wo“ das beste Ergebnis.

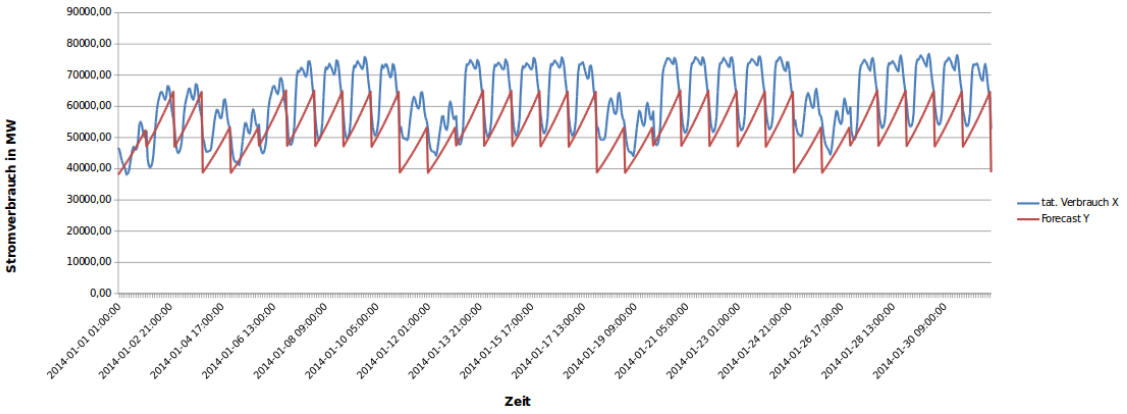


Abbildung 68: Vergleich von tatsächlichem Stromverbrauch und vorhergesagtem Verbrauch. Durchlauf: „Fe_Wo“

Interpretation des zweiten Teilversuchs Der zweite Teilversuch (siehe auch Tabelle 87) enthält die Datensätze von 2013 als Trainingsdaten. In den ersten Tagen des Prädiktionszeitraums sehen die Graphen der Durchläufe „Fe“ und „Fe_Wo“ sehr ähnlich aus. Am Anfang nähern sich die Prädiktionswerte von beiden Durchläufen „Fe“ und „Fe_Wo“ mehr dem tatsächlichen Verbrauch an als bei „Wo“. Danach weist der Durchlauf „Fe“ eine stärkere Abweichung im Vergleich zum Durchlauf „Fe_Wo“ auf.

Wie aus Tabelle 87 zu entnehmen ist, produziert Durchlauf „Fe_Wo“ bessere Ergebnisse als die Durchläufe „Fe“ und „Wo“. Die Werte von R-Squared, CV MAE, CV RMSE, MAE und RMSE betragen jeweils 0.55, 0,15, 0,18, 9473,37 und 11216,39. Die Ergebnisse aller drei Durchläufe sind nicht befriedigend. Aber grafisch - wie auch in den Fehlerkennzahlen ersichtlich - erreicht hier der Durchlauf „Fe_Wo“ das beste Ergebnis.

Vergleich erster Teilversuch und zweiter Teilversuch Von beiden Durchläufen „Fe_Wo“ des ersten Teilversuchs und „Fe_Wo“ des zweiten Teilversuchs erreicht der Durchlauf des ersten Teilversuchs das bessere Ergebnis. Dies kann einerseits grafisch und andererseits auch durch den quantitativen Vergleich der Kennzahlen nachvollzogen werden. Damit ist belegt, dass eine Beachtung des kompletten Trainingssets zu einer besseren Prognosegenauigkeit führt. Es wird auch weiterhin bei der Modellbildung wie ursprünglich definiert das gesamte Trainingsset verwendet.

Support Vector Machine Alle Durchläufe beziehen sich dabei auf die gesamten Trainingsdaten. Die Scripte zur Support Vector Machine sind auf der beigelegten CD im Ordner `svn/Forecast/1.Hypothese/SVM/` abgelegt. Tabelle 88 zeigt, welche Dateien in den Ordnern vorhanden sind.

Die View `CONSUMPTION_TRAINING` enthält die Trainingsdaten, mit denen die Modelle gebildet werden. Um das Trainingsset für den Durchlauf „Arbeitswoche/Feiertage + Wochenen-

Datei	Inhalt
build_and_forecast_svm_yr_qt _mth_dom_dow_hod_weekend _workday.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
build_and_forecast_svm _yr_qt_mth_dom_dow_hod _weekend.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
build_and_forecast_svm _yr_qt_mth_dom_dow_hod _workday.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
weekend_workday_g0 _001#c1000.xlsx	Mehrere Dateien nach dem Muster (Durchlaufart)_g(gamma-Wert mit Unterstrich als Komma)#c(C-Wert mit Unterstrich als Komma).xlsx; Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 mitsamt Fehlerkennzahlen
Diagramm.xlsx	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 88: Relevante Dateien für die Support Vector Machine

de“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend_workday.sql` das folgende SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WEEKEND" as
      ATTRIBUTE7, "WORKDAY" as ATTRIBUTE8 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 69: SQL-Statement für den gesamten Trainingsdatensatz bei SVM

Um das Trainingsset für den Durchlauf „Wochenende“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend.sql` diese SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WEEKEND" as
      ATTRIBUTE7 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 70: SQL-Statement für den Trainingsdatensatz bezogen auf das Wochenende bei SVM

Um das Trainingsset für den Durchlauf „Arbeitswoche/Feiertage“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_workday.sql` folgende SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WORKDAY" as
      ATTRIBUTE7 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 71: SQL-Statement für den Trainingsdatensatz bezogen auf die Arbeitswoche und Feiertage bei SVM

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, wird folgendermaßen vorgegangen: Es erfolgt die Ausführung des für den Durchlauf benötigten Scripts in SAP HANA. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Diese Daten müssen anschließend exportiert und in die entsprechende Spalte der Datei `Januar2014.xlsx` eingetragen werden. Die Datei `Januar2014.xlsx` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. Des Weiteren sind hier auch die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten.

In der folgenden Tabelle erfolgt die Gegenüberstellung der verschiedenen Parameterpaare, die bei den Durchläufen entstehenden, und entsprechenden Fehlerkennzahlen.

	Anpassungsparameter		**Fehlerkennzahlen**				
Durchlauf	γ	**C**	$\mathbf{R^2}$	**MAE**	**RMSE**	**CV_MAE**	**CV_RMSE**
WE	0,01	1000	0,62	8299,08	9815,41	0,13	0,16
WD+WE	0,01	1000	0,71	74074,97	74343,04	1,20	1,21
WD	0,01	100	0,51	9409,81	11270,41	0,15	0,18
WD	0,01	1000	0,68	74832,26	75085,68	1,22	1,22
WD+WE	0,001	1000	0,53	9856,44	11845,94	0,16	0,19
WD	0,001	1000	0,53	9856,44	11845,94	0,16	0,19

Tabelle 89: Durchläufe SVM für die 1. Hypothese

Tabelle 89 zeigt die Ergebnisse der drei Versuche. Zuerst wird ein Durchlauf in Bezug auf das Wochenende (WE) und die Zeit (H) durchgeführt. Es werden die als optimal ermittelten Parameter γ und C aus der Basishypothese wiederverwendet. Abbildung 72 zeigt das Ergebnis.

Der nächste Teildurchlauf bezieht sich auf die Betrachtung von Wochenende (WE), Werktagen (WD) und Zeit (H).In diesem Fall werden zwei Durchläufe gemacht, wobei der Parameter C in beiden gleich unverändert bei $C = 1000$ liegt. Dagegen wird beim Parameter γ je nach Durchlauf eine Anpassung vorgenommen. Der erste Durchlauf wird mit folgenden Parametern durchgeführt: $\gamma = 0.01$ und $C = 1000$. Laut ausgerechneten Messkriterien in der Abbildung 89 liegt R-Squared in diesem Fall bei 0,71. Bei Betrachtung der Visualisierung (siehe Abbildung 73) wird deutlich, dass das Modell durch die Verschiebung

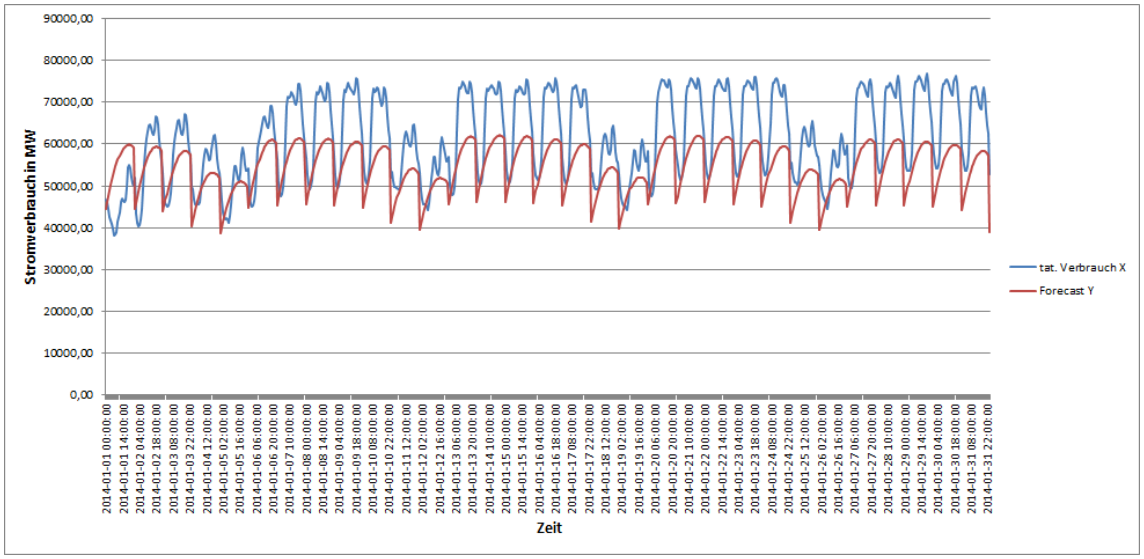


Abbildung 72: Support Vector Machine mit $\gamma = 0.01$ und $C = 1000$

nach unten nicht ausreichend genau ist. Auch die Fehlerkennzahlen CV (MAE) 72074,97 und CV (RMSE) 74343,04 sind hoch. Das Modell wird als nicht ausreichend eingestuft.

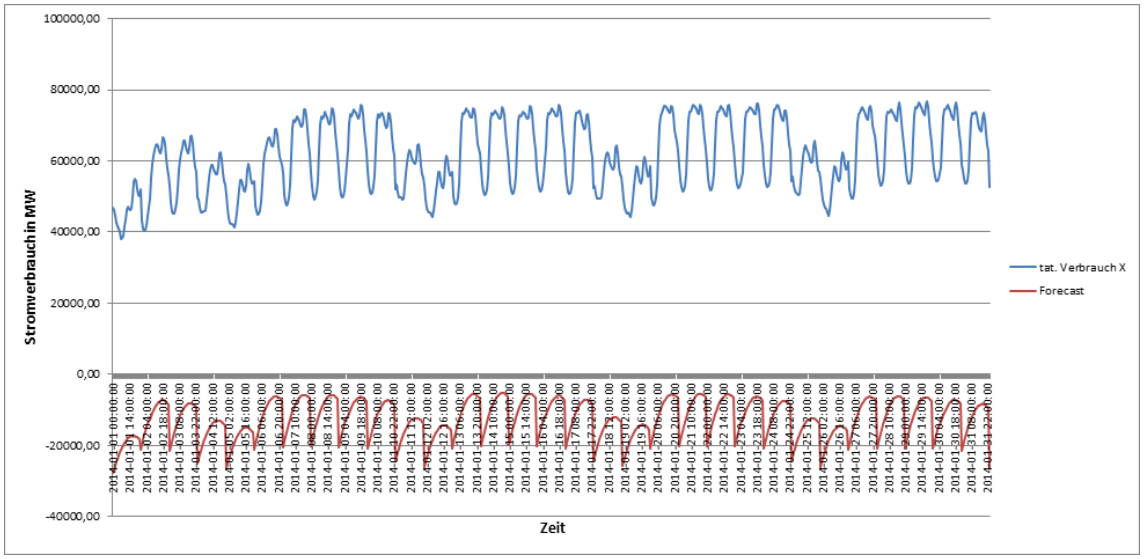


Abbildung 73: Support Vector Machine, WE/WD mit $\gamma = 0.01$ und $C = 1000$

Im 2. Durchlauf wird der Parameter γ von 0,01 auf 0,001 gesetzt. Die Messkriterien CV (MAE) / CV (RMSE) sind um fast 60000 Einheiten verbessert (siehe Tabelle 89). Anhand der Visualisierung der Ergebnisse aus dem 2. Durchlauf 74 ist ersichtlich, dass die Anpassung des Parameters γ die Funktion beeinflusst und das Modell zu den genaueren Ergebnissen bringt.

Der nächste Teildurchlauf bezieht sich auf die Betrachtung von Werktagen (WD) und Zeit (H). In diesem Fall werden 3 Durchläufe durchgeführt (siehe Tabelle 89). Bei den

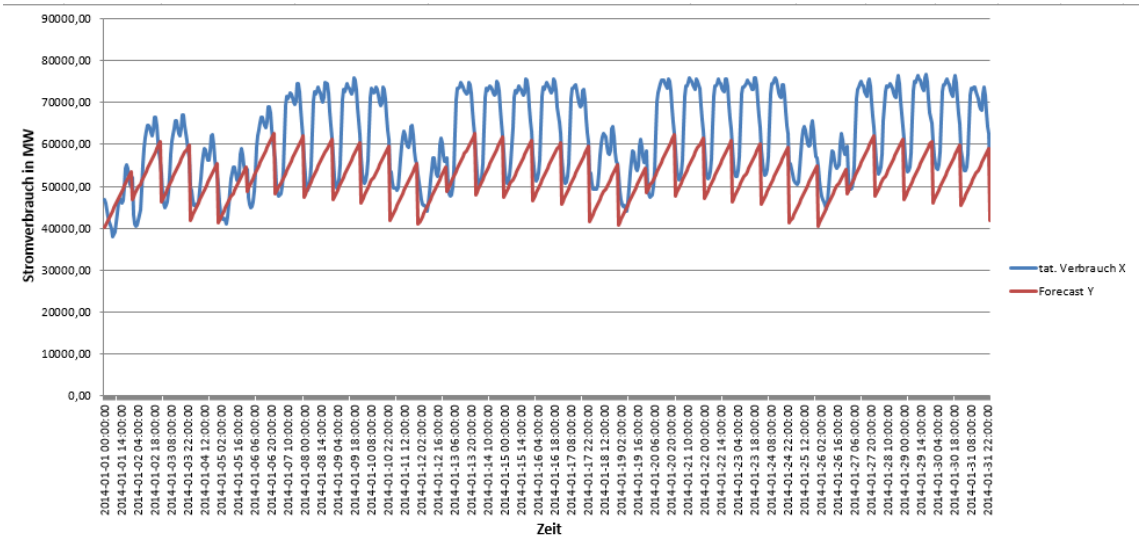


Abbildung 74: Support Vector Machine, WE/WD mit $\gamma = 0.001$ und $C = 1000$

ersten beiden Prognosen wird der Parameter γ konstant bei 0,01 gehalten, während der Parameter C angepasst wird. Beim 1. Test (siehe Abbildung 75) liegt C bei 100. Daraus ergeben sich folgende Werte bei den Fehlerkennzahlen: R-Squared = 0,5, CV (MAE) = 9409,81, CV (RMSE) = 11270,41. Beim 2. Test (siehe Abbildung 76) liegt C bei 1000. Dementsprechend ändern sich die Werte der Fehlerkennzahlen R-Squared = 0,68, CV (MAE) = 74832,26, CV (RMSE) = 75085,68. Diese Änderung des Parameters C bringt zwar bessere Ergebnisse bezüglich der Fehlerkennzahl R-Squared, allerdings wird bei der Analyse der Visualisierung von Prognosen deutlich, dass das Modell bei $C = 100$ genauere Ergebnisse liefert.

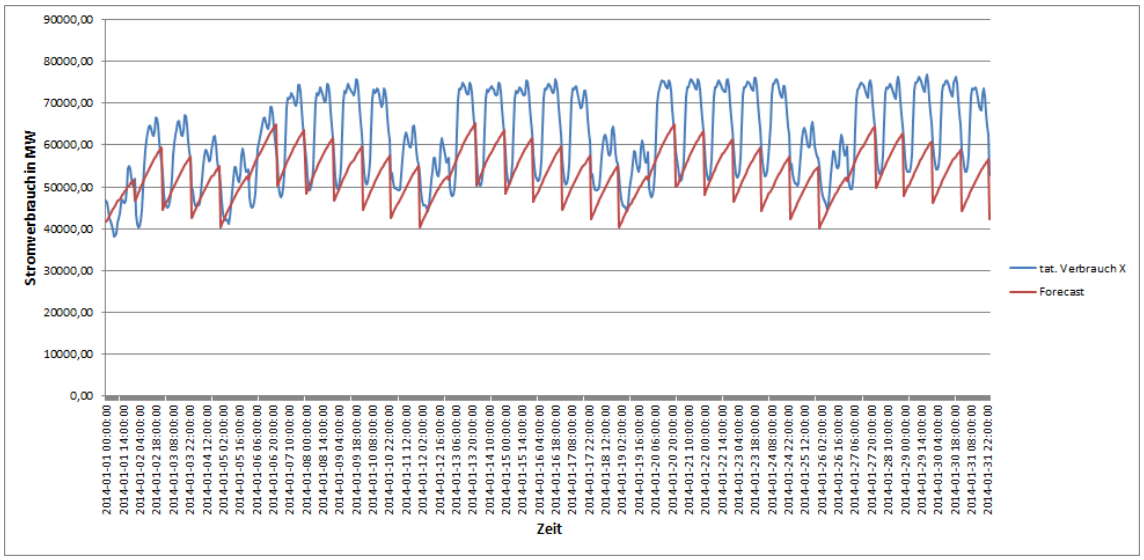


Abbildung 75: Support Vector Machine, WD mit $\gamma = 0.01$ und $C = 100$

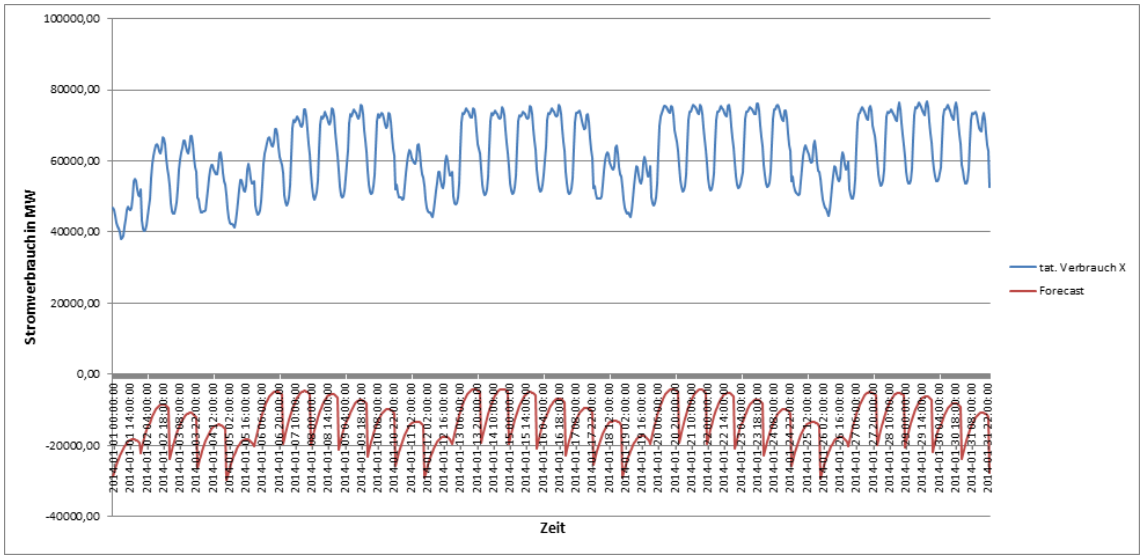


Abbildung 76: Support Vector Machine, WD mit $\gamma = 0.01$ und $C = 1000$

Beim letzten Durchlauf wird eine Anpassung des Parameters γ von 0,01 auf 0,001 durchgeführt. Der CV MAE liegt in dem Fall bei 0,16, was im Vergleich zum Testdurchlauf mit $\gamma = 0,01$ und $C = 1000$ deutlich besser ist. Die Analyse der Visualisierung zeigt dementsprechend auch einen besseren Kurvenverlauf (siehe Abbildung 77).

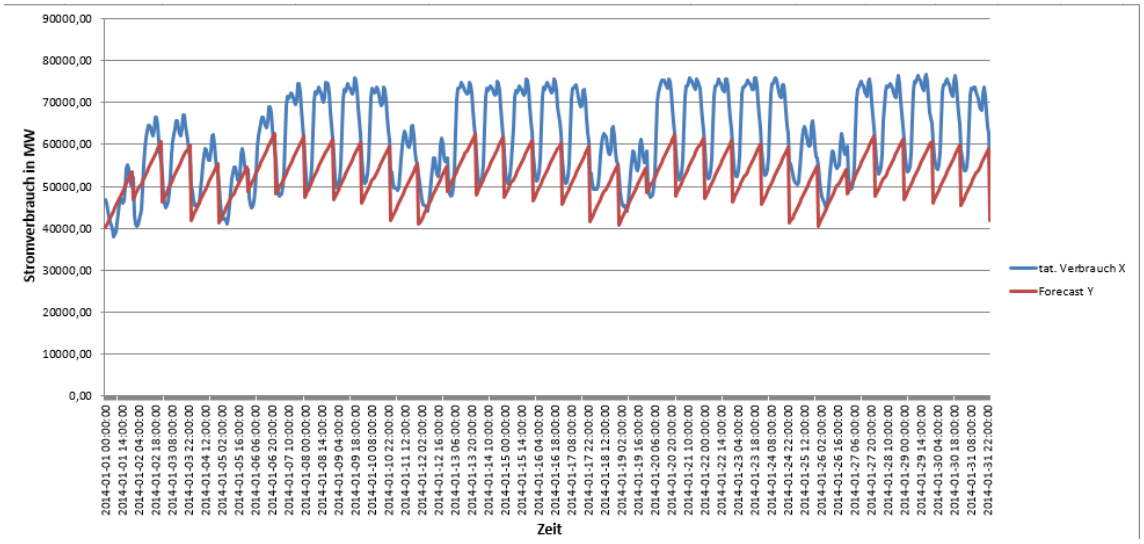


Abbildung 77: Support Vector Machine, WD mit $\gamma = 0.001$ und $C = 1000$

Es ergibt sich, dass der erste Durchlauf unter Betrachtung des Wochenendes *WE* und der Zeit *H* mit den Eingabeparametern $\gamma = 0,01$ und $C = 1000$ zum besten Ergebnis führt.

Zusammenfassung der Ergebnisse Es erfolgt ein Vergleich der besten Ergebnisse der im Wettbewerb zueinander stehenden Algorithmen multiple Lineare Regression, exponenti-

elle Regression und Support Vector Machine. Alle Ergebnisse sind in Bezug zur bereits definierten ersten Hypothese (siehe auch Kapitel 3.2). Vorab wird festgestellt, dass die Algorithmen am besten operieren, wenn alle verfügbaren Trainingsdaten (im Zeitraum von 2009 bis 2013) für die Modellbildung eingebunden werden. Das ergeben entsprechende Testläufe exponentiellen Regression, die die Modellbildung für den Zeitraum 2013 und für den Zeitraum 2009 und 2013 in den Vergleich stellen.

Die folgenden drei Durchläufe beziehen sich den besten Durchlauf des jeweiligen Algorithmus.

Die Berechnung mit der multiplen linearen Regression von (Durchlauf Fe_Wo) ergibt diese Fehlerkennzahlen: R-Squared = 0,54, MAE = 7585,87, RMSE = 9165,80, CV MAE = 0,12 und CV RMSE = 0,15.

Die Werte der Fehlerkennzahlen der exponentiellen Regression von (Durchlauf Fe_Wo) mit allen Datensätzen von 2009-2013 lauten R-Squared = 0,54, MAE = 8810,64, RMSE = 10540,96, CV MAE = 0,14 und CV RMSE = 0,17.

Die Ergebnisse der Support Vector Machine mit den Anpassungsfaktoren $\gamma = 0,01$ und $C = 1000$ hingegen sind am genauesten, wenn nur die Wochenenden in die Berechnung mit einbezogen werden. Die Fehlerkennzahlen betragen hier R-Squared = 0,62, MAE = 8299,08, RMSE = 9815,41, CV MAE= 0,13 und CV RMSE = 0,16.

Im Vergleich der drei Ergebnisse wird deutlich, dass die multiple lineare Regression das beste Ergebnis liefert.

12.5 2. Hypothese

Die zweite Hypothese lautet, dass die Prädiktionen für den Stromverbrauch, angereichert mit der jeweiligen Durchschnittstemperatur der Stunde, genauere Ergebnisse liefern als die erste Hypothese (siehe Kapitel 12.4). Die Datenbasis für diese Hypothese besteht aus den historischen Energieverbrauchsdaten der Entso-E vom 01.01.2009 bis zum 31.12.2013. Zusätzlich werden diese Daten mit der jeweiligen gemessenen Durchschnittstemperatur des Deutschen Wetterdienstes angereichert. Außerdem werden die Informationen aus der ersten Hypothese (Angaben zu Feiertagen sowie Wochenenden) genutzt. Um hier zu differenzieren, werden zunächst alle Datenfelder (Durchschnittstemperatur, Feiertage und Wochenende) evaluiert. Anschließend wird evaluiert, ob die Nutzung ausschließlich der Temperatur (ohne Feiertage sowie Wochenenden) bessere Ergebnisse liefert. Im folgenden Abschnitt erfolgt zunächst die Beschreibung des Datenflusses. Die Ergebnisse zu den einzelnen Algorithmen werden im darauffolgenden Abschnitt dokumentiert.

Datenfluss Der Datenfluss für die Energieverbrauchsdaten verhält sich analog zu den Ausführungen in den Kapiteln 12.3.1 und 12.4. Es folgt die Erläuterung, wie sich die Temperaturdaten zusammenstellen. Die Quelltabelle für die Temperaturdaten ist `OLIMP.DWD_Weather_AirTemp`. Die Datenfelder sind in Tabelle 90 zu finden.

TimeStamp	Air_Temp	Lat	Lon
01.01.2009 00:00:00	-7,61	8,7892	13,6289
01.01.2009 00:00:00	-2,3	3,4775	9,8956
...	...	...	

Tabelle 90: Tabelle `OLIMP.DWD_WEATHER_AIRTEMP`

In der Spalte `TimeStamp` ist der Zeitpunkt der Messung enthalten. Die Spalte `Air_Temp` enthält die gemessene Temperatur. In den Spalten `Lat` und `Lon` sind die GPS-Koordinaten der Wetterstationen enthalten. Es liegen also pro Stunde mehrere Messwerte vor, die in Wetterstationen in ganz Deutschland erfasst worden sind. Mit Hilfe dieser Tabelle wird zunächst eine View angelegt, welche sicherstellt, dass für den gesamten Zeitraum vom 01.01.2009 00:00:00 Uhr bis zum 31.01.2014 23:00:00 Uhr stündliche Einträge vorhanden sind. Der SQL-Code für diese View ist in Listing 78 zu finden.

Hierbei wird mit einem Left Outer Join der beiden Tabellen `_SYS_BI.M_TIME_DIMENSION` und `OLIMP.DWD_Weather_AirTemp` sichergestellt, dass für jeden stündlichen Zeitpunkt zwischen 01.01.2009 00:00:00 und ca. Juli 2014 ein Dateneintrag vorhanden ist. Zusätzlich wird in dieser View bereits der Mittelwert aller Temperaturen pro Zeitpunkt gebildet. Im nächsten Schritt werden nun fehlende Temperaturdaten ergänzt. Dies wird durch die Prozedur in Listing 79 realisiert.

Die Prozedur durchsucht die zuvor erstellte View `View_Airtemp_All_Dirty` (siehe Listing

```
1  CREATE VIEW "OLIMP"."VIEW_AIRTEMP_ALL_DIRTY" ( "ID",
2              "AIRTEMP" ) AS SELECT
3              (t.HOUR_COUNT-8783) as ID,
4              (avg(c."Air_Temp")) as AIRTEMP
5  FROM "_SYS_BI"."M_TIME_DIMENSION" t
6  LEFT OUTER JOIN "OLIMP"."OLIMP::DWD_Weather_AirTemp" c ON c."TimeStamp" = t.
       DATETIMESTAMP
7  WHERE t.YEAR_INT >= 2009
8  AND t.YEAR_INT <= 2014
9  GROUP by t.HOUR_COUNT-8783
10 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 78: View `VIEW_AIRTEMP_ALL_DIRTY`

78) nach `NULL`-Einträgen in der Spalte `AIRTEMP`. Wenn dies der Fall ist, wird der vorige Zeilenwert von `AIRTEMP` eingefügt, welcher nicht `NULL` war. Zusätzlich erstellt die Prozedur eine neue Tabelle `PRE.AIRTEMP_ALL_CLEAN`, in der sich die bereinigten Temperaturwerte für jeden Zeitpunkt zwischen 01.01.2009 00:00:00 und ca. Juli 2014 befinden. In Tabelle 91 ist ein Auszug der Tabelle ersichtlich.

ID	AIRTEMP
1	-4,59
2	-4,48
...	...

Tabelle 91: Tabelle `PRE.AIRTEMP_ALL_CLEAN`

Die Spalte `ID` ist dabei die fortlaufende stündliche Repräsentation der Zeit. Die Spalte `AIRTEMP` ist die Durchschnittstemperatur des jeweiligen Zeitpunktes. Mit Hilfe dieser Datentabelle kann nun die passende View für die Algorithmen erstellt werden. Zu diesem Zweck wird die View aus Listing 55 erneut erweitert. Neben den Angaben zu Feiertagen und Wochenenden der ersten Hypothese enthält die View nun auch Angaben über die Durchschnittstemperatur der jeweiligen Stunden. Der SQL-Code der mit den Wetterdaten angereicherten View ist in Listing 80 ersichtlich.

Analog hierzu wurde auch die View für die Testdaten (01.01.14 00:00:00 bis 31.01.2014 23:00:00 Uhr) erweitert. Der SQL-Code hierzu ist in Listing 81 zu finden. Analog zu den bereits erläuterten Angaben zu den Testviews enthält diese View nun Angaben zu Feiertagen, Wochenenden sowie der Durchschnittstemperatur.

Versuchsdurchführung Die Vorhersagen für die Evaluation dieser Hypothese teilen sich in zwei Versuche auf. Mit Hilfe der View `CONSUMPTION_TRAINING` (siehe Listing 80) werden Modelle mit den entsprechenden Algorithmen gebildet. Dabei werden folgende Versuche betrachtet:

1. Prädiktion mit Angaben zur Temperatur

```
1  CREATE PROCEDURE OLIMP.CLEAN_AIRTEMP_CONSUMPTION LANGUAGE SQLSCRIPT
2  AS
3          CURSOR c_cursor1 FOR SELECT * FROM "OLIMP"."VIEW_AIRTEMP_ALL_DIRTY";
4          v_prev_airtemp DOUBLE;
5
6  BEGIN
7          v_prev_airtemp := 0;
8          OPEN c_cursor1();
9
10         call OLIMP.DROP_TABLE_IF_EXISTS('AIRTEMP_ALL_CLEAN', 'PRE');
11         CREATE COLUMN TABLE PRE.AIRTEMP_ALL_CLEAN(ID INTEGER NOT NULL,
               AIRTEMP DOUBLE NOT NULL);
12
13         IF c_cursor1::ISCLOSED
14         THEN
15                 CALL ins_msg_proc('WRONG: cursor not open');
16         ELSE
17                 FOR cur_row as c_cursor1 DO
18                         IF cur_row.AIRTEMP IS NULL
19                         THEN
20                                 SELECT AIRTEMP INTO v_prev_airtemp FROM "
                                       OLIMP"."VIEW_AIRTEMP_ALL_DIRTY" WHERE ID
                                       < cur_row.ID AND AIRTEMP IS NOT NULL
                                       ORDER BY ID DESC LIMIT 1;
21
22                                 CALL ins_msg_proc('Airtemp for ID ' ||
                                       cur_row.ID || ' is null, prev: ' ||
                                       v_prev_airtemp);
23                                 INSERT INTO PRE.AIRTEMP_ALL_CLEAN VALUES(
                                       cur_row.ID, v_prev_airtemp);
24                         ELSE
25                                 INSERT INTO PRE.AIRTEMP_ALL_CLEAN VALUES(
                                       cur_row.ID, cur_row.AIRTEMP);
26                                 v_prev_airtemp := 0;
27                         END IF;
28                 END FOR;
29                 CLOSE c_cursor1;
30         END IF;
31
32 END
```

Abbildung 79: Prozedur zum Bereinigen der Temperaturdaten

2. Prädiktion mit Angaben zur Temperatur sowie zu Feiertagen und Wochenenden

Mit Hilfe des ersten Versuchs soll zunächst der Einfluss der Temperatur auf die Prädiktion evaluiert werden. Im zweiten Versuch werden anschließend die Angaben zu Feiertagen und Wochenenden aus der ersten Hypothese mit in die Prädiktionsbildung eingebunden.

Multiple Lineare Regression Alle Durchläufe zur multiplen linearen Regression können auf der beigelegten CD im Ordner `svn/Forecast/2.Hypothese/MultipleLineareRegression/` gefunden werden. Im Ordner `Nur_Temperatur` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz für Temperaturdaten. Im

```
1  CREATE VIEW "PAL"."CONSUMPTION_TRAINING" ( "ID",
2              "CONSUMPTION",
3              "HOUR_COUNT",
4              "HOUR_OF_DAY",
5              "DAY_OF_WEEK",
6              "DAY_OF_MONTH",
7              "MONTH",
8              "QUARTER",
9              "YEAR",
10             "WORKDAY",
11             "AIRTEMP",
12             "WEEKEND" ) AS SELECT
13             c.ID-1 as ID,
14             c.CONSUMPTION as CONSUMPTION,
15             c.ID as HOUR_COUNT,
16             t.HOUR_INT as HOUR_OF_DAY,
17             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
18             t.DAY_INT as DAY_OF_MONTH,
19             t.MONTH_INT as "MONTH",
20             t.QUARTER_INT as "QUARTER",
21             t.YEAR_INT as "YEAR" ,
22             t.WORKDAY as "WORKDAY",
23             w."AIRTEMP" as "AIRTEMP",
24             CASE WHEN t.DAY_OF_WEEK_INT = 5
25  or t.DAY_OF_WEEK_INT = 6
26  THEN 1
27  ELSE 0
28  END as "WEEKEND"
29  FROM "PRE"."CONSUMPTION_ALL_CLEAN" as c JOIN "_SYS_BI"."M_TIME_DIMENSION" as
        t ON c.ID+8783 = t."HOUR_COUNT"
30  LEFT OUTER JOIN "PRE"."AIRTEMP_ALL_CLEAN" as w ON c."ID" = w."ID"
31  ORDER BY ID ASC WITH READ ONLY
```

Abbildung 80: Erweiterte View `CONSUMPTION_TRAINING`

Ordner `Temperatur_Feiertage_und_Wochenende` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz für Temperatur, Feiertage und Wochenenden. Im Ordner `Split` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz für Temperaturdaten. In diesem Durchlauf soll festgestellt werden, ob eine Teilung der Trainingsdaten in Arbeitswochen und Wochenenden bessere Prädiktionsergebnisse liefert. Dementsprechend finden sich im Ordner `Split` die Unterordner `Woche` und `Wochenende`. Im Anschluss werden die Prädiktionen dieser beiden Durchläufe zusammengefügt. Die Ergebnisse hierzu sind im Ordner `Zusammengefuegt` zu finden. In allen diesen Ordnern sind folgende Dateien vorhanden:

Um das Trainingsset für die Durchläufe einzuschränken, werden in der jeweiligen Modellbildung die SQL-Codes aus den Listings 82 und 83 verwendet.

Für die im obigen Abschnitt erwähnte Trennung von Arbeitstagen und Wochenenden werden zwei Views erstellt, welche ausschließlich Trainingsdaten für Wochenenden (`CONSUMPTION_TRAINING_WEEKEND` und Arbeitstage (`CONSUMPTION_TRAINING_WEEK` enthal-

```
CREATE VIEW "PAL"."CONSUMPTION_FORECAST" ( "ID",
            "HOUR_COUNT",
            "HOUR_OF_DAY",
            "DAY_OF_WEEK",
            "DAY_OF_MONTH",
            "MONTH",
            "QUARTER",
            "YEAR",
            "WORKDAY",
            "AIRTEMP",
            "WEEKEND" ) AS SELECT
            t.HOUR_COUNT-8783-43824-1 as ID,
            t.HOUR_COUNT-8783 as HOUR_COUNT,
            t.HOUR_INT as HOUR_OF_DAY,
            t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
            t.DAY_INT as DAY_OF_MONTH,
            t.MONTH_INT as "MONTH",
            t.QUARTER_INT as "QUARTER",
            t.YEAR_INT as "YEAR",
            t.WORKDAY as "WORKDAY",
            w.AIRTEMP as "AIRTEMP",
            CASE WHEN t.DAY_OF_WEEK_INT = 5
or t.DAY_OF_WEEK_INT = 6
THEN 1
ELSE 0
END as "WEEKEND"
FROM "_SYS_BI"."M_TIME_DIMENSION" as t
LEFT OUTER JOIN "PRE"."AIRTEMP_ALL_CLEAN" as w on t.HOUR_COUNT-8783 = w."ID"
WHERE t.DATETIMESTAMP >= '2014-01-01 00:00:00'
AND t.DATETIMESTAMP <= '2014-01-31 24:00:00'
ORDER BY ID ASC WITH READ ONLY
```

Abbildung 81: Erweiterte View `CONSUMPTION_FORECAST`

Datei	**Inhalt**
build_modell_mlr.sql	Script zum Erstellen des Modells
build_forecast_mlr.sql	Script zum Erstellen der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellte Fehlerkennzahlen, bezogen auf die Trainingsdaten
forecast.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014.
Diagramm.xlsx	Grafische Darstellung von tatsächlichem und vorhergesagtem Stromverbrauch.

Tabelle 92: Relevante Dateien für die lineare Regression

```
INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "AIRTEMP
    " FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 82: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "WORKDAY" FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 83: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen

ten. Diese Views werden entsprechend als Quelltabellen für die Trainingsdaten genutzt. In der Modellbildung wird dies durch die in Listing 84 und 85 dargestellten SQL-Codes dargestellt.

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
    "WORKDAY", "AIRTEMP" FROM "PAL"."CONSUMPTION_TRAINING_WEEK";
```

Abbildung 84: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen in der Woche

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
    "WORKDAY", "AIRTEMP" FROM "PAL"."CONSUMPTION_TRAINING_WEEKEND";
```

Abbildung 85: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur, Wochenenden und Feiertagen an Wochenenden

Listing 82 zeigt, wie der gesamte Trainingsdatensatz mit den entsprechenden Angaben zur Temperatur in die Modellbildung eingebunden wird. Listing 83 zeigt, wie der gesamte Trainingsdatensatz mit den entsprechenden Angaben zu Temperatur, Feiertagen und Wochenenden in die Modellbildung eingebunden werden. Listing 84 zeigt, wie der gesamte Trainingsdatensatz (nur Arbeitswochen) mit den entsprechenden Angaben zu Temperatur, Feiertagen in die Modellbildung eingeht. Listing 85 zeigt, wie der gesamte Trainingsdatensatz (nur Wochenenden) mit den entsprechenden Angaben zu Temperatur und Feiertagen in die Modellbildung eingebunden wird. Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Zunächst wird das Script `build_modell_mlr.sql` in SAP HANA ausgeführt, welches das mathematische Modell des Algorithmus auf Basis der Trainingsdaten anlegt. Anschließend muss das Script `build_forecast_mlr.sql` ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und aus Übersichtsgründen werden diese Tabellen jeweils einzeln exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten

für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset, enthalten. Die Datei `Diagramm.xlsx` zeigt die grafische Darstellung von tatsächlichem und vorhergesagtem Stromverbrauch. Die Parametereinstellungen werden wie folgt gewählt:

Parameter	Einstellung	Erläuterung
THREAD _NUMBER	8	Modellbildung wird mit 8 Threads durchgeführt.
PMML_EXPORT	1	Gibt an, dass das Modell im PMML-Format vorliegt.
ADJUSTED_R2	1	R-Squared und R-Squared-Adjusted werden berechnet.
VARIABLE _SELECTION	0	Alle in der View vorhandenen Variablen werden zur Modellbildung einbezogen.

Tabelle 93: Parametereinstellungen der Multiplen Linearen Regression (für alle Durchläufe)

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/2.Hypothese/MultipleLineareRegression`. Wie bereits weiter oben erwähnt, finden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Die Fehlerkennzahlen zu beiden Prädiktionen sind in Tabelle 94 zusammengefasst dargestellt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Temp	0,39	8165,57	9984,82	0,13	0,16
Temp_Fe_Wo	0,55	7421,00	8988,63	0,12	0,14
Spl_Wo	0,47	7706,69	9339,93	0,12	0,14
Spl_Woe	0,56	6010,57	7042,07	0,11	0,13
Split-zusam	0,57	7268,98	8804,53	0,11	0,14

Tabelle 94: Ergebnisse der 2. Hypothese

Dabei bezieht sich der Durchlauf „Temp“ auf die ausschließliche Verwendung der Temperaturdaten. Der Durchlauf „Temp_Fe_Wo“ bezieht sich auf die Verwendung der Temperatur sowie den Angaben zu Wochenenden und Feiertagen. Der Durchlauf „Spl_Wo“ bezieht sich auf die Verwendung der Angaben zur Temperatur und Feiertagen in der Arbeitswoche (Montag - Freitag). Der Durchlauf „Spl_Woe“ bezieht sich auf die Verwendung der Angaben zur Temperatur an den Wochenenden (Samstag und Sonntag). Im Durchlauf „Split-zusam“ werden die Ergebnisse der Durchläufe „Spl_Wo“ und „Spl_Woe“ zusammengefügt. In allen oben genannten Durchläufen wird das gesamte Trainingsdatenset verwendet (01.01.2009 00:00:00 - 31.12.2013 23:00:00 Uhr).

Abbildung 86 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Temp“.

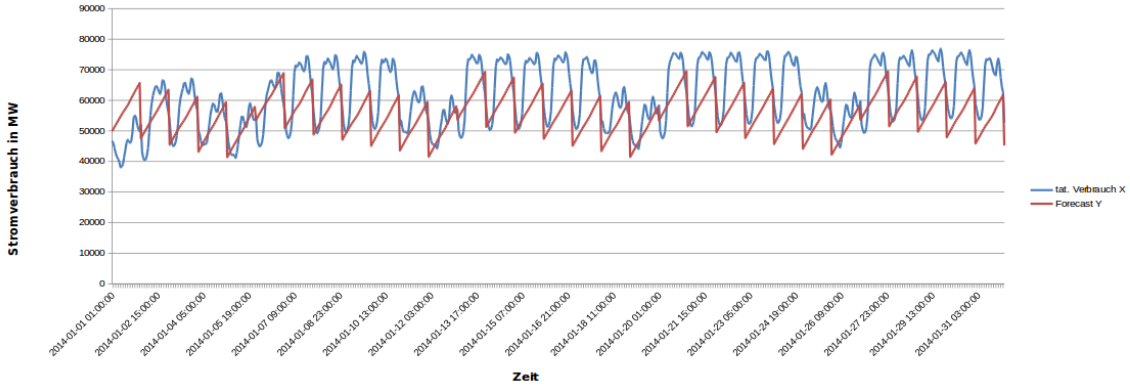


Abbildung 86: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Temp“

Abbildung 87 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Temp_Fe_Wo“.

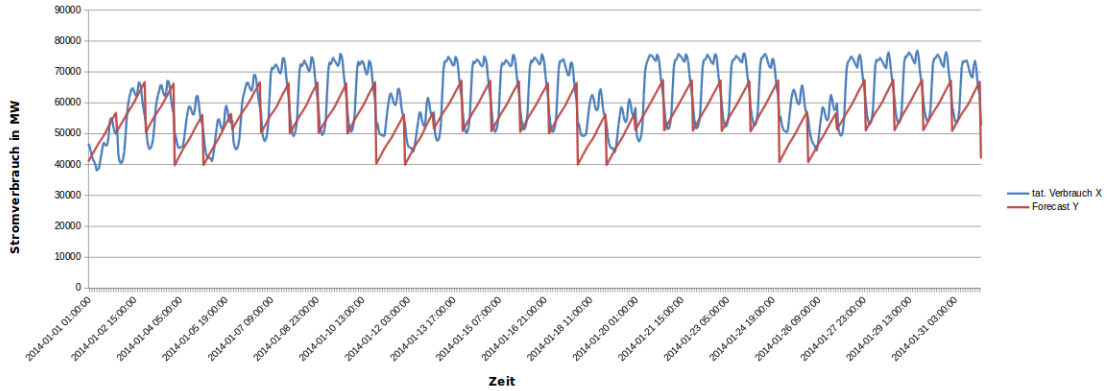


Abbildung 87: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Temp_Fe_Wo“

Abbildung 88 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Spl_Wo“.

Abbildung 89 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Spl_Woe“.

Abbildung 90 zeigt die grafische Zusammenfassung der beiden Durchläufe „Spl_Wo“ und „Spl_Woe“.

Vergleicht man zunächst die Abbildungen 86 und 87 der beiden Durchläufe „Temp“ und „Temp_Fe_Wo“ dann fällt auf, dass insbesondere im Durchlauf „Temp_Fe_Wo“ die Stromverbrauchstiefen verhältnismäßig korrekt modelliert werden, während die Stromverbrauchshöhen zu niedrig vorhergesagt werden. An den Wochenenden müsste hier die gesamte Prädiktion um einen „Sockelbetrag“ angehoben werden: Hier werden die Stromver-

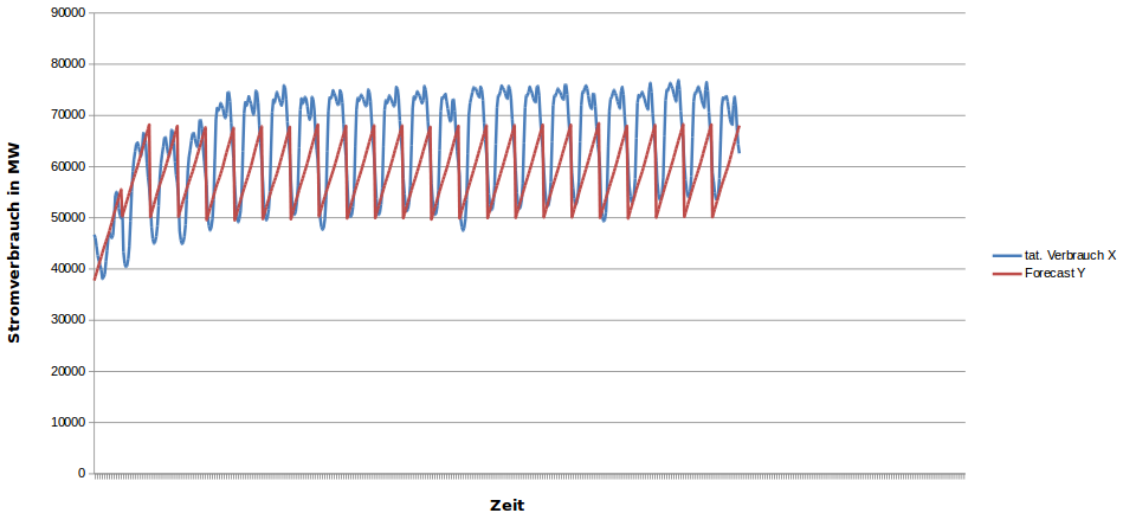


Abbildung 88: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Spl_Wo“

brauchstiefen und Stromverbrauchshöhen zu niedrig vorhergesagt. Beim Durchlauf „Temp“ reißt die Prädiktionskurve teilweise sehr stark aus, so dass hier die Stromverbrauchstiefen und Stromverbrauchshöhen nicht korrekt vorhergesagt werden. Beiden Prädiktionen fehlt die korrekte Modellierung der Stromverbrauchskurve des jeweiligen Tages: Wie in der Datenbasis sowie der 1. Hypothese bereits festgestellt, existiert auch hier lediglich eine Spitze, ohne dass die Schwankungen im Verlaufe eines Tages weiter beachtet werden. Dies wird auch an den Fehlerkennzahlen der beiden Durchläufe in Tabelle 94 deutlich. Der Durchlauf „Temp“ erreicht hier einen R-Squared von 0,39 und einen MAE von 8165,57 Punkten. Der Einfluss der Temperatur scheint also einen vernachlässigbaren Einfluss auf die Modellbildung der multiplen linearen Regression zu haben. Vergleicht man diesen Durchlauf mit den Durchlauf „Temp_Fe_Wo“ wird deutlich, dass hier der Einfluss der Feiertage sowie der Wochenenden stärker ist. Der Durchlauf Temp_Fe_Wo erreicht hier einen R-Squared von 0,55 sowie einen MAE von 7421,00 Punkten. Ebenso liegt der RMSE mit 8988,63 Punkten im Vergleich zum Durchlauf „Temp“ um 996,18 Punkte niedriger. Insgesamt ist jedoch der Einfluss der Temperatur auf die Prädiktionsqualität geringfügiger als erwartet: Der Durchlauf „Fe_Wo“ der ersten Hypothese liefert einen R-Squared von 0,544 Punkten und ist im Vergleich zum Durchlauf „Temp_Fe_Wo“ der zweiten Hypothese mit einem R-Squared von 0,55 Punkten nahezu gleichauf. Auch der Vergleich der MAE‘s der beiden Durchläufe zeigt nur eine geringfügige Verbesserung der Prädiktion durch die Verwendung der Temperatur als Prädikator auf. Der Durchlauf „Fe_Wo“ aus der ersten Hypothese erreicht einen MAE von 7585,87 Punkten, verglichen mit dem Durchlauf „Temp_Fe_Wo“ mit einem MAE von 7421,00 Punkten ist nur eine geringere Verbesserung zu erkennen. Die Einbindung der Temperatur erzielt demnach zwar bessere aber keine signifikanten Verbesserungen der Prädiktionsqualität.

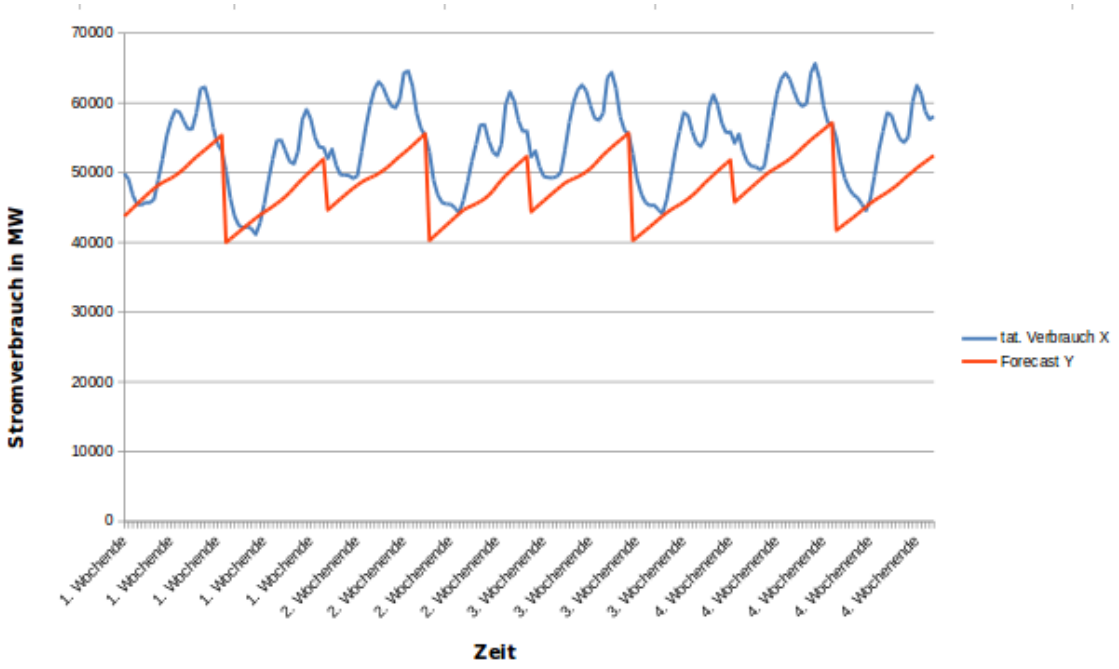


Abbildung 89: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Spl_Woe“

Die beiden Durchläufe „Spl_Wo“ und „Spl_Woe“ teilen die Trainingsmenge jeweils in Arbeitswoche (Montag bis Freitag) und Wochenende (Samstag und Sonntag) und bilden das darauf basierende Modell. Entsprechend produzieren die aus diesen beiden Durchläufen resultierenden Modelle auch nur Vorhersagen für die Wochenenden des Januar 2014 (Durchlauf „Spl_Wo“) und die Arbeitswochen (Durchlauf „Spl_Woe“). Die Grafiken 88 und 89 zeigen die dazu produzierten Ergebnisse. Wie im vorigen Abschnitt bereits erwähnt, werden die Stromverbrauchstiefen im Durchlauf „Spl_Wo“ verhältnismäßig korrekt vorhergesagt, während die Stromverbrauchshöhen zu niedrig prognostiziert werden und die korrekte Modellierung des Tagesverbrauch gänzlich fehlt. Dies gilt ebenso für den Durchlauf „Spl _Woe“: Hier fehlt zusätzlich der bereits erwähnte „Sockelbetrag“, welcher der Präditkion zu einer besseren Qualität verhelfen könnte. Die Ergebnisse dieser beiden Prädiktionen werden im Anschluss wieder zusammengefasst. Dies ist in Abbildung 90 sichtbar und ist in der Tabelle 94 unter dem Durchlauf „Split-zusam“ dokumentiert. Vergleicht man die beiden Einzeldurchläufe miteinander, so wird ersichtlich, dass der Durchlauf „Spl_Woe“ (Wochenende), gemessen an den Fehlerkennzahlen bessere Ergebnisse liefert als der Durchlauf „Spl_Wo“. Insbesondere ist im Durchlauf „Spl_Woe“ der MAE mit 6010,57 Punkten um 1696,12 Punkte niedriger als der Durchlauf „Spl_Wo“. Werden beide Ergebnisse zusammengefügt, so ergibt sich für diesen Durchlauf - gemessen an den Fehlerkennzahlen - das beste Ergebnis der gesamten Versuchsreihe. Der R-Squared liegt hier bei 0,57, der MAE bei 7268,98 Punkten und der RMSE bei 8804,53 Punkten. Insbesondere im Vergleich zu den Durchläufen „Temp“ und „Temp_Fe_Wo“ erzielt der Durchlauf leichte, jedoch keine

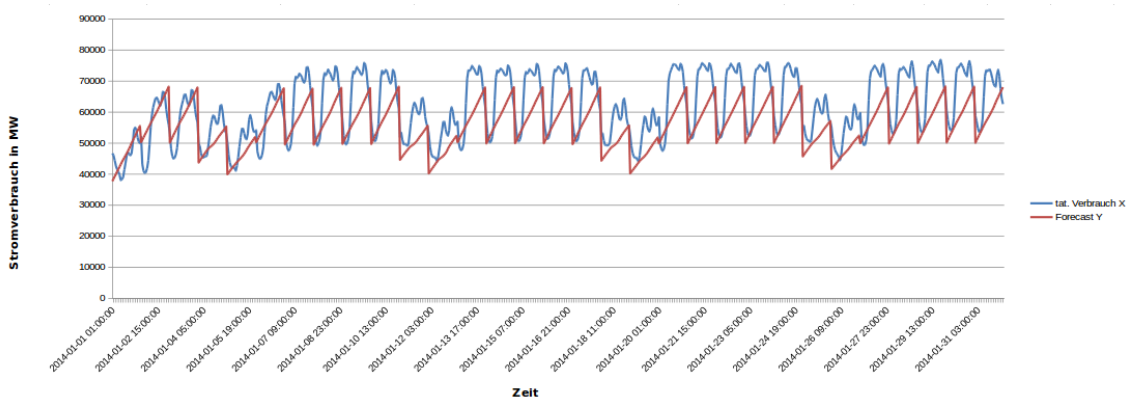


Abbildung 90: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Zusammenfassung der Durchläufe „Spl_Wo“ und „Spl_Woe“.

signifikanten Verbesserungen. Zusammenfassend lässt sich für diese Hypothese aussagen, dass ein Split der Trainingsdaten in Wochenenden und Arbeitswochen zu leichten Verbesserungen der Prädiktionsgenauigkeit führt. Dennoch fehlt allen Prädiktionen der bereits in der ersten Hypothese festgestellte fehlende Sockelbetrag, welcher die Qualität der Prädiktionen nochmals anheben könnte. Insgesamt erzielen keine der fünf durchgeführten Prädiktionen zu adäquate Ergebnisse.

Feldversuche mit der multiplen linearen Regression Da die Ergebnisse der Multiplen Linearen Regression für die erste und zweite Hypothese bisher unbefriedigende Ergebnisse liefert, werden in diesem Abschnitt einige Feldversuche dokumentiert, mit der eine Verbesserung der Ergebnisse erzielt werden könnte. Zunächst soll überprüft werden, ob eine Änderung des Zeitraumes der Trainingsdaten eine Verbesserung des Prädiktionsergebnisses liefert. Hierzu werden folgende Versuche durchgeführt:

1. Trainingsdaten beinhalten lediglich das Jahr 2013
2. Trainingsdaten beinhalten lediglich den Dezember 2013
3. Trainingsdaten beinhalten lediglich den Januar 2013
4. Trainingsdaten beinhalten den Monat Januar der Jahre 2009-2013

Um das Trainingsdatenset einzuschränken, werden in der Modellbildung folgende SQL-Statements genutzt:

```
1  INSERT INTO PAL_MLR_DATA_TBL SELECT TRAIN.NEXTVAL, "CONSUMPTION", "
       HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "
       WORKDAY", "AIRTEMP", "WEEKEND" FROM "PAL"."CONSUMPTION_TRAINING" WHERE "
       YEAR" = 2013 ORDER BY "ID" ASC;
```

Abbildung 91: SQL-Code der Trainingsdaten für das Jahr 2013

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT TRAIN.NEXTVAL, "CONSUMPTION", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "WORKDAY", "AIRTEMP", "
    WEEKEND" FROM "PAL"."CONSUMPTION_TRAINING" WHERE "YEAR" = 2013 AND "
    MONTH" = 12 ORDER BY "ID" ASC;
```

Abbildung 92: SQL-Code der Trainingsdaten für Dezember 2013

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT TRAIN.NEXTVAL, "CONSUMPTION", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "WORKDAY", "AIRTEMP", "
    WEEKEND" FROM "PAL"."CONSUMPTION_TRAINING" WHERE "YEAR" = 2013 AND "
    MONTH" = 01 ORDER BY "ID" ASC;
```

Abbildung 93: SQL-Code der Trainingsdaten für Januar 2013

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT TRAIN.NEXTVAL, "CONSUMPTION", "
    HOUR_OF_DAY", INSERT INTO PAL_MLR_DATA_TBL SELECT TRAIN.NEXTVAL, "
    CONSUMPTION", "HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "WORKDAY", "
    AIRTEMP", "WEEKEND" FROM "PAL"."CONSUMPTION_TRAINING" WHERE "MONTH" = 01
     ORDER BY "ID" ASC;
```

Abbildung 94: SQL-Code der Trainingsdaten für Januar 2009-2013

```
1 DROP SEQUENCE PAL.TRAIN;
2 CREATE SEQUENCE PAL.TRAIN START WITH 0 MINVALUE 0;
```

Abbildung 95: Erstellung der Sequenz zum korrekten durchnummerieren der Zeilen

Listing 91 zeigt dabei, wir die Trainingsdaten für das Jahr 2013 eingeschränkt wird. Listing 92 zeigt, wie das Trainingsdatenset auf den Dezember 2013 eingeschränkt wird. Listing 93 zeigt, wie das Trainingsdatenset auf den Januar 2013 eingeschränkt wird. Listing 94 zeigt, wie das Trainingsdatenset auf den Monat Januar für die Jahre 2009 - 2013 eingeschränkt wird. In jeder dieser SQL-Codes wird für die erforderliche fortlaufende ID (beginnend bei 0) eine zuvor generierte Sequenz verwendet, die für jeden Durchlauf im Anschluss wieder zurückgesetzt wird. Der SQL-Code hierzu ist in Listing 95 ersichtlich. Die fortlaufende ID kann anschließend über den Befehl `TRAIN.NEXTVAL` verwendet werden. Die Fehlerkennzahlen sowie die Skripte der zuvor aufgeführten Durchläufe sind in Tabelle 95 ersichtlich und können auf der beigelegt CD im Ordner `svn/Forecast/2.Hypothese/MultipleLineareRegression/Feldtest` gefunden werden.

Durchlauf	**R-Squared**	**MAE**	**RMSE**	**CV_MAE**	**CV_RMSE**
2013_komplett	0,56	8693,42	10423,99	0,14	0,16
Dez_2013	0,21	9484,81	11485,12	0,15	0,18
Jan_2013	0,57	8278,78	9868,47	0,13	0,16
Jan_2009_2013	0,60	6485,72	7775,87	0,10	0,12

Tabelle 95: Ergebnisse der 2. Hypothese

Abbildung 96 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „2013_komplett“. Dieser Durchlauf erstellt aus den Trainingsdaten für das Jahr 2013 entsprechende Prädiktionen für Januar 2014.

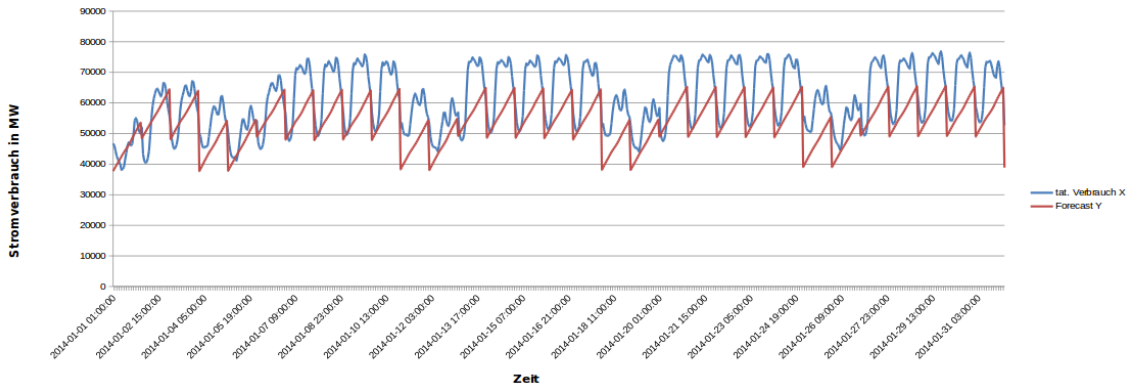


Abbildung 96: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „2013_komplett“

Abbildung 97 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Dez_2013“. Dieser Durchlauf erstellt aus den Trainingsdaten für den Zeitraum Dezember 2013 entsprechende Prädiktionen für Januar 2014.

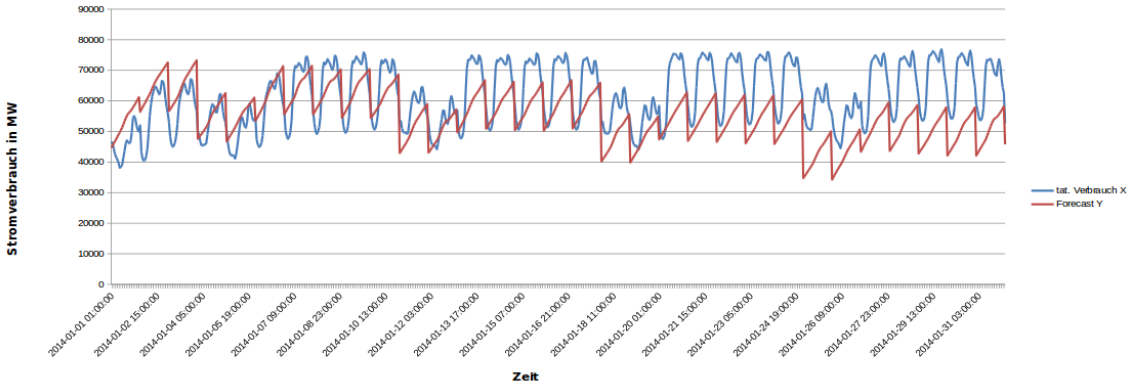


Abbildung 97: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Dez_2013“

Abbildung 98 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Jan_2013“. Dieser Durchlauf erstellt aus den Trainingsdaten für den Zeitraum Januar 2013 entsprechende Prädiktionen für Januar 2014.

Abbildung 99 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten für den Durchlauf „Jan_2009_2013“. Dieser Durchlauf

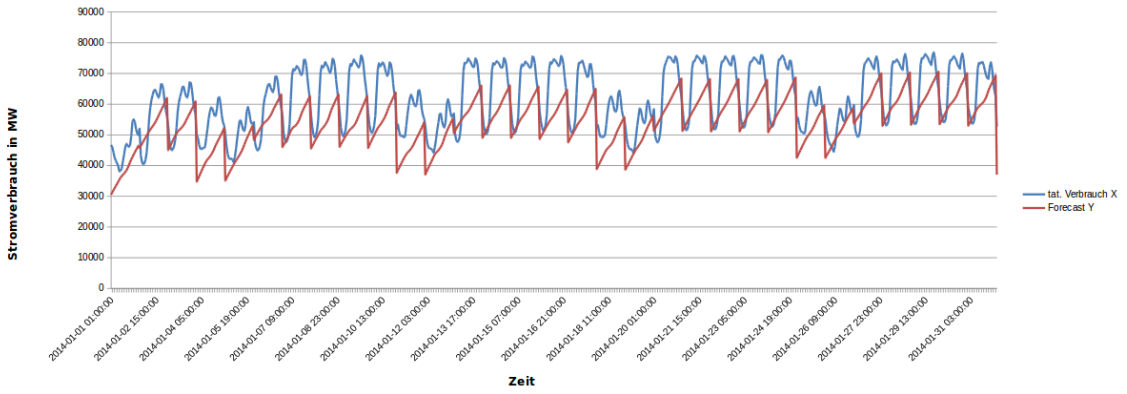


Abbildung 98: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Jan_2013“

erstellt aus den Trainingsdaten für den Zeitraum Januar 2009 - 2013 entsprechende Prädiktionen für Januar 2014.

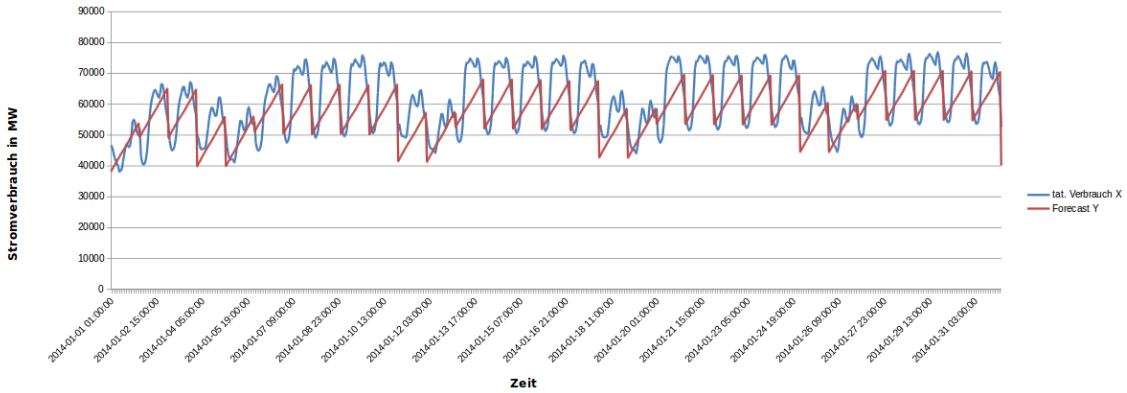


Abbildung 99: Vergleich von tatsächlichem und vorhergesagtem Stromverbrauch. Durchlauf: „Jan_2009_2013“

Bezugnehmend auf die Fehlerkennzahlen der Feldversuche in Tabelle 95 führt lediglich der Durchlauf „Jan_2009_2013“ zu leicht verbesserten Prädiktionsergebnissen, verglichen mit den vorigen Durchläufen dieser Hypothese (siehe Tabelle 94). Der Durchlauf „Jan_2009 _2013“ erzielt hier einen R-Squared von 0,60. Ebenso erzielt dieser Durchlauf beim MAE den niedrigsten Wert mit 6485,72 Punkten. Verglichen mit dem bisher besten Durchlauf dieser Hypothese („Split-zusam“) erzielt der der Durchlauf „Jan_2009_2013“ eine leichte Verbesserung von 783,26 Punkten. Dieses Verhalten spiegelt sich auch im RMSE wieder - hier erzielt der Durchlauf „Jan_2009_2013“ eine Verbesserung von 1028,66 Punkten. Die Prädiktionsgenauigkeit erhöht sich also leicht - aber nicht signifikant, wenn der Algorithmus ausschließlich mit Trainingsdaten der gleichen Periode angereichert wird. Hierzu zeigt auch der Durchlauf „Jan_2013“ mit wesentlich schlechteren Werten im MAE (8278,78 Punkte) und RMSE (9868,47 Punkte), dass mehrere historische Perioden des gleichen

Betrachtungszeitraumes zu besseren Ergebnissen in der Prädiktion führen. In diesem Zusammenhang führt die Betrachtung einer anderen Periode (Durchlauf „Dez_2013“) zu wesentlich schlechteren Prädiktionsergebnissen mit einem MAE von 9484,81 Punkten und einem RMSE von 11485,12 Punkten. Auch der R-Squared dieses Durchlaufes fällt mit 0,21 Punkten besonders schlecht aus. Auch die Modellbildung auf Basis der Trainingsdaten für den Zeitraum des kompletten Jahres 2013 hat zu keine Verbesserung der Prädiktion geführt. Der vergleichbare Durchlauf hierfür ist „Temp_Fe_Wo“, welcher jedoch die Trainingsdaten von 2009 - 2013 in die Modellbildung einbezieht. Insbesondere fällt hier auf, dass der Durchlauf „Temp_Fe_Wo“ einen um 2063,81 verringerten MAE und einen um 1435,36 Punkten verringerten RMSE hat und dadurch eine höhere Prädiktionsqualität hat. Dies schlägt sich auch auf Betrachtung der grafischen Darstellung der Prädiktionen (siehe Abbildung 96, 97, 98 und 99) nieder: Hier fehlt wie in allen vorigen beschriebenen Durchläufen wieder der Sockelbetrag, welcher die Prädiktionsergebnisse weiter verbessern könnte.
Mit diesen Feldversuchen sollte getestet werden, ob eine Veränderung des Zeitraumes der Trainingsdaten auf denen die Modellbildung basiert, zu besseren Ergebnisse führt. Bis auf leichte - aber nicht signifikante - Verbesserungen beim Durchlauf „Jan_2009_2013“ tragen diese Versuche keine weiteren Erkenntnisse bei. Aus diesem Grund wird auch weiterhin der gesamte Trainingsdatensatz für die Evaluation der Hypothesen verwendet.

Exponentielle Regression Alle Durchläufe zur exponentiellen Regression können auf der beigelegten CD im Ordner `svn/Forecast/2.Hypothese/Exponentialregression/` gefunden werden. Im Ordner `Nur_Temperatur` befinden sich die Daten zur Modell- und Prädiktionsbildung für den ersten Versuch mit Temperatur. Im Ordner `Temperatur_Feiertag_Wochentag` befinden sich die Daten zur Modell- und Prädiktionsbildung für den zweiten Versuch mit Temperatur, Feiertag und Wochentag. In allen dieser Ordner sind folgende Dateien vorhanden: Um das Trainingsset für die jeweiligen Durchläufe einzu-

Datei	**Inhalt**
Exponentialregression.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellten Fehlerkennzahlen, bezogen auf die Trainingsdaten
forecast.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014
Diagramm.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 96: Relevaten Dateien für die Exponentiale Regression

schränken, wird in der Datei `Exponentialregression.sql` folgende SQL-Code verwendet:

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "AIRTEMP
    " FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 100: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur

Listing 100 zeigt, wie der gesamte Trainingsdatensatz und die Temperatur in die Modellbildung eingebunden werden.

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "WORKDAY" FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 101: SQL-Code für den gesamten Trainingsdatensatz mit Temperatur-Feiertag-Wochentag

Listing 101 zeigt, wie der gesamte Trainingsdatensatz und die Wochentage, Temperatur und Feiertage in die Modellbildung eingebunden werden.

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Es muss das Script `Exponentialregression.sql` in SAP HANA ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und Übersichtsgründen werden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten.

Ergebnisse des Durchlaufes mit Nur_Temperatur: Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/2.Hypothese/Exponentialregression/Nur_Temperatur`.

Ergebnisse des Durchlaufes mit Temperatue_Feiertag_Wochentag_: Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/2.Hypothese/Exponentialregression/Temperatur_Feiertag_Wochentag`.

Ergebnisse der Durchläufe: Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/2.Hypothese/Exponentialregression`. Wie bereits weiter oben erwähnt finden sich in diesen Ordnern alle relevanten Ergebnisse der Durchläufe. Die Fehlerkennzahlen zu beiden Prädiktionen sind in Tabelle 97 zusammengefasst dargestellt.

Dabei bezieht sich der Durchlauf „Temp“ auf die ausschließliche Verwendung der Temperaturdaten. Der Durchlauf „Temp_Fe_Wo“ bezieht sich auf die Verwendung der Tempera-

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Temp	0,37	9006,00	10906,44	0,15	0,18
Temp_Fe_W0	0,54	8080,12	9770,44	0,13	0,16
Split	...	...	...	...	
Spl_W	0,44	8465,50	10264,29	0,13	0,16
Spl_Wo	0,55	6684,64	7683,15	0,12	0,14
Spl_zusamm	0,55	8005,92	9664,11	0,13	0,16

Tabelle 97: Fehlerkennzahlen der Anwendung des Modells auf die Testdaten

tur, sowie den Angaben zu Wochenenden und Feiertagen. Der Durchlauf „Spl_W“ bezieht sich auf die Verwendung der Angaben zur Temperatur und Feiertagen in der Arbeitswoche (Montag - Freitag). Der Durchlauf „Spl_Wo“ bezieht sich auf die Verwendung der Angaben zur Temperatur an den Wochenenden (Samstag und Sonntag). Im Durchlauf „Splzusamm“ werden die Ergebnisse der Durchläufe „Spl_W“ und „Spl_Wo“ zusammengefügt. In allen oben genannten Durchläufen wird das gesamte Trainingsdatenset verwendet (01.01.2009 00:00:00 - 31.12.2013 23:00:00 Uhr).

Abbildung 102 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädikationswerten für den Durchlauf „Temp“.

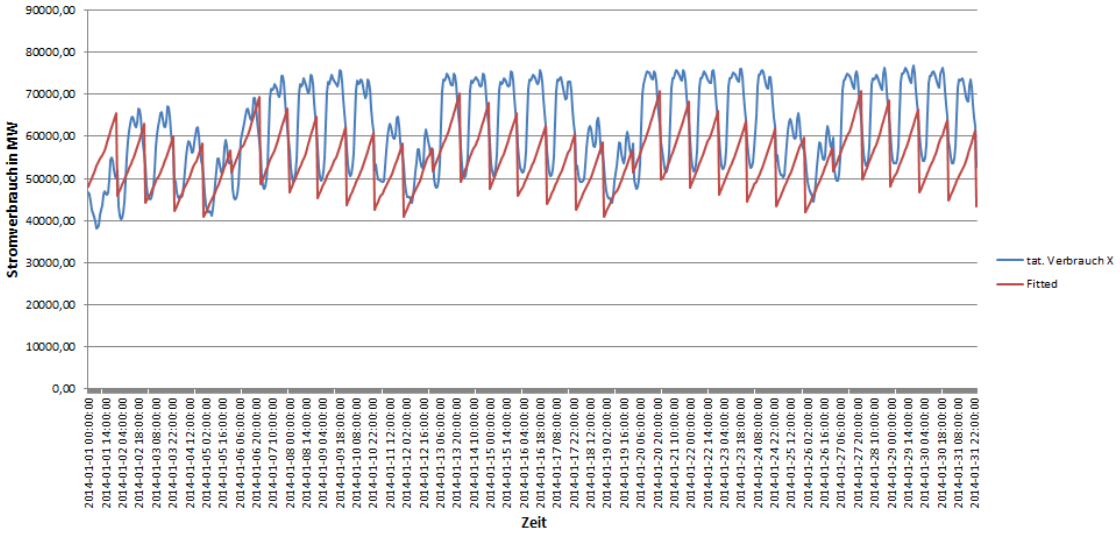


Abbildung 102: Diagramm Vergleich Ist und Forecast. Durchlauf: „Temp“

Abbildung 103 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädikationswerten für den Durchlauf „Temp_Wo_W“.

Abbildung 104 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädikationswerten für den Durchlauf „Spl_W“.

Abbildung 105 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädikationswerten für den Durchlauf „Spl_Wo“.

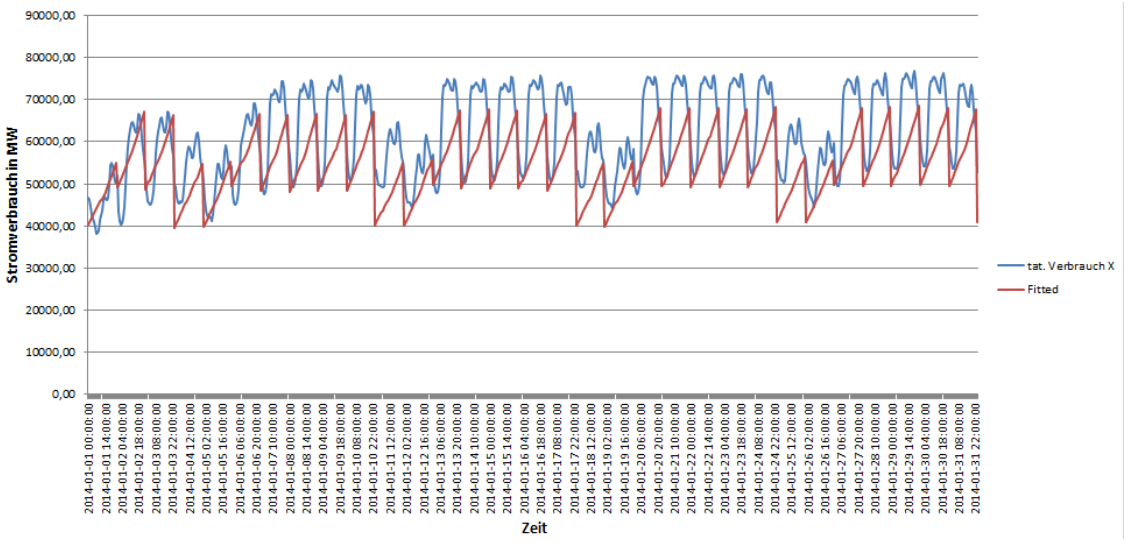


Abbildung 103: Diagramm Vergleich Ist und Forecast. Durchlauf: „Temp_Wo_W“

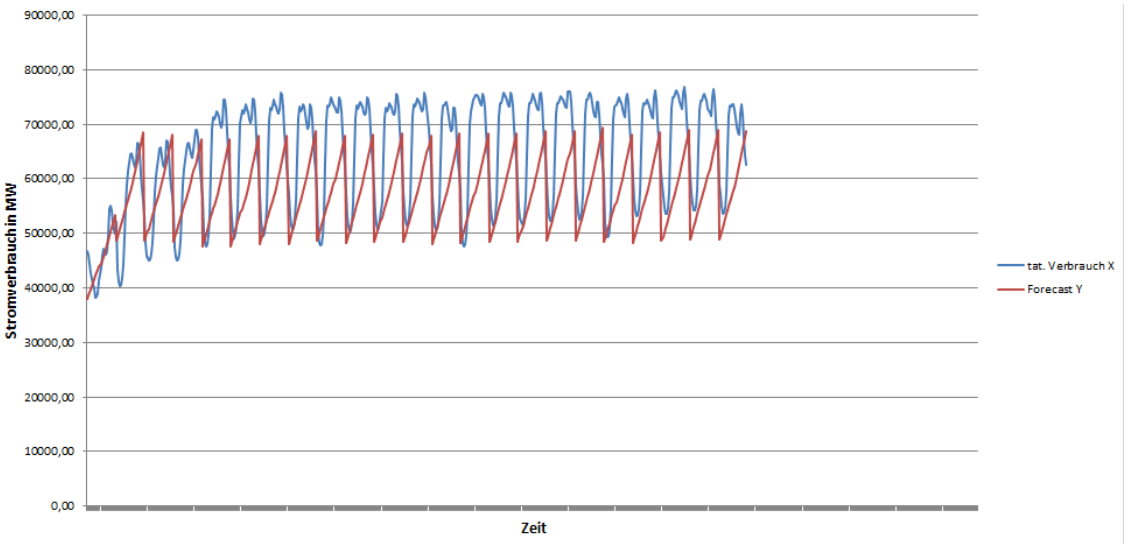


Abbildung 104: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_W“

Abbildung 106 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädikationswerten für den Durchlauf „Spl_zusamm“.

Wie die Grafiken zeigen, haben alle Durchläufe den gleichen Schrumpfung- und Wachstumstrend wie die tatsächlichen Verbräuche, aber es fehlt noch ein „Sockelbetrag“ in allen Prädikationen, welcher die gesamte Prädikation näherer an den tatsächlichen Verbrauch rücken würde. Der Durchlauf „Spl_zusamm“ zeigt eine bessere Grafik im Vergleich zu allen anderen Durchläufen im gesamten Zeitraum.

Die Fehlerkennzahlen der Tabelle 97 zeigen, dass die Ergebnisse aller Durchläufe nicht befriedigend sind. Der Durchlauf „Spl_zusamm“ ergibt einen R-squared 0,55, MAE 8005,92

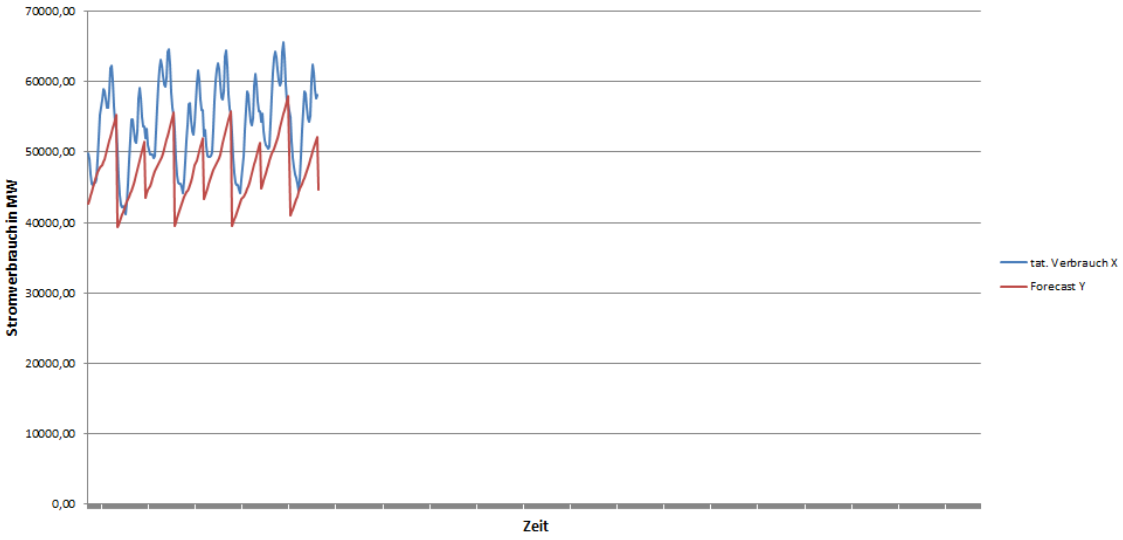


Abbildung 105: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_Wo“

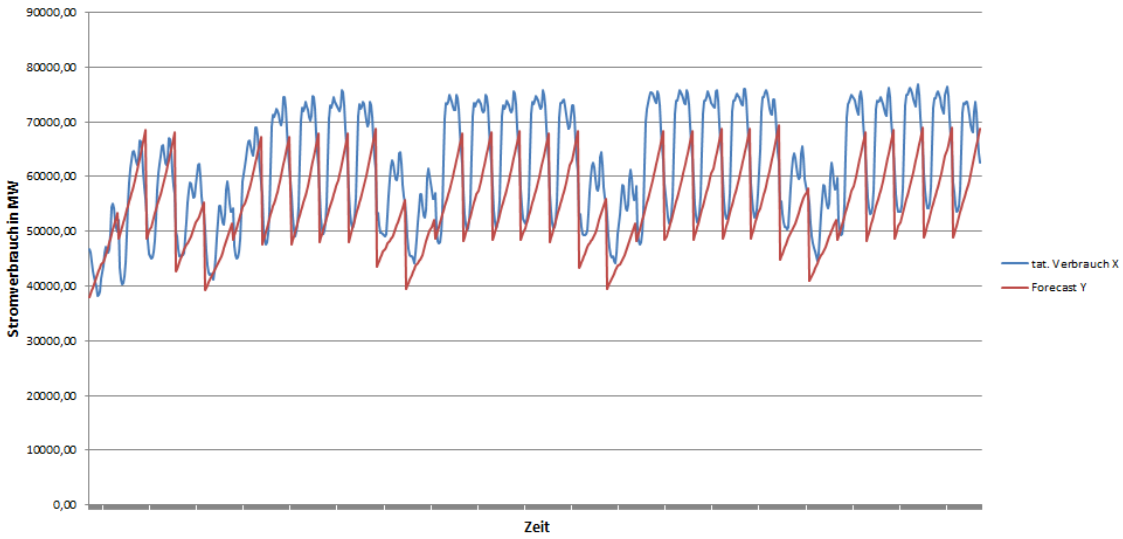


Abbildung 106: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_zusamm“

und RMSE 9664,11. Das beste Ergebnis liefert der Durchlauf „Spl_Wo“ mit den Werten R-Squared = 0,55 sowie MAE = 6684,64 und RMSE = 7683,15.

Aber der Durchlauf „Spl_zusamm“ ergibt ein besseres Diagramm als der Durchlauf „Spl _Wo“. Der Grund dafür ist die Datenmenge, die für die Prädikation genommen werden. Im Durchlauf „Spl_Wo“ werden nur das Wochenende und im Durchlauf „Spl_zusamm“ werden die Verbrauchsdaten aus dem ganzem Monat betrachtet.

Mit Betrachtung des Diagramms und der Werte der Fehlerkennzahlen lässt sich zusammenfassend für diesen Durchlauf sagen, dass der Durchlauf „Spl_zusamm“ das beste Ergebnis liefert. Dieser Durchlauf enthält für diese Hypothese auch die meisten Daten, da für die Prädiktion die Informationen zu Wochentag sowie zum Wochenende genutzt werden.

Dennoch besteht hinsichtlich der Prädiktionsqualität durchaus Steigerungsbedarf.

Support Vector Machine Alle Durchläufe zur Support Vector Machine können auf der beigelegten CD im Ordner `svn/Forecast/2.Hypothese/SVM/` gefunden werden. In diesen Ordnern sind folgende Dateien vorhanden:

Datei	Inhalt
build_and_forecast_svm_yr_qt_mth_dom _dow_hod_temp.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
build_and_forecast_svm_yr_qt_mth_dom _dow_hod_weekend_temp.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
temp_g0_001#c1000.xlsx	Mehrere Dateien nach dem Muster (Durchlaufart)_g(gamma-Wert mit Unterstrich als Komma)#c(C-Wert mit Unterstrich als Komma).xlsx; Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 mitsamt Fehlerkennzahlen
Diagramm.xlsx	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 98: Relevante Dateien für die Support Vector Machine

Um das Trainingsset für den Durchlauf „Arbeitswoche/Feiertage + Wochenende“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend_temp.sql` folgende SQL-Code verwendet:

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "AIRTEMP" as
      ATTRIBUTE7, "WORKDAY" as ATTRIBUTE8, "WEEKEND" as ATTRIBUTE9 FROM "PAL"."
      CONSUMPTION_TRAINING";
```

Abbildung 107: SQL-Code für den gesamten Trainingsdatensatz bei SVM

Um das Trainingsset für den Durchlauf „Wochenende“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_temp.sql` SQL-Code aus Listing 108 verwendet.

In der folgenden Tabelle sollen die verschiedenen Parameterpaare den bei den Durchläufen entstehenden Fehlerkennzahlen gegenübergestellt werden. Dabei ist **T** als Temperatur, **WD** als Arbeitswoche und **WE** als Wochenende zu verstehen.

Im ersten Durchlauf wird nur die Temperatur als zusätzlicher Faktor für die Prognose betrachtet. Hierbei werden zuerst drei Teildurchläufe mit dem $\gamma = 0,001$ und jeweils ein

```
1 INSERT INTO PAL_SVM_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "AIRTEMP" as
      ATTRIBUTE7 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 108: SQL-Code für den Trainingsdatensatz bezogen auf die Temperatur bei SVM

	Anpassungsparameter		**Fehlerkennzahlen**				
Durchlauf	**γ**	**C**	**R^2**	**MAE**	**RMSE**	**CV (MAE)**	**CV (RMSE)**
T	0,001	10	0,39	10375,18	12652,8	0,17	0,21
T	0,001	100	0,39	9299,45	11387,10	0,15	0,18
T	0,001	1000	0,39	8825,68	10634,47	0,14	0,17
T	0,01	10	0,40	9391,04	11501,37	0,15	0,19
T	0,01	100	0,43	8887,67	10670,12	0,14	0,17
T	0,01	1000	0,58	76650,95	76946,06	1,24	1,25
T+WD+WE	0,001	10	0,49	10165,91	12427,66	0,17	0,20
T+WD+WE	0,001	100	0,51	9556,52	11672,40	0,16	0,19
T+WD+WE	0,001	1000	0,53	9379,71	11381,24	0,15	0,18
T+WD+WE	0,01	100	0,58	9312,43	11266,56	0,15	0,18
T+WD+WE	0,01	1000	0,72	74910,53	75171,31	1,22	1,22

Tabelle 99: Durchläufe SVM für die 2. Hypothese

$C = 10$, $C = 100$ und $C = 1000$. Dabei ergab $C = 1000$ das beste Ergebnis, das in Abbildung 109 zu sehen ist.

Dies spiegelt sich auch in den Fehlerkennzahlen wieder. So produziert der dritte Teildurchlauf im Vergleich zu den anderen beiden den geringsten MAE mit 8825,68 und den geringsten RMSE mit 10634,47 (siehe Tabelle 99). Denn bei diesen Fehlerkennzahlen deutet ein geringerer Wert eine stärkere Annährung an.

Anschließend werden im selben Durchlauf drei weitere Teildurchläufe gestartet, wo γ auf 0,01 gesetzt wird und C wie bei den ersten Teildurchläufen gesetzt wird. Hier ergibt die Kombination von $\gamma = 0,01$ und $C = 100$ das beste Ergebnis (siehe Abbildung 110), was sich auch in den Fehlerkennzahlen $MAE = 8887,67$ und $RMSE = 10670,12$ widerspiegelt (siehe Tabelle 99). Es fällt jedoch auf, dass sich der R-Squared dieses Paares mit 0,43 im Mittelfeld befindet, aber der beste R-Squared mit 0,58 in den anderen Fehlerkennzahlen schlechte Werte produziert.

Insgesamt ist im ersten Durchlauf festzustellen, dass $\gamma = 0,001$ und $C = 1000$ sowie $\gamma = 0,01$ und $C = 100$ ähnlich gute Ergebnisse liefern. Dies wird sowohl in den Fehlerkennzahlen MAE und RMSE (siehe Tabelle 99) als auch in den dazugehörigen Diagrammen

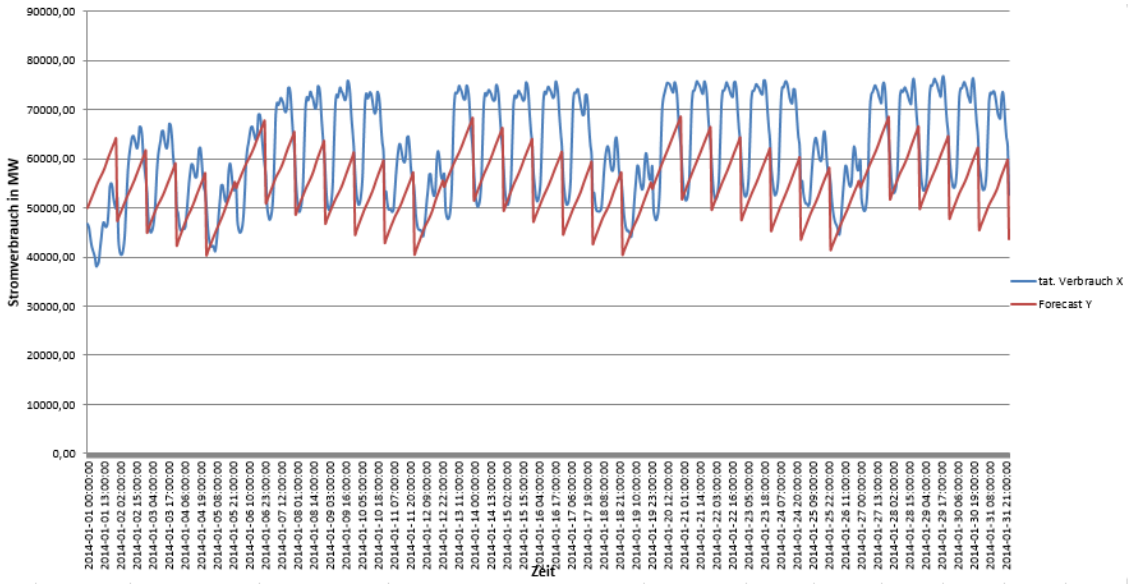


Abbildung 109: Support Vector Machine, T mit $\gamma = 0,001$ und $C = 1000$

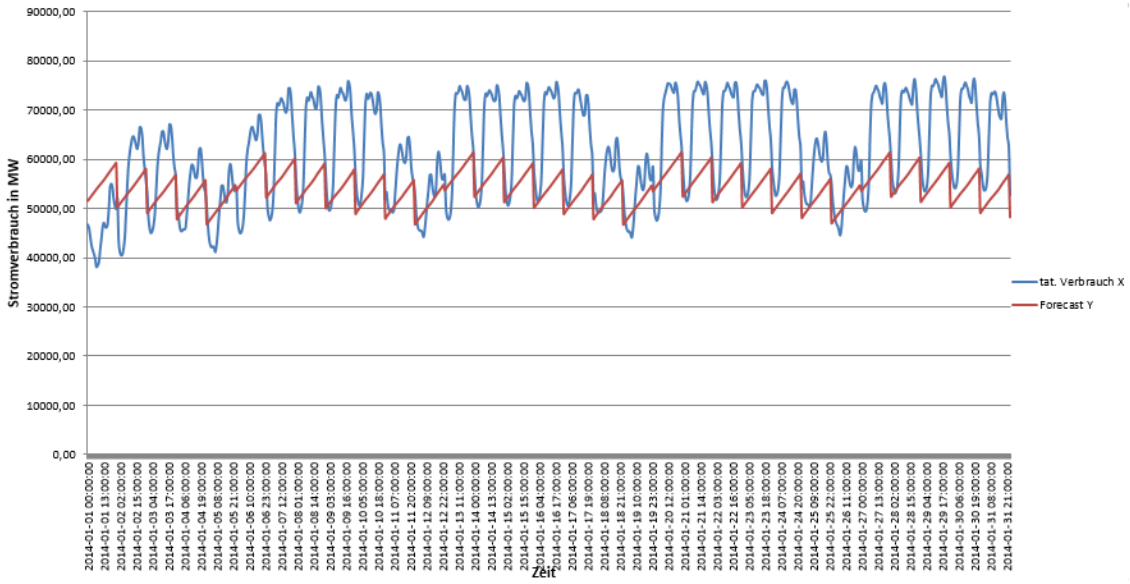


Abbildung 110: Support Vector Machine, T mit $\gamma = 0,01$ und $C = 100$

(siehe Abbildungen 109 und 110) deutlich.

Der zweite Durchlauf fügt zusätzlich die Faktoren Wochentag und Wochenende in die Prognose ein. Analog zum ersten Durchlauf werden bei den drei Teildurchläufen $\gamma = 0,001$ und C jeweils auf $C = 10$, $C = 100$ und $C = 1000$ gesetzt. Dabei ergibt die Parameterkombination $\gamma = 0,001$ und $C = 1000$ (siehe Abbildung 111) die besten Ergebnisse mit einem R^2 von 0,53 sowie MAE = 9379,71 und RMSE = 11381,24 (siehe Tabelle 99). Die anderen beiden Teildurchläufe haben durchweg schlechtere Ergebnisse.

Anschließend wird im selben Durchlauf zwei weitere Teildurchläufe gestartet mit den Pa-

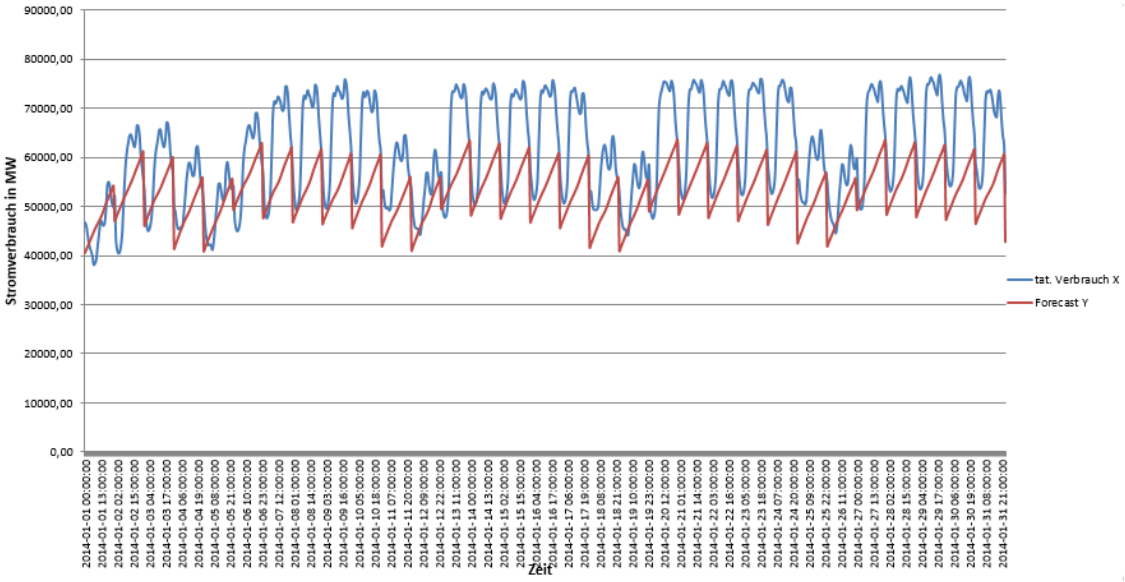


Abbildung 111: Support Vector Machine, T+WD+WE mit $\gamma = 0,001$ und $C = 1000$

rameterkombinationen $\gamma = 0,01$ und $C = 100$ sowie $\gamma = 0,01$ und $C = 1000$. Hierbei hat sich die Kombination $\gamma = 0,01$ und $C = 100$ durchgesetzt (siehe Abbildung 112), da hier die Fehlerkennzahlen mit Ausnahme des R-Squared bessere Werte erreichen (siehe Tabelle 99).

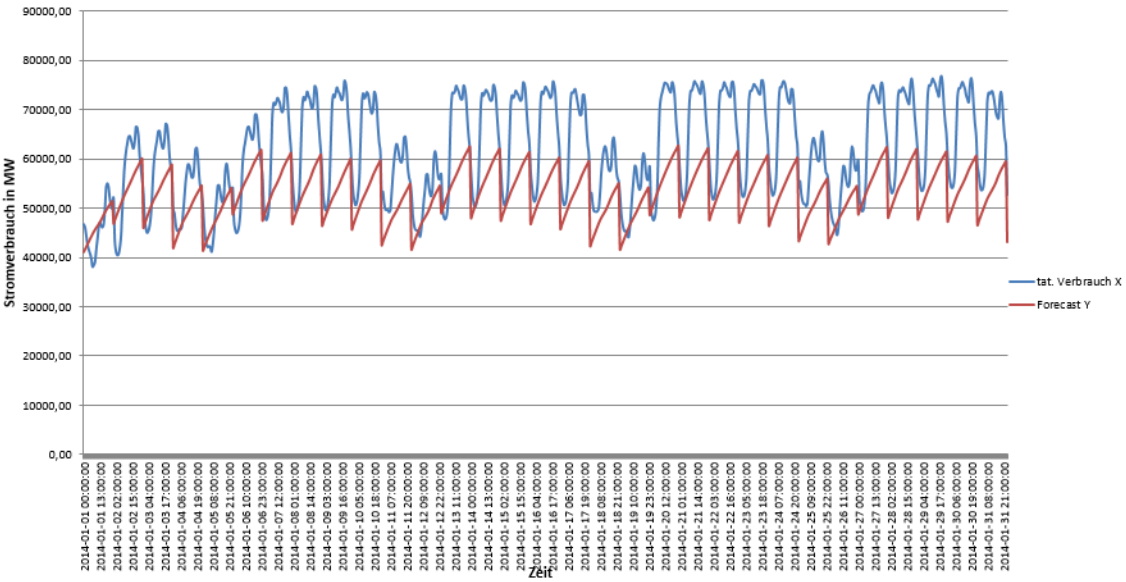


Abbildung 112: Support Vector Machine, T+WD+WE mit $\gamma = 0,01$ und $C = 100$

Auch im zweiten Durchlauf ist ein nahezu identischer Verlauf bei den beiden besten Kombinationspaaren festzustellen, sowohl anhand der Kennzahlen, die sich geringfügig voneinander unterscheiden, als auch anhand der Diagramme, deren Verlauf auch nur geringfügig abweicht.

Beim Betrachten aller Ergebnisse der zweiten Hypothese ist festzustellen, dass die Kombination $\gamma = 0,001$ und $C = 1000$ aus dem Durchlauf T sowohl bei den Diagrammen als auch bei den Fehlerkennzahlen das beste Ergebnis im Zuge dieser Hypothese liefert. Denn einerseits weist das Diagramm in diesem Fall die größte Ähnlichkeit mit den tatsächlichen Daten auf und andererseits sind die wichtigsten Fehlerkennzahlen MAE sowie RMSE unter allen Durchläufen dieser Hypothese am niedrigsten.

Zusammenfassung der Ergebnisse der zweiten Hypothese

Es erfolgt ein Vergleich der besten Ergebnisse der im Wettbewerb zueinander stehenden Algorithmen multiple lineare Regression, exponentielle Regression und Support Vector Machine. Alle Ergebnisse sind in Bezug zur bereits definierten ersten Hypothese (siehe auch Kapitel 3.2). Vorab wird festgestellt, dass die Algorithmen am besten operieren, wenn alle verfügbaren Trainingsdaten (im Zeitraum von 2009 bis 2013) für die Modellbildung eingebunden werden.

Die folgenden drei Durchläufe beziehen sich den besten Durchlauf des jeweiligen Algorithmus. Den besten Durchlauf bei der exponentiellen Regression weist (Durchlauf Spl _zusamm) mit den Fehlerkennzahlen R-Squared = 0,55, MAE = 8005,92, RMSE = 9664,11, CV MAE = 0,13 und CV RMSE = 0,16 auf.

Den besten Durchlauf bei der multiplen linearen Regression weist (Durchlauf Spl_zusamm) mit den Fehlerkennzahlen R-Squared = 0,57, MAE = 7268,98 und RMSE= 8804,53, CV MAE = 0,12 und CV RMSE = 0,14 auf.

Die Ergebnisse der Support Vector Machine mit den Anpassungsfaktoren $C = 1000$ und $\gamma = 0,001$ hingegen sind am genauesten, wenn nur die Temperatur in die Berechnung mit einbezogen wird. Die Fehlerkennzahlen betragen hier R-Squared = 0,39, MAE = 8825,68, RMSE = 10634,47, CV MAE = 0,14 CV RMSE = 0,17.

Im Vergleich der drei Ergebnisse wird deutlich, dass die multiple lineare Regression das beste Ergebnis liefert.

12.6 3. Hypothese

Die dritte Hypothese besagt, dass die Vorhersage für den Stromverbrauch, die mit der jeweiligen Durchschnittstemperatur sowie den Energiepreisen für Strom angereichert wird, genauere Ergebnisse liefert als die zweite Hypothese (siehe Kapitel 12.5. Die Datenbasis für diese Hypothese besteht aus den historischen Energieverbrauchsdaten der Entso-E vom 01.01.2009 bis zum 31.12.2013. Zusätzlich werden diese Daten mit den jeweiligen Strompreisen für Haushalte und Industrie auf Jahresbasis angereichert. Dazu kommen die Informationen aus der ersten Hypothese (Angaben zu Feiertagen sowie Wochenende) und der zweiten Hypothese (Stündlichen Angaben zu den Durchschnittstemperaturen) hinzu. Um hier zu differenzieren, welchen Einfluss die Strompreise der Industrie sowie der Haushalte auf die Prädiktion hat, werden im ersten Versuch zunächst die Haushaltspreise in die Modellbildung eingehen. Im zweiten Versuch wird der Strompreis der Industrie isoliert betrachtet. Im Dritten Versuch werden beide Strompreise in die Modellbildung eingebunden. Im folgenden Abschnitt wird zunächst der Datenfluss beschrieben. Die Ergebnisse zu den einzelnen Algorithmen werden in den darauf folgenden Abschnitten dokumentiert.

Datenfluss Der Datenfluss für die Energieverbrauchsdaten verhält sich analog zu den Ausführungen in den Kapiteln 12.3.1, 12.4 und 12.5. Es folgt die Erläuterung, wie sich die Strompreisdaten zusammensetzen. Die Quelltabelle für die Strompreise der Haushalte ist `PRE.PRICE_HOUSEHOLD_ALL_CLEAN`. Die Quelltabelle für die Strompreise der Industrie ist `PRE.PRICE_INDUSTRIY_ALL_CLEAN`. Die Datenfelder dieser Tabellen sind in den Tabellen 100 und 101 zu finden.

Year	Price
2009	0,2282
2010	0,2375
...	...

Tabelle 100: Tabelle `PRE.PRICE_HOUSEHOLD_ALL_CLEAN`

Year	Price
2009	0,0975
2010	0,0921
...	...

Tabelle 101: Tabelle `PRE.PRICE_INDUSTRY_ALL_CLEAN`

Anders als bei anderen Datenquellen muss hier kein aufwendiger ETL-Prozess ausgeführt werden, da beide Tabellen lediglich aus 6 Zeilen bestehen, die mit der jeweiligen Angabe des Strompreises zu jedem Jahr befüllt sind. Die Korrekturen der Datensätze werden hier händisch umgesetzt. Die Spalte `Year` enthält den Zeitpunkt der Messung. Die Spalte `Price` enthält den erfassten Strompreis.

Mit Hilfe dieser beiden Datentabellen kann nun die passende View für die Algorithmen erstellt werden. Zu diesem Zweck wird die View aus Listing 80 erneut erweitert. Neben den Angaben zu Feiertagen und Wochenenden der ersten Hypothese enthält die View auch die Angaben zu der Durchschnittstemperatur der jeweiligen Stunden, sowie die jährlichen Strompreisdaten für Haushalte und der Industrie. Der SQL-Code der mit den Wetterdaten angereicherten View ist in Listing113 dokumentiert.

```
1  CREATE VIEW "PAL"."CONSUMPTION_TRAINING" ( "ID",
2              "CONSUMPTION",
3              "HOUR_COUNT",
4              "HOUR_OF_DAY",
5              "DAY_OF_WEEK",
6              "DAY_OF_MONTH",
7              "MONTH",
8              "QUARTER",
9              "YEAR",
10             "WORKDAY",
11             "AIRTEMP",
12             "PRICE_HOUSEHOLD",
13             "PRICE_INDUSTRY",
14             "WEEKEND" ) AS SELECT
15             c.ID-1 as ID,
16             c.CONSUMPTION as CONSUMPTION,
17             c.ID as HOUR_COUNT,
18             t.HOUR_INT as HOUR_OF_DAY,
19             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
20             t.DAY_INT as DAY_OF_MONTH,
21             t.MONTH_INT as "MONTH",
22             t.QUARTER_INT as "QUARTER",
23             t.YEAR_INT as "YEAR" ,
24             t.WORKDAY as "WORKDAY",
25             w."AIRTEMP" as "AIRTEMP",
26             x."PRICE" as "PRICE_HOUSEHOLD",
27             y."PRICE" as "PRICE_INDUSTRY",
28             CASE WHEN t.DAY_OF_WEEK_INT = 5
29 or t.DAY_OF_WEEK_INT = 6
30 THEN 1
31 ELSE 0
32 END as "WEEKEND"
33 FROM "PRE"."CONSUMPTION_ALL_CLEAN" as c JOIN "_SYS_BI"."M_TIME_DIMENSION" as
       t ON c.ID+8783 = t."HOUR_COUNT"
34 LEFT OUTER JOIN "PRE"."AIRTEMP_ALL_CLEAN" as w ON c."ID" = w."ID"
35 LEFT OUTER JOIN "PRE"."PRICE_HOUSEHOLD_ALL_CLEAN" as x ON x."YEAR" = t."
     YEAR_INT"
36 LEFT OUTER JOIN "PRE"."PRICE_INDUSTRY_ALL_CLEAN" as y ON y."YEAR" = t."
     YEAR_INT"
37 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 113: Erweiterte View `CONSUMPTION_TRAINING`

Die beiden Tabellen `PRE.PRICE_HOUSEHOLD_ALL_CLEAN` und `PRE.PRICE_INDUSTRIY_ALL_CLEAN` werden über einen Left Outer Join mit der Tabelle `PRE.CONSUMPTION_ALL_CLEAN` in die View eingebunden. Ebenso wird die View für die Testdaten mit den neuen Tabellen angereichert. Dies kann in Listing 114 nachvollzogen werden.

```
1  CREATE VIEW "PAL"."CONSUMPTION_FORECAST" ( "ID",
2              "HOUR_COUNT",
3              "HOUR_OF_DAY",
4              "DAY_OF_WEEK",
5              "DAY_OF_MONTH",
6              "MONTH",
7              "QUARTER",
8              "YEAR",
9              "WORKDAY",
10             "AIRTEMP",
11             "PRICE_HOUSEHOLD",
12             "PRICE_INDUSTRY",
13             "WEEKEND" ) AS SELECT
14             t.HOUR_COUNT-8783-43824-1 as ID,
15             t.HOUR_COUNT-8783 as HOUR_COUNT,
16             t.HOUR_INT as HOUR_OF_DAY,
17             t.DAY_OF_WEEK_INT as DAY_OF_WEEK,
18             t.DAY_INT as DAY_OF_MONTH,
19             t.MONTH_INT as "MONTH",
20             t.QUARTER_INT as "QUARTER",
21             t.YEAR_INT as "YEAR",
22             t.WORKDAY as "WORKDAY",
23             w.AIRTEMP as "AIRTEMP",
24             x.PRICE AS "PRICE_HOUSEHOLD",
25             y.PRICE AS "PRICE_INDUSTRY",
26             CASE WHEN t.DAY_OF_WEEK_INT = 5
27 or t.DAY_OF_WEEK_INT = 6
28 THEN 1
29 ELSE 0
30 END as "WEEKEND"
31 FROM "_SYS_BI"."M_TIME_DIMENSION" as t
32 LEFT OUTER JOIN "PRE"."AIRTEMP_ALL_CLEAN" as w on t.HOUR_COUNT-8783 = w."ID"
33 LEFT OUTER JOIN "PRE"."PRICE_HOUSEHOLD_ALL_CLEAN" as x on x."YEAR" = t."
       YEAR_INT"
34 LEFT OUTER JOIN "PRE"."PRICE_INDUSTRY_ALL_CLEAN" as y on y."YEAR" = t."
       YEAR_INT"
35 WHERE t.DATETIMESTAMP >= '2014-01-01 00:00:00'
36 AND t.DATETIMESTAMP <= '2014-01-31 24:00:00'
37 ORDER BY ID ASC WITH READ ONLY
```

Abbildung 114: Erweiterte View `CONSUMPTION_FORECAST`

Die Beschreibung des Datenflusses für die Evaluation der dritten Hypothese ist hiermit abgeschlossen. Folgend erfolgt die Beschreibung der Versuchsdurchführung.

Versuchsdurchführung Die Vorhersagen für die Evaluation dieser Hypothese teilen sich in drei Versuche auf. Mit Hilfe der View `CONSUMPTION_TRAINING` (siehe Listing 113) werden Modelle mit den entsprechenden Algorithmus gebildet. Dabei werden folgende Versuche betrachtet:

1. Prädiktion mit Angaben zu Strompreisen von Haushalten in Deutschland.
2. Prädiktion mit Angaben zu Strompreisen der Industrie in Deutschland.

3. Prädiktion mit Angaben zu Strompreisen von Haushalten und der Industrie in Deutschland.

Die Versuche beziehen sich dabei auf die gesamten Trainingsdaten (01.01.2009 - 31.12.2013). In den folgenden Abschnitten erfolgt die Beschreibung der Ergebnisse der Algorithmen.

Multiple Lineare Regression Alle Durchläufe zur Multiplen Linearen Regression können auf der beigelegten CD im Ordner `svn/Forecast/3.Hypothese/MultipleLineareRegression/` gefunden werden. Im Ordner `Haushalt` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten in Deutschland. Im Ordner `Industrie` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten der Industrie in Deutschland. Genauso sind entsprechend dazu im Ordner `Haushalt_Industrie` die Daten von Haushalten und Industrie zu finden. Im Ordner `Split` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten sowie Industriepreisen innerhalb Deutschlands zu finden. Es soll festgestellt werden, ob eine Unterscheidung zwischen Arbeitstagen und Wochenenden bei den Trainingsdaten bessere Vorhersageergebnisse liefert. Dementsprechend befinden sich im Ordner `Split` die Unterordner `Woche` und `Wochenende`. Im Anschluss werden die Prädiktionen dieser beiden Versuche zusammengefügt. Die Ergebnisse hierzu sind im Ordner `Zusammengefuegt` zu finden. In jedem dieser Ordner sind folgende Dateien vorhanden:

Datei	**Inhalt**
build_modell_mlr.sql	Script zum Erstellen des Modells
build_forecast_mlr.sql	Script zum Erstellen der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellte Fehlerkennzahlen, bezogen auf die Trainingsdaten
forecast.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014
Diagramm.xlsx	Grafische Darstellung von tatsächlichen und vorhergesagten Stromverbrauch

Tabelle 102: Relevante Dateien für die Multiple Lineare Regression

Um das Trainingsset für die jeweiligen Durchläufe einzuschränken, wird in der Datei `build_modell_lmr.sql` folgendes SQL-Statement verwendet:

Für die im vorherigen Abschnitt erwähnte Trennung von Arbeitstagen und Wochenenden werden zwei Views erstellt, welche ausschließlich Trainingsdaten für Wochenende (`CONSUMPTION_TRAINING_WEEKEND`) und Arbeitstage (`CONSUMPTION_TRAINING_WEEK`) enthalten. Diese Views werden entsprechend als Quelltabellen für die Trainingsdaten genutzt. In der Modellbildung wird dies durch die in Listing 118 und 119 aufgeführten SQL-Statements dargestellt.

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
     "WORKDAY", "AIRTEMP", "PRICE_HOUSEHOLD", "WEEKEND" FROM "PAL"."
    CONSUMPTION_TRAINING";
```

Abbildung 115: SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten in Deutschland

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
     "WORKDAY", "AIRTEMP", "PRICE_INDUSTRY", "WEEKEND" FROM "PAL"."
    CONSUMPTION_TRAINING";
```

Abbildung 116: SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten der Industrie in Deutschland

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
     "WORKDAY", "AIRTEMP", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY", "WEEKEND"
    FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 117: SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
    "WORKDAY", "AIRTEMP", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING_WEEK";
```

Abbildung 118: SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland

```
1 INSERT INTO PAL_MLR_DATA_TBL SELECT "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "YEAR",
    "WORKDAY", "AIRTEMP", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING_WEEKEND";
```

Abbildung 119: SQL-Statement für den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland

Listing 115 zeigt, wie der gesamte Trainingsdatensatz und die Strompreisdaten für Haushalte in Deutschland eingebunden werden. Im Listing 116 wird illustriert, wie der gesamte Trainingsdatensatz mit den Strompreisdaten der Industrie in Deutschland eingebunden wird. Analog dazu zeigt Listing 117, wie der gesamte Trainingsdatensatz mit den Strompreisdaten der Haushalte und der Industrie eingebunden wird. Genauso ist im Listing 118 zu sehen, wie der gesamte Trainingdatensatz (nur Arbeitswoche) mit den Strompreisdaten der Haushalte und der Industrie eingebunden wird. Listing 119 zeigt wie der gesamte

Trainingdatensatz (nur Wochenenden) mit den Strompreisdaten der Haushalte und der Industrie eingebunden wird. Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden, muss folgendermaßen vorgegangen werden: Zunächst muss das Script `build_modell_mlr.sql` in SAP HANA ausgeführt werden, welche das mathematische Modell des Algorithmus auf Basis der Trainingsdaten anlegt. Anschließend muss das Script `build_forecast_mlr.sql` ausgeführt werden. Hiernach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und Übersichtsgründen werden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared) bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen bezogen auf das Testdatenset enthalten. Die Datei `Diagramm.xlsx` zeigt die grafische Darstellung von tatsächlichen und vorhergesagten Stromverbrauch. Die Parametereinstellungen werden wie folgt gewählt:

Parameter	**Einstellung**	**Erläuterung**
THREAD_NUMBER	8	Modellbildung wird mit 8 Threads durchgeführt.
PMML_EXPORT	1	Gibt an, dass das Modell im PMML-Format vorliegt.
ADJUSTED_R2	1	R-Squared und R-Squared-Adjusted werden berechnet.
VARIABLE_SELECTION	0	Alle in der View vorhandenen Variablen werden zur Modellbildung einbezogen.

Tabelle 103: Parametereinstellungen der Multiple Lineare Regression (für alle Durchläufe)

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/3.Hypothese/MultipleLineareRegression`. Wie bereits zuvor erwähnt finden sich in diesen Ordnern alle relevanten Ergebnisse der Test-Durchläufe. Die Fehlerkennzahlen zu den Prädiktionen sind in Tabelle 104 zusammengefasst dargestellt. Dabei bezieht sich der Durchlauf „H-Halt“ auf die Haushaltstrompreise, während sich der Durchlauf „Ind“ auf die Industriestrompreise bezieht, der Durchlauf „H-Halt-Ind“ bezieht sich wiederum auf die gemeinsame Nutzung von Haushalt- und Industriestrompreise. Der Durchlauf „Spl-Woch“ bezieht sich auf die Haushalt- und Industriestrompreise in der Woche, der Durchlauf „Spl-WEnde“ bezieht sich auf die Haushalt- und Industriestrompreise am Wochenende. Im Versuch „Split-zusam“ werden die Ergebnisse aus den Versuchen „Spl-Woch“ und „Spl-Wochenende“ zusammengefügt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
H-Halt	0,55	7387,86	8948,69	0,11	0,14
Ind	0,55	8119,12	9779,82	0,13	0,15
H-Halt-Ind	0,55	8079,52	9735,82	0,13	0,15
Spl-Woch	0,46	8292,96	10067,01	0,12	0,15
Spl-WEnde	0,56	6930,39	7913,76	0,12	0,14
Split-zusam	0,57	7941,33	9557,88	0,12	0,15

Tabelle 104: Ergebnisse der Hypothese

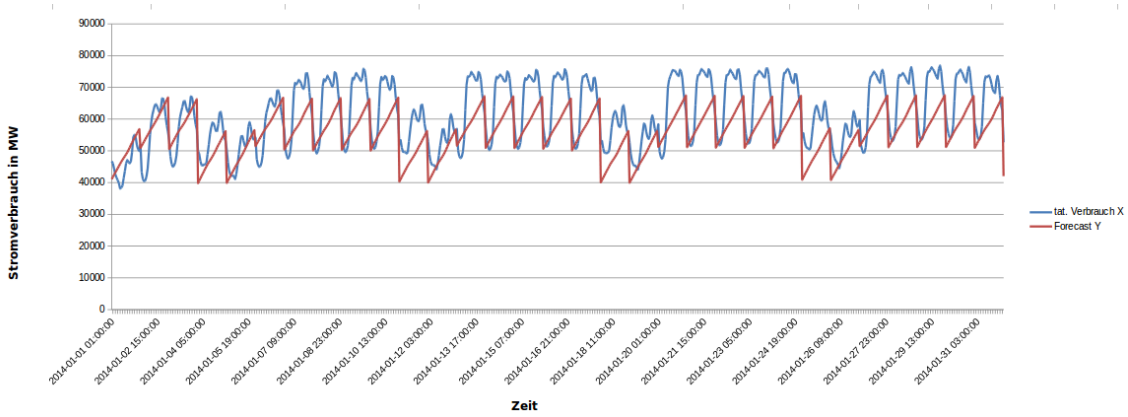


Abbildung 120: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „H-Halt“

Abbildung 120 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten unter Verwendung der Haushaltstrompreisdaten.

Abbildung 121 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten unter Verwendung der Industriestrompreisdaten.

Abbildung 122 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten unter Verwendung von Hauhalt- und Industriestrompreisdaten.

Abbildung 123 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten unter Verwendung der Hauhalt- und Industriestrompreisdaten in der Woche.

Abbildung 124 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Vergleich zu den Prädiktionswerten unter Verwendung der Hauhalt- und Industriestrompreisdaten am Wochenende.

Abbildung 125 zeigt die grafische Darstellung des tatsächlichen Stromverbrauches im Ver-

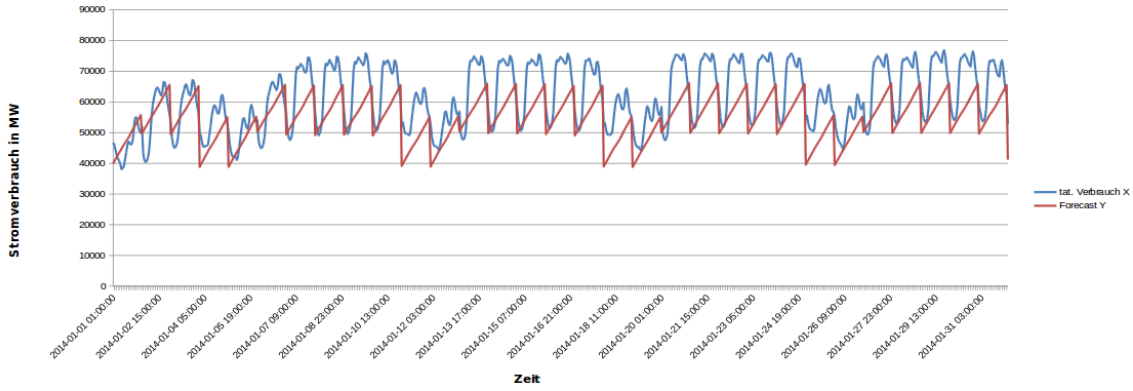


Abbildung 121: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Ind“

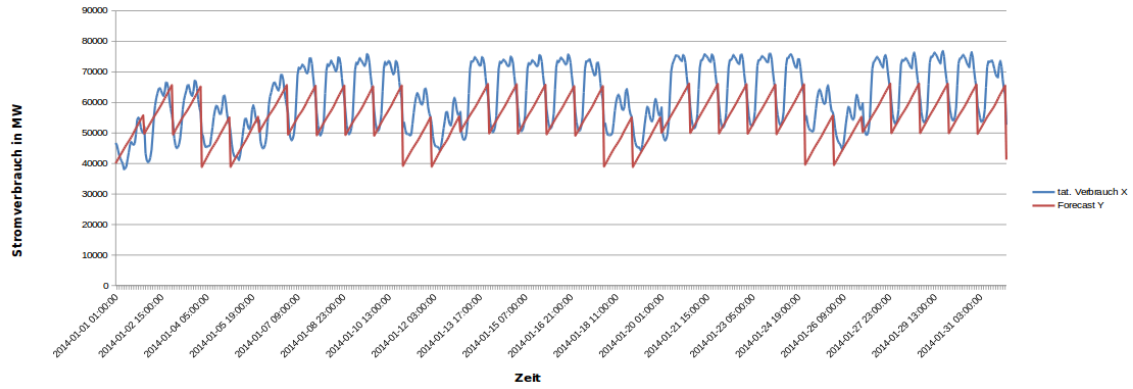


Abbildung 122: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „H-Halt-Ind“

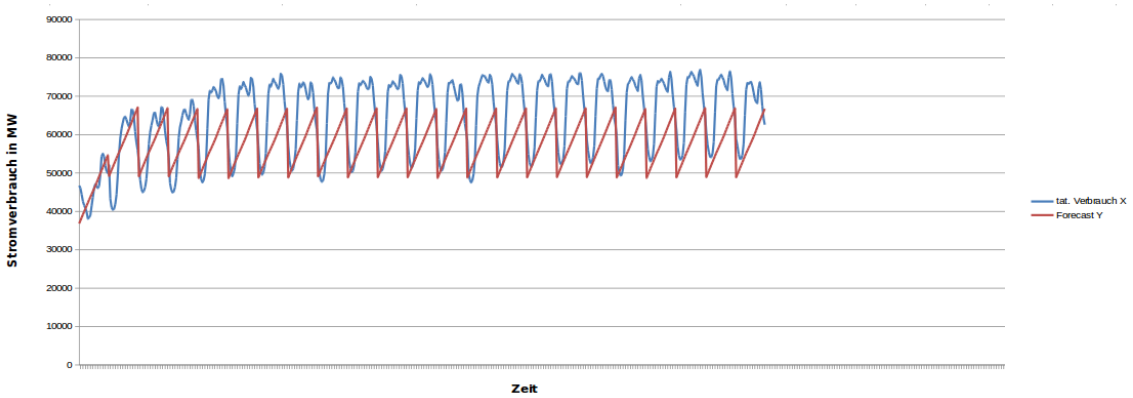


Abbildung 123: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Spl-Woch“

gleich zu den Prädiktionswerten unter Verwendung der Hauhalt- und Industriestrompreisdaten in der Woche sowie am Wochenende.

Abbildung 124: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Spl-WEnde“

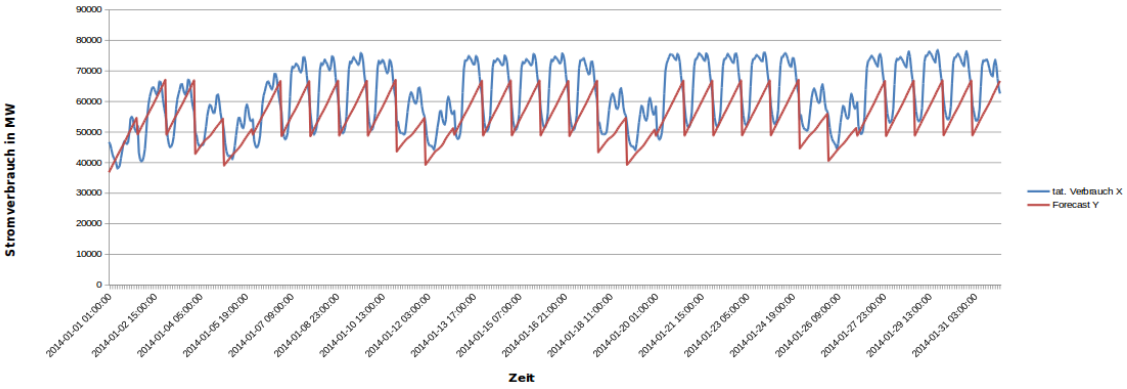


Abbildung 125: Vergleich von tatsächlichen und vorhergesagten Stromverbrauch. Versuch: „Split-zusam“

Vergleicht man zunächst die Abbildungen 120, 121 und 122 fällt in alle drei Durchläufen wieder der fehlende Sockelbetrag (wie bereits in der Datenbasis sowie in der Datenbasis sowie der ersten und zweiten Hypothese festgestellt) an den Wochenenden auf. Bei weiterer Betrachtung fällt auf, dass dieser Sockelbetrag in den Arbeitswochen nicht erforderlich ist. Hier werden die Tiefen annähernd richtig erkannt, jedoch sind die Höhen - im Vergleich zum tatsächlichen Stromverbrauch - zu niedrig prognostiziert. Die Kombination dieser beiden Problemfaktoren führt insgesamt zu nicht befriedigenden Ergebnissen. Ebenso wird der Stromverbrauch im Verlaufe eines Tages nicht korrekt modelliert. Diese Probleme werden auch an der Fehlerkennzahlen der Durchläufe in Tabelle 104 deutlich. Der Durchlauf „H-Halt“ erreicht hier einen R-Squared von 0,55, einen MAE von 7387,86 Punkten und einen RMSE von 8948,69 Punkten. Der Einfluss der Haushaltsstrompreises scheint also einen geringen Einfluss auf die Modellbildung der Multiplen Linearen Regression zu haben. Auch der Durchlauf „Ind“ produziert einen R-Squared von 0,55 und hat - im Vergleich zum Durchlauf „H-Halt“ einen noch schlechteren MAE von 8119,12 Punkten

erzielt. Ebenso liegt hier der RMSE mit 9779,82 Punkten noch schlechter als der Versuch „H-Halt“. Demnach hat der Einfluss des Industriestrompreises einen vernachlässigbaren Einfluss auf die Modellbildung der Multiplen Linearen Regression. Im Versuch „H-Halt-Ind“ werden die beiden Faktoren Haushaltsstrompreis und Industriestrompreis kombiniert. Dieser Versuch erzielt ebenso einen R-Squared von 0,55 und einen MAE von 8079,52 und einen RMSE von 9735,82 Punkten. Gemessen an diesen Durchläufen erzielt der Versuch „H-Halt“ das beste Ergebnis, jedoch ist dieses Ergebnis nicht zufriedenstellend. Die Durchläufe „Spl-Woch“ und „Spl-WEnde“ erstellen entsprechende Prädiktionen für die Wochen beziehungsweise Wochenenden des Monats Januar 2014. Wie bereits angedeutet fehlt der Prädiktion für das Wochenende eine „Sockelbeitrag“ (Versuch „Spl-WEnde“). Dieses Verhalten gilt jedoch nicht für den Versuch „Spl-Woch“. In diesem Versuch werden die Tiefen verhältnismäßig korrekt berechnet, während die berechneten Höhen im Vergleich zum tatsächlichen Verbrauch zu niedrig ausfallen. Bezugnehmend auf die Fehlerkennzahlen in Tabelle 104 erreicht der Durchlauf „Spl-Woch“ einen R-Sqared von 0,46, einen MAE von 8292,96 und einen RMSE von 10067,01 Punkten. Der Durchlauf „Spl-WEnde“ erreicht einen R-Squared von 0,56, einen MAE von 6930,39 und einen RMSE von 7913,76 Punkten. Im Versuch „Split-zusam“ werden die Ergebnisse der Durchläufe „Split-Woch“ und „Split-WEnde“ wieder zusammengefügt. Dieser Durchlauf erzielt einen R-Squard von 0,57, welcher im Vergleich zum Durchlauf „H-Halt“ zwar zunächst besser erscheint. Vergleicht man jedoch MAE und RMSE von „H-Halt“ und „Split-zusam“, wird deutlich, dass hier der Durchlauf „H-Halt“ bessere Ergebnisse erzielt. Hierbei liegt der MAE vom Versuch „H-Halt“ um 553,47 niedriger als der Versuch „Split-zusam“. Dies gilt auch für den RMSE: Hier liegt der RMSE von Versuch „H-Halt“ um 609,19 Punkte niedriger als der Versuch „Split-zusam“. Zusammenfassend lässt sich bezogen auf die Multiple Lineare Regression für diese Hypothese sagen, dass der Durchlauf „H-Halt“ das beste Ergebnisse liefert, wobei auch dieser Durchlauf keine signifikant guten Ergebnisse ergibt. Insbesondere hat in diesem Kontext auch die Aufteilung der Trainingsdatenmenge in Arbeitswochen und Wochenenden keine Verbesserung der Prädiktionswerte ergeben.

Exponentielle Regression Alle Durchläufe zur Exponentiellen Regression können auf der beigelegten CD im Ordner `svn/Forecast/3.Hypothese/Exponentialregression/` zu finden. In allen Unter-Ordnern sind entsprechend folgende Dateien vorhanden:

Im Ordner `Haushalt` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten in Deutschland. Im Ordner `Industrie` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten der Industrie in Deutschland. Im Ordner `Gesamt` befinden sich die Daten zur Modell- und Prädiktionsbildung über den gesamten Trainingsdatensatz mit Strompreisdaten von Haushalten und der Industrie in Deutschland. Im Ordner `Split` befinden sich die Daten zur Modell- und Prädiktionsbildung über den

gesamten Trainingsdatensatz mit Woche, Wochenende und Woche_ Wochentag.

Datei	**Inhalt**
Exponentialregression.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
fehlerkennzahlen.csv	Vom Algorithmus erstellte Fehlerkennzahlen, bezogen auf die Trainingsdaten
forecast.csv	Vorhersagewerte für Januar 2014
Januar2014.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014
Diagramm.xlxs	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 105: Relevaten Dateien für die Exponentiale Regression

Um das Trainingsset für die jeweiligen Durchläufe einzuschränken, wird in der Datei `Exponentialregression.sql` folgendes SQL-Statement verwendet:

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "WORKDAY", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING";
```

Abbildung 126: SQL-Statement für den gesamten Trainingsdatensatz mit Gesamtpreis

Listing 126 zeigt, wie der gesamte Trainingsdatensatz und Gesamtpreis von Industrie und Haushalt in die Modellbildung eingebunden werden.

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "WORKDAY", "PRICE_HOUSEHOLD" FROM "PAL"."
    CONSUMPTION_TRAINING";
```

Abbildung 127: SQL-Statement für den gesamten Trainingsdatensatz mit Haushaltspreis

Listing 127 zeigt, wie der gesamte Trainingsdatensatz und Haushaltpreis in die Modellbildung eingebunden werden.

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "WORKDAY", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING";
```

Abbildung 128: SQL-Statement für den gesamten Trainingsdatensatz mit Industriepreis

Listing 128 zeigt, wie der gesamte Trainingsdatensatz und Industriepreis in die Modellbildung eingebunden werden.

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER","AIRTEMP"
    , "WORKDAY", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING_WEEK";
```

Abbildung 129: SQL-Statement für den gesamten Trainingsdatensatz mit Woche ohne Wochenende

Listing 129 zeigt, wie der gesamte Trainingsdatensatz von Wochentagen ohne Wochenende in die Modellbildung eingebunden werden

```
1 INSERT INTO PAL_ER_DATA_TBL select "ID", "CONSUMPTION", "HOUR_COUNT", "
    HOUR_OF_DAY", "DAY_OF_WEEK", "DAY_OF_MONTH", "MONTH", "QUARTER", "WEEKEND
    ","AIRTEMP", "PRICE_HOUSEHOLD", "PRICE_INDUSTRY" FROM "PAL"."
    CONSUMPTION_TRAINING_WEEKEND";
```

Abbildung 130: SQL-Statement für den gesamten Trainingsdatensatz mit Wochenende

Listing 130 zeigt, wie der gesamte Trainingsdatensatz vom Wochenende in die Modellbildung eingebunden werden.

Um nun ein Modell für die ausgewählten Trainingsdaten zu bilden muss folgendermaßen vorgegangen werden: Es wird das Script `Exponentialregression.sql` in SAP HANA ausgeführt. Danach liegen die Vorhersagen für den Januar 2014 als Tabellen in SAP HANA vor. Aufgrund der Menge an getätigten Vorhersagen und aus Übersichtsgründen werden diese Tabellen jeweils exportiert und liegen in dem entsprechenden Ordner im CSV- bzw. Excel-Format vor: Die Datei `fehlerkennzahlen.csv` beinhaltet die von dem SAP HANA System generierten Kennzahlen (z.B. R-Squared), bezogen auf die Modellbildung mit den Trainingsdaten. Die Datei `forecast.csv` beinhaltet die von dem Algorithmus generierten Prädiktionsdaten für den Zeitraum Januar 2014. Die Datei `Januar2014.xlxs` beinhaltet den Vergleich des tatsächlichen Stromverbrauches und der Prädiktion. In dieser Datei sind ebenfalls die Fehlerkennzahlen, bezogen auf das Testdatenset enthalten.

Ergebnisse der Durchläufe Die folgenden Aussagen beziehen sich auf die Ergebnisse im Ordner `svn/Forecast/3.Hypothese/Exponentialregression`. Es finden sich in diesen Ordnern alle relevanten Ergebnisse zu den Durchläufen. Die Fehlerkennzahlen zu den Prädiktionen sind in Tabelle 106 zusammengefasst dargestellt.

Dabei bezieht sich der Durchlauf „Hhalt“ auf die Haushaltstrompreise, der Durchlauf „Indus“ bezieht sich auf die Industriestrompreise, der Durchlauf „Gesamt “ bezieht sich auf die gemeinsame Nutzung von Haushalt- und Industriestrompreise, der Durchlauf „SplW“ bezieht sich auf die Haushalt- und Industriestrompreise in der Woche, der Durchlauf „SplW0“ bezieht sich auf die Haushalt- und Industriestrompreise am Wochenende. Im Versuch „Splzusamm“ werden die Ergebnisse aus den Versuchen „SplW“ und „SplWo“ zusammengefügt.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Gesamt	0,55	8879,59	10616,65	0,14	0,17
H_halt	0,54	8041,53	9727,77	0,13	0,16
Indus	0.54	8923,39	10661,59	0,14	0,17
Split	...	...	...	...	...
Spl_W	0,44	9200,19	11072,8	0,14	0,17
Spl_Wo	0,55	7679,22	8609,51	0,14	0,16
Spl_zusamm	0,56	8807,68	10492,62	0,14	0,17

Tabelle 106: Fehlerkennzahlen der Anwendung des Modells auf die Testdaten

Abbildung 131 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „Gesamt“.

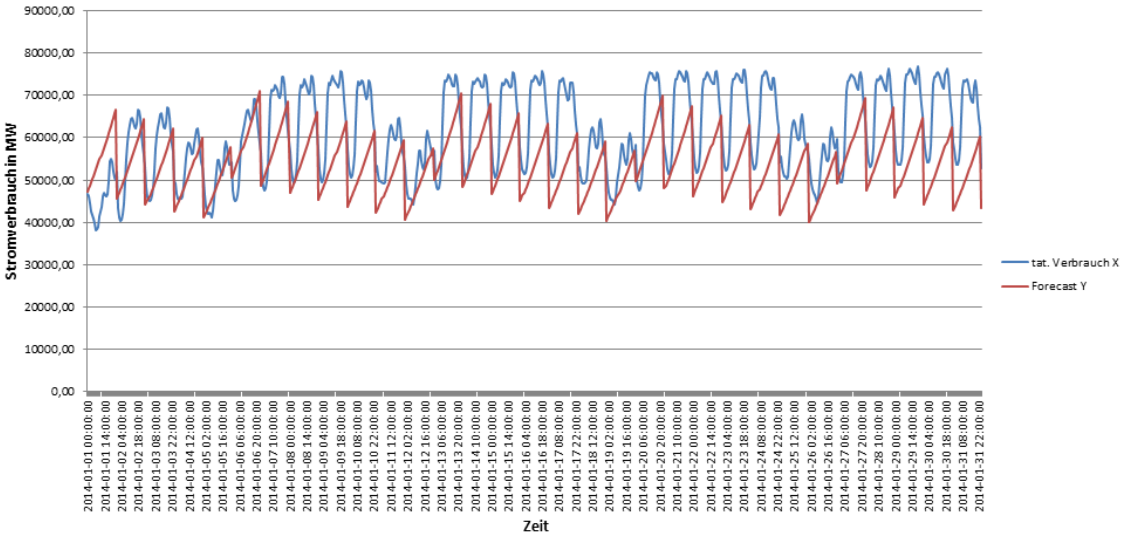


Abbildung 131: Diagramm Vergleich Ist und Forecast. Durchlauf: „Gesamt “

Abbildung 132 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „H_halt “.

Abbildung 133 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „Indus “.

Abbildung 135 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „Spl_W “.

Abbildung 135 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „Spl_Wo “.

Abbildung 136 zeigt die grafische Darstellung der tatsächlichen Stromverbräuche im Vergleich zu den Prädiktionswerten für den Durchlauf: „Spl_zusamm “.

Die Tabelle 106 dokumentiert, dass der Durchlauf „Hhalt“ einen R-Squared von 0,54, einen MAE von 8041,53 Punkten und einen RMSE von 9727,7 Punkten erreicht. Der Einfluss des

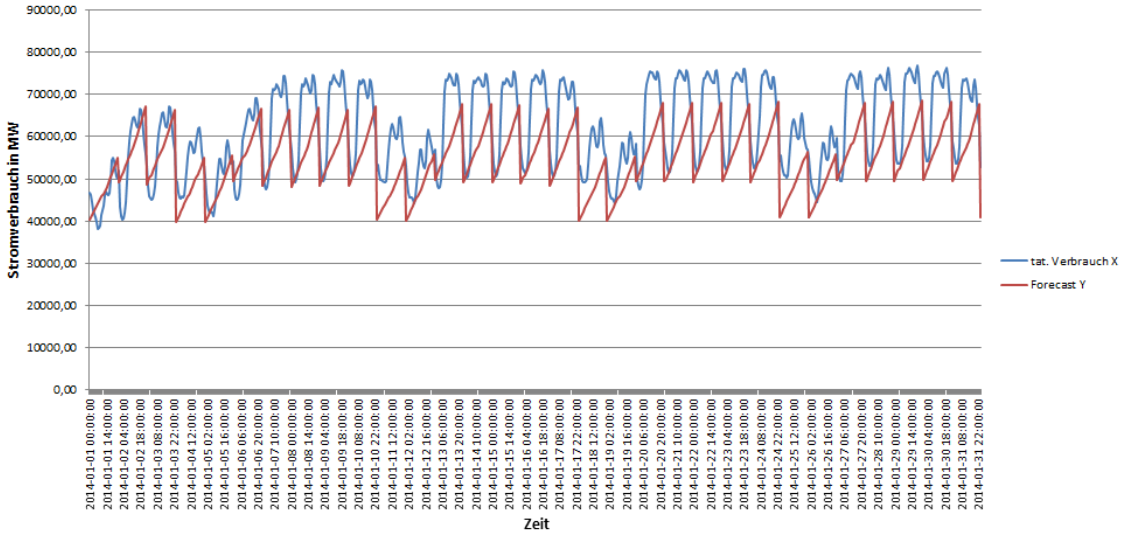


Abbildung 132: Diagramm Vergleich Ist und Forecast. Durchlauf : „H_halt “

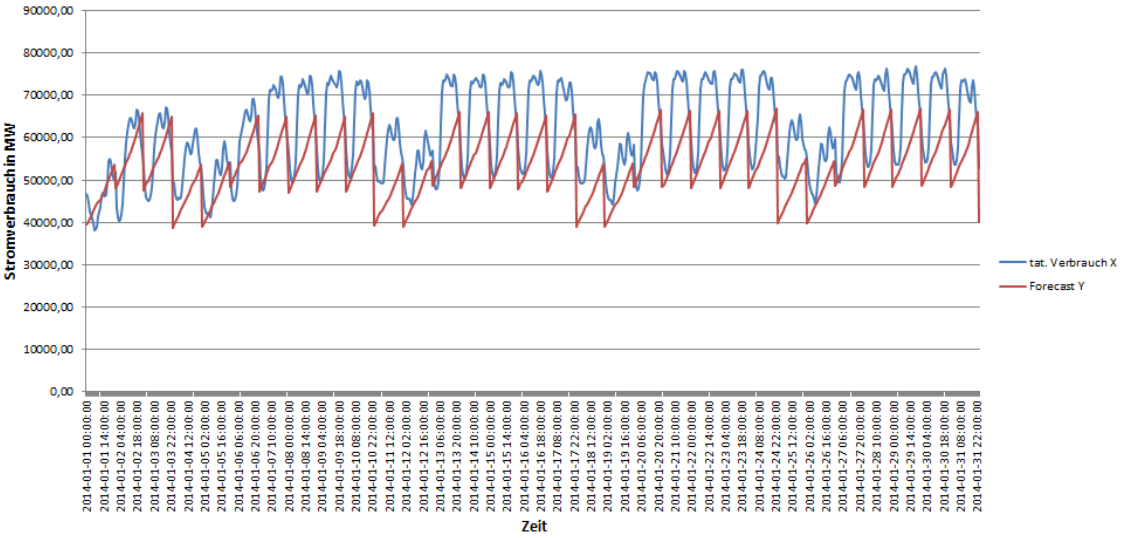


Abbildung 133: Diagramm Vergleich Ist und Forecast. Durchlauf: „Indus “

Haushaltsstrompreises scheint also einen geringen Einfluss auf die Modellbildung der exponentiellen Regression zu haben. Auch der Durchlauf „Indus“ produziert einen R-Squared von 0,54 und hat - im Vergleich zum Durchlauf „Hhalt“ einen noch schlechteren MAE von 8923,39 Punkten erzielt. Ebenso liegt hier der RMSE mit 10661,59 Punkten noch schlechter als der Durchlauf „Hhalt“. Demnach hat der Einfluss des Industriestrompreises einen vernachlässigbaren Einfluss auf die Modellbildung der exponentiellen Regression. Im Durchlauf „Gesamt “ werden die beiden Faktoren Haushaltsstrompreis und Industriestrompreis kombiniert. Dieser erzielt ebenso einen R-Squared von 0,55 und einen MAE von 8879,59 und einen RMSE von 10616,65 Punkten. Gemessen an diesen Durchläufen erzielt dieser „Hhalt“ das beste Ergebnis, jedoch ist dieses Ergebnis immer nicht signifikant gut.

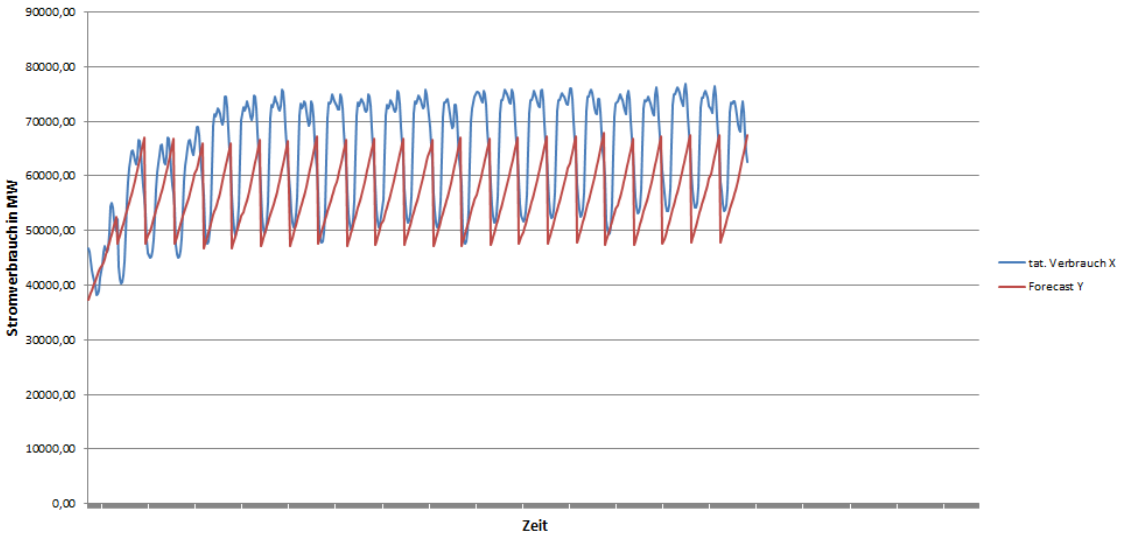


Abbildung 134: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_W “

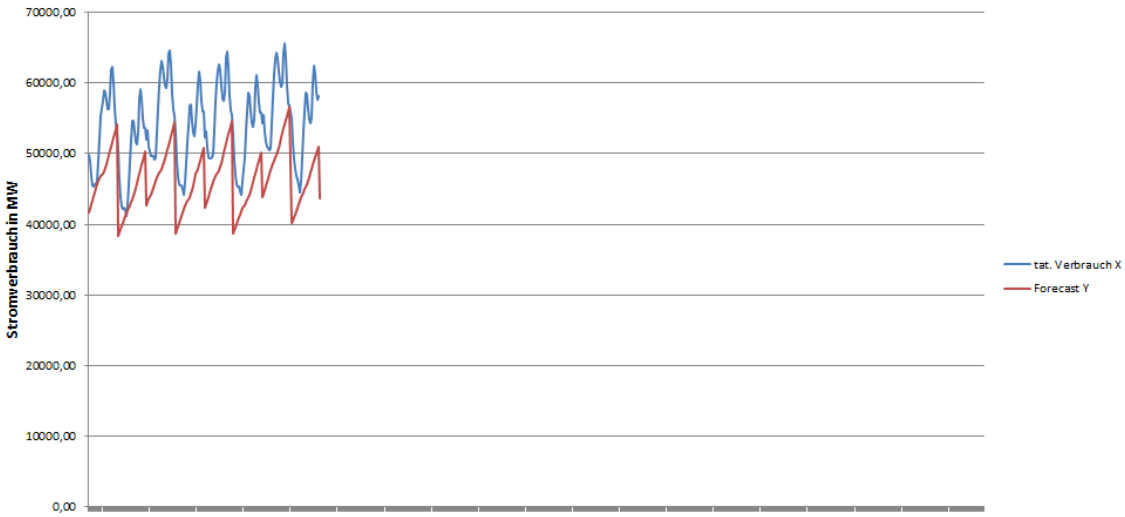


Abbildung 135: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_Wo “

Die Durchläufe „SplW“ und „SplWo“ erstellen entsprechende Prädiktionen für die Wochen beziehungsweise Wochenenden des Monats Januar 2014. Es fehlt der „Sockelbeitrag“ (Durchlauf „SplWo“). Dieses Verhalten gilt jedoch nicht für den Durchlauf „SplW“. Bezugnehmend auf die Fehlerkennzahlen in Tabelle 106 erreicht der Durchlauf „SplW“ einen R-Sqared von 0,44, einen MAE von 9200,19 und einen RMSE von 11072,8 Punkten. Der Durchlauf „SplWo“ erreicht einen R-Squared von 0,55, einen MAE von 7679,22 und einen RMSE von 8609,51 Punkten. Im Durchlauf „Splzusamm“ werden die Ergebnisse der Durchläufe „SplW“ und „SplWo“ wieder zusammengefügt. Dieser Durchlauf erzielt einen R-Squard von 0,56, welcher im Vergleich zum Durchlauf „Hhalt“ zwar zunächst besser erscheint. Vergleicht man jedoch MAE und RMSE von „Hhalt“ und „Splzusamm“, wird deutlich, dass hier der Durchlauf „Hhalt“ bessere Ergebnisse erzielt. Hierbei liegt der MAE

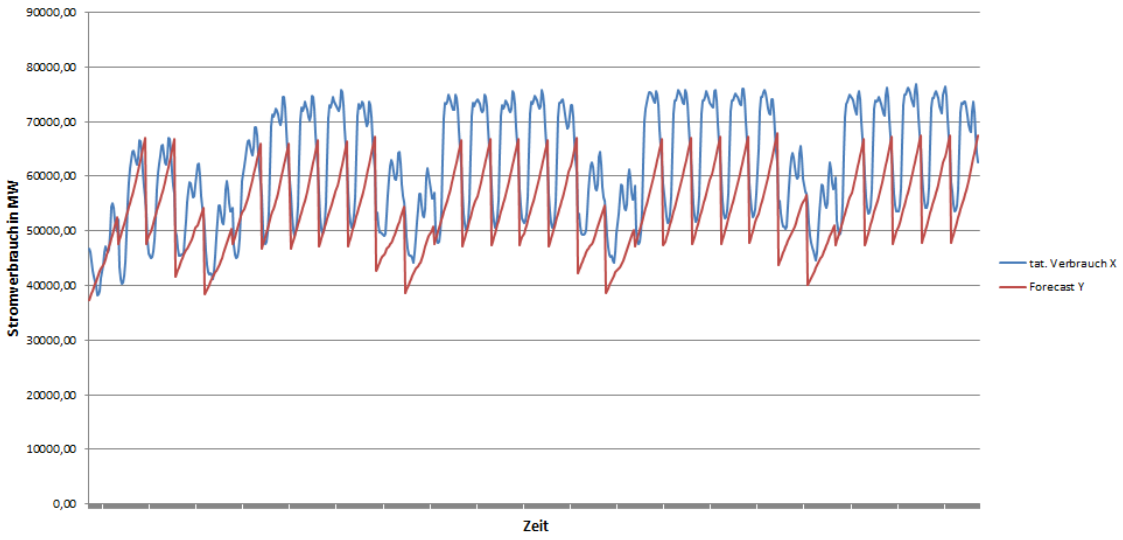


Abbildung 136: Diagramm Vergleich Ist und Forecast. Durchlauf: „Spl_zusamm “

vom Durchlauf „Hhalt“ um 766,15 niedriger als der Durchlauf „Splzusamm“. Dies gilt auch für den RMSE: Hier liegt der RMSE von Durchlauf „Hhalt“ um 764,85 Punkte niedriger als der Durchlauf „Splzusamm“.

Zusammenfassend lässt sich bezogen auf die exponentielle Regression für diese Hypothese feststellen, dass der Durchlauf „Hhalt“ das beste Ergebnisse liefert, wobei auch dieser Durchlauf keine signifikant guten Ergebnisse ergibt. Insbesondere hat in diesem Kontext auch die Aufteilung der Trainingsdatenmenge in Arbeitswochen und Wochenenden keine Verbesserung der Prädiktionswerte ergeben.

Support Vector Machine Alle Durchläufe zur Support Vector Machine können auf der beigelegten CD im Ordner `svn/Forecast/3.Hypothese/SVM/` gefunden werden. In diesen Ordnern sind folgende Dateien vorhanden:

Um das Trainingsset für den Durchlauf „Arbeitswoche/Feiertage + Wochenende“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend_workday_temp_industry_household.sql` folgendes SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM2_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WEEKEND" as
      ATTRIBUTE7, "WORKDAY" as ATTRIBUTE8, "AIRTEMP" as ATTRIBUTE9, "
      PRICE_INDUSTRY" as ATTRIBUTE10, "PRICE_HOUSEHOLD" as ATTRIBUTE11 FROM "
      PAL"."CONSUMPTION_TRAINING";
```

Abbildung 137: SQL-Statement für den gesamten Trainingsdatensatz bei SVM

Um das Trainingsset für den Durchlauf „Wochenende“ einzuschränken, wird in der

Datei	Inhalt
build_and_forecast_svm_yr_qt_mth _dom_dow_hod_weekend_workday _temp_household.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
build_and_forecast_svm_yr_qt_mth _dom_dow_hod_weekend_workday _temp_industry.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
build_and_forecast_svm_yr_qt_mth _dom_dow_hod_weekend_workday _temp_industry_household.sql	Script zum Erstellen des Modells und der Vorhersage für Januar 2014
household_industry_g0_001 #c1000.xlsx	Mehrere Dateien nach dem Muster (Durchlaufart)_g(gamma-Wert mit Unterstrich als Komma)#c(C-Wert mit Unterstrich als Komma).xlsx; Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 mitsamt Fehlerkennzahlen
Diagramm.xlsx	Vergleich der Vorhersagewerte mit den tatsächlichen Werten für Januar 2014 als Diagramm

Tabelle 107: Relevante Dateien für die Support Vector Machine

Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend_workday_temp_household.sql` folgendes SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM2_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WEEKEND" as
      ATTRIBUTE7, "WORKDAY" as ATTRIBUTE8, "AIRTEMP" as ATTRIBUTE9, "
      PRICE_HOUSEHOLD" as ATTRIBUTE10 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 138: SQL-Statement für den Trainingsdatensatz bezogen auf den Haushalt bei SVM

Um das Trainingsset für den Durchlauf „Wochenende“ einzuschränken, wird in der Datei `build_and_forecast_svm_yr_qt_mth_dom_dow_hod_weekend_workday_temp_industry.sql` folgendes SQL-Statement verwendet:

```
1 INSERT INTO PAL_SVM2_TRAININGSET_TBL
2 SELECT ID, "CONSUMPTION" as VALUEE, "YEAR" as ATTRIBUTE1, "QUARTER" as
      ATTRIBUTE2, "MONTH" as ATTRIBUTE3, "DAY_OF_MONTH" as ATTRIBUTE4, "
      DAY_OF_WEEK" as ATTRIBUTE5, "HOUR_OF_DAY" as ATTRIBUTE6, "WEEKEND" as
      ATTRIBUTE7, "WORKDAY" as ATTRIBUTE8, "AIRTEMP" as ATTRIBUTE9, "
      PRICE_INDUSTRY" as ATTRIBUTE10 FROM "PAL"."CONSUMPTION_TRAINING";
```

Abbildung 139: SQL-Statement für den Trainingsdatensatz bezogen auf die Industrie bei SVM

In der folgenden Tabelle sollen die verschiedenen Parameterpaare den bei den Durchläufen entstehenden Fehlerkennzahlen gegenübergestellt werden.

	Anpass-ungspara-meter		**Fehlerkennzahlen**				
Durchlauf	**γ**	**C**	**R^2**	**MAE**	**RMSE**	**CV(MAE)**	**CV(RMSE)**
HH	0,01	1000	0,71	9704,17	11283,55	0,16	0,18
HH+IN	0,01	1000	0,71	76032,01	76303,62	1,23	1,24
HH+IN	0,001	1000	0,53	9708,99	11706,28	0,16	0,19
IN	0,01	1000	0,71	76730,14	76993,75	1,25	1,25
IN	0,001	1000	0,53	9483,94	11482,53	0,15	0,18

Tabelle 108: Durchläufe SVM für die 3. Hypothese

Wie bereits erwähnt sind bei der Hypothese 3 drei Teildurchläufe durchzuführen. Der Eingabeparameter C wird konstant in allen Tests bei 1000 gehalten, während Parameter /*gamma* je nach Durchlauf zwischen 0,01/0,001 variiert. Der erste Teildurchlauf bezieht sich auf die Anwendung von Daten aus Hypothesen 1,2 sowie zusätzlich Strompreisen bei Haushalten (*HH*). Bei diesem Durchlauf sind die Eingabeparameter γ= 0,01 und C = 1000. Bei der Analyse von Fehlerkennzahlen fällt auf, dass R-Squared einen hohen und die CV (MAE) einen niedrigen Wert aufweist (siehe Tabelle 108). Auch die beiden Fehlerkennzahlen MAE und RMSE haben vergleichsweise niedrige Werte. Die Visualisierung in der Abbildung 140 zeigt den Kurvenverlauf. Beim zweiten Teildurchlauf sind die

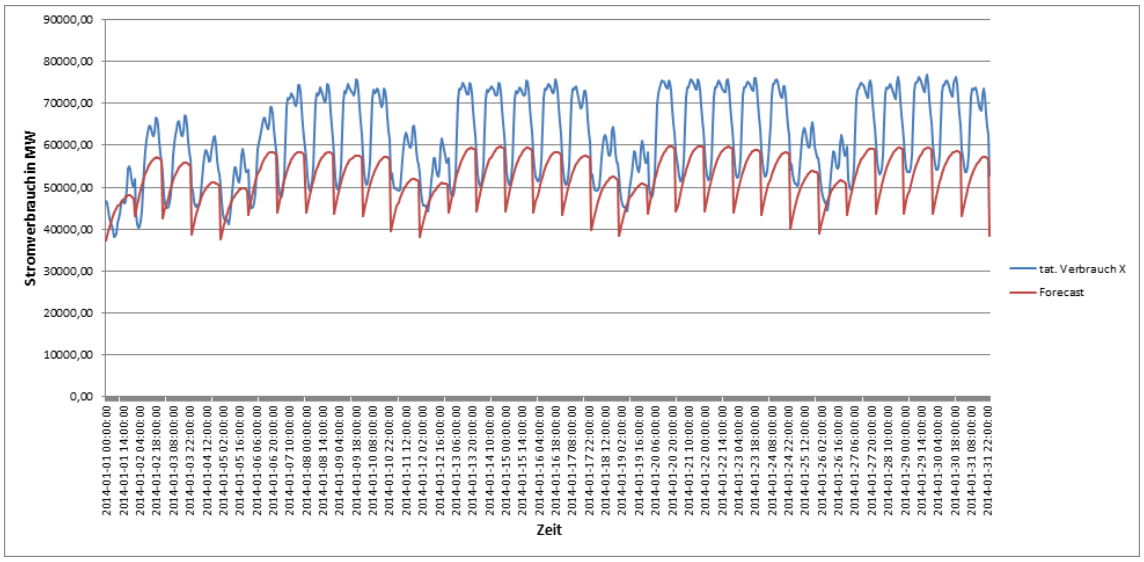


Abbildung 140: Support Vector Maschine mit $\gamma = 0,01$ und $C = 1000$

Daten aus Hypothesen 1, 2 zu beziehen und die Strompreise für die Industrie (*IN*) als zusätzliche Datenquelle zu nutzen. In diesem Fall ist der Eingabeparameter C jeweils bei 2 Tests konstant bei 1000, während die Anpassung beim Eingabeparameter γ (0,01 und

0,001) vorzunehmen ist. Beim ersten Durchlauf beträgt der Eingabeparameter $\gamma = 0,01$. Wie aus der Abbildung 141 hervorgeht, verschiebt sich die Funktion nach unten. Anzumerken ist es dabei, dass die Fehlerkennzahl R-Squared 0,71 ist und die Funktion visuell der Funktion mit dem tatsächlichen Verbrauch ähnelt. Allerdings weisen die Fehlerkennzahlen MAE und RMSE sehr hohe Werte von über 70000 auf. Somit wird erkennbar, dass das Modell weder genau noch zuverlässig ist.

Bei dem zweiten Durchlauf wird der Parameter γ angepasst und beträgt nun 0,001. Das hat eine Funktionsverschiebung nach oben zur Folge (siehe Abbildung 142. Allerdings sinkt dabei der R-Squared und beträgt R-Squared 0,53, was auch am Kurvenverlauf sichtbar ist. Die Fehlerkennzahlen MAE und RMSE haben sich durch die vorgenommene Änderung am Parameter deutlich verkleinert (siehe Tabelle 108.

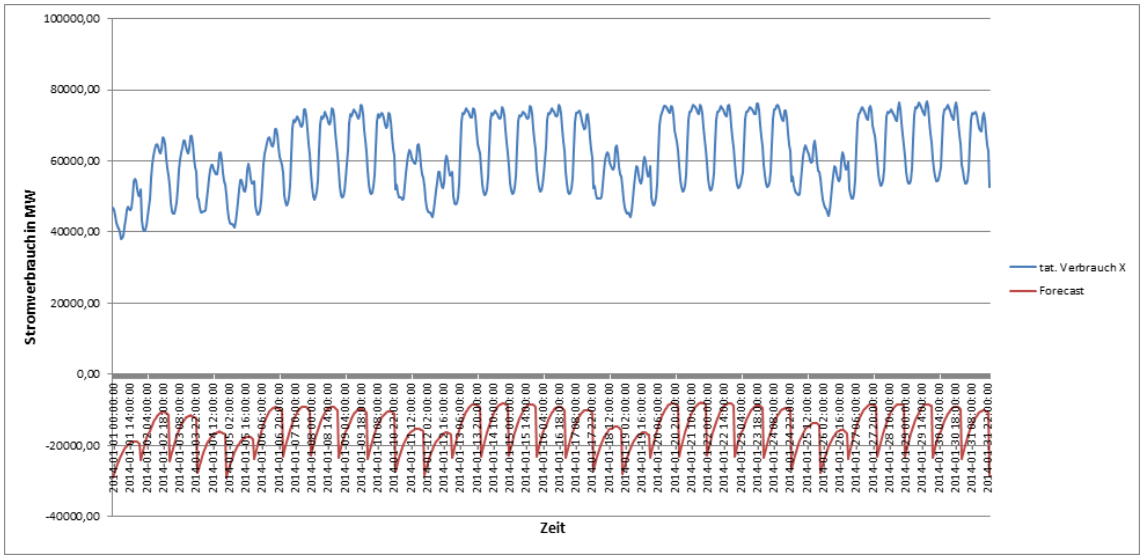


Abbildung 141: Support Vector Maschine, (IN) mit $\gamma = 0,01$ und $C = 1000$

Im Abschluss wird der dritte Testdurchlauf mit Daten aus Hypothesen 1, 2 und zusätzlich mit Strompreisen für die Industrie (IN)/Haushalt(HH) angereichert. Der Eingabeparameter C ist jeweils bei zwei Tests konstant bei 1000, während die Anpassung beim Eingabeparameter γ (0,01 und 0,001) vorzunehmen ist. Beim ersten Durchlauf beträgt der Parameter $\gamma = 0,01$. Nach der Abbildung 143, verschiebt sich die Funktion durch diesen γ-Wert nach unten. Es ist zu bemerken, dass die Fehlerkennzahl R-Squared = 0,71 ist und die Funktion rein abbildungstechnisch der Funktion mit dem tatsächlichen Verbrauch sehr ähnelt. Allerdings weisen die Fehlerkennzahlen MAE und RMSE sehr hohe Werte von über 70000 auf. Somit wird erkennbar, dass das auch dieses Modell weder genau noch zuverlässig ist.

Beim zweiten Durchlauf wird der Parameter γ angepasst und beträgt nun 0,001. Das hat die Funktionsverschiebung nach oben zur Folge (siehe Abbildung 145). Allerdings sinkt

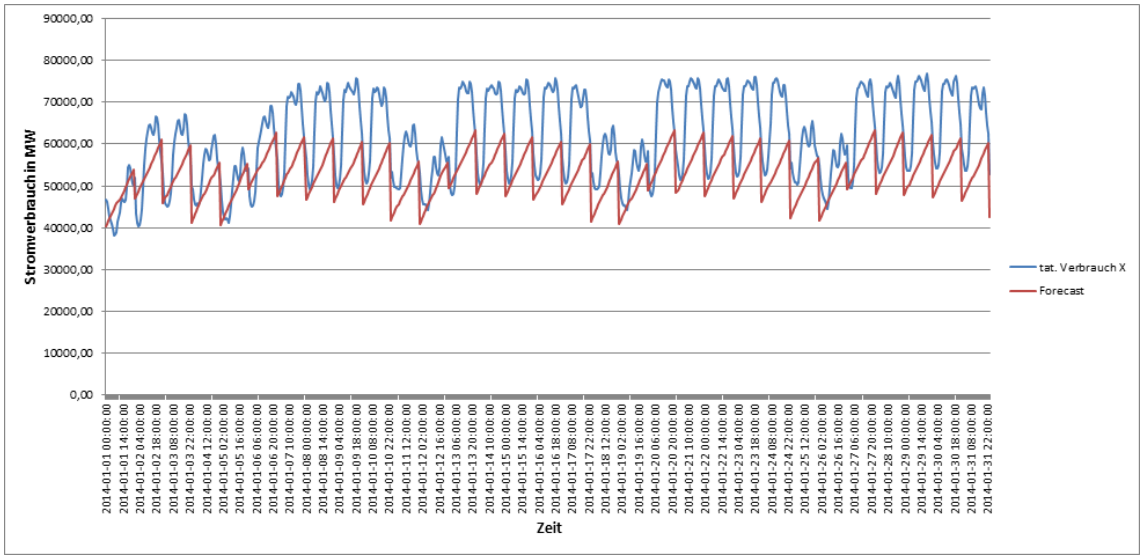


Abbildung 142: Support Vector Maschine,(IN) mit $\gamma = 0,001$ und $C = 1000$

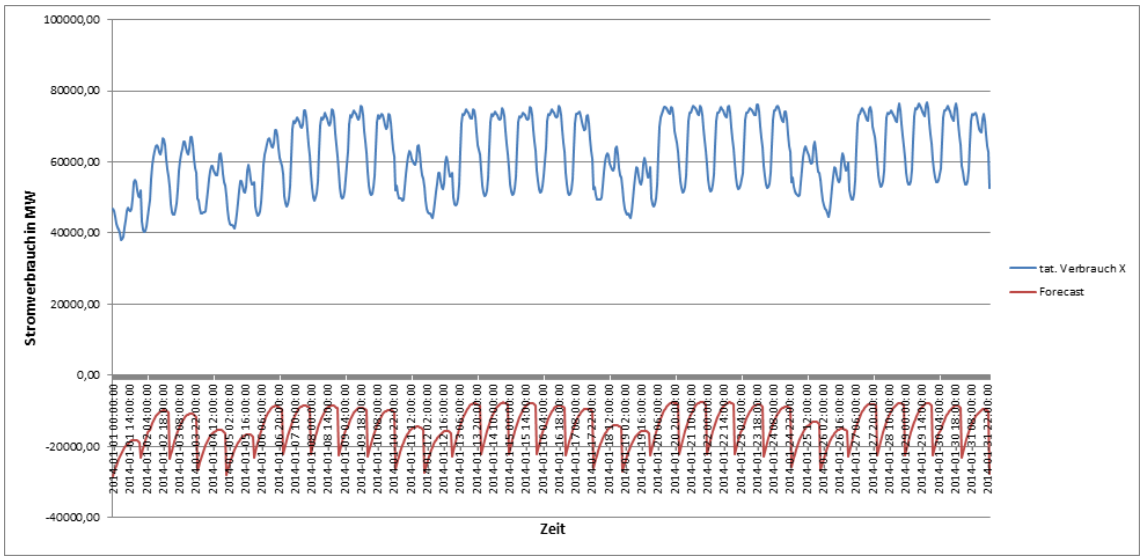


Abbildung 143: Support Vector Maschine, (IN/HH) mit $\gamma = 0,01$ und $C = 1000$

dabei der R-Squared und beträgt R-Squared = 0,53, was direkt am Kurvenverlauf sichtbar ist. Die Fehlerkennzahlen MAE und RMSE haben sich durch die vorgenommene Änderung am Parameter deutlich verkleinert (siehe Tabelle108).

Aus oben beschriebenen Ergebnissen lässt sich ableiten, dass der erste Durchlauf unter der Betrachtung der Daten aus Hypothesen 1,2 und Strompreisen für Haushalt (HH) mit den Eingabeparametern $\gamma = 0,01$ und $C = 1000$ zum besten Ergebnis führt.

Zusammenfassung der Ergebnisse Es hat sich gezeigt, dass die Ergebnisse der Testdurchläufe für die Energieverbrauchsvorhersage in allen getesteten Durchläufen zu keinen

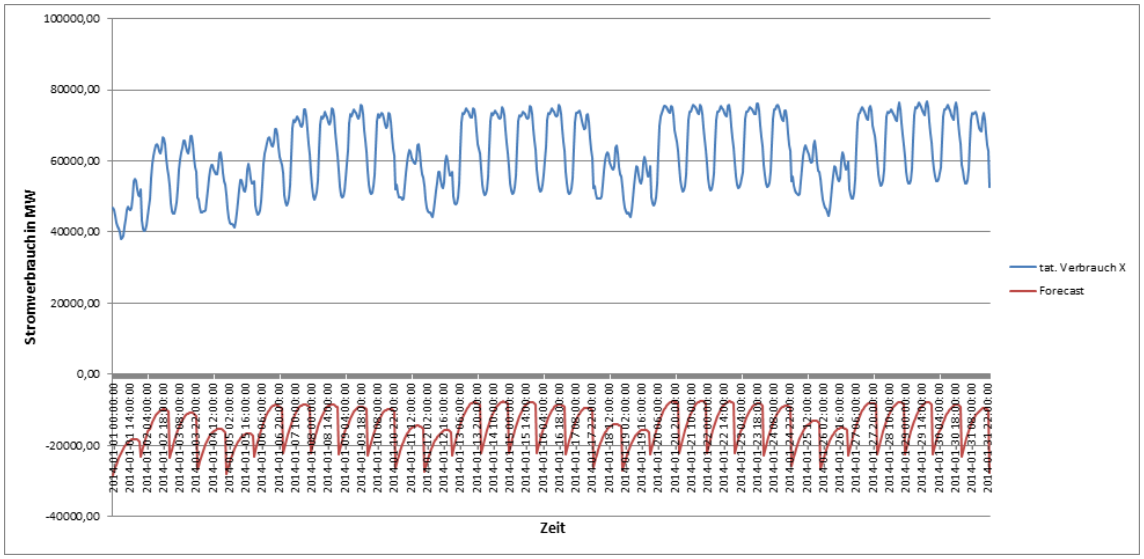


Abbildung 144: Support Vector Maschine, (IN/HH) mit $\gamma = 0,01$ und $C = 1000$

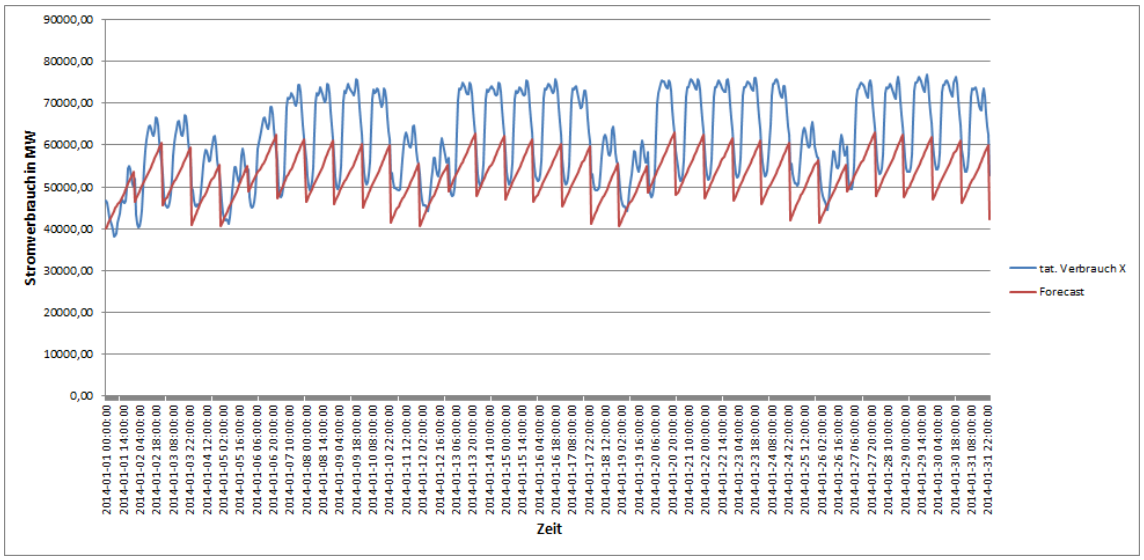


Abbildung 145: Support Vector Maschine,(IN/HH) mit $\gamma = 0,001$ und $C = 1000$

befriedigenden Ergebnissen kommen. Der Einsatz von verschiedenen Algorithmen hat gezeigt, dass die Fehlerkennzahlen immer in einem Bereich liegen, der vermuten lässt, dass die Vorhersagen nicht zuverlässig sind. Es kann mit den eingesetzten Methoden kein Zusammenhang zwischen den zusätzlich hinzugezogenen Parametern und dem tatsächlichen Stromverbrauch hergestellt werden:

Die dritte Hypothese lautet, dass die Prädiktionen für den Stromverbrauch genauere Ergebnisse liefern, wenn sie mit Informationen zu Wochenenden, Feiertagen (Workday), Luft Temperatur, Haushaltstrompreise und Industriestrompreis angereichert werden. Um diese Hypothese zu untersuchen, werden drei verschiedene Algorithmen, die multiple lineare Re-

gression, die exponentielle Regression und die Support Vector Machine in vier Variationen getestet:

Variation1 Haushaltpreise

Variation2 Industriepreise

Variation3 Gesamt (Haushaltpreise und Industriepresie)

Variation4.1 Split (Haushalt und Industriepreise in der Woche)

Variation4.2 Split(Haushalt und Industriepreise am Wochenende)

Variation4.3 Split(Haushalt und Industriepreise in der Woche und am Wochenende manuell in Excel)

Um die bestmögliche Kombination an Daten zu bestimmen, werden die Algorithmen multiplen linearen Regression und exponentiellen Regression jeweils sechs Mal durchgelaufen: Einmal nach einer Anreicherung der Daten nur mit Haushaltstrompreis (H-Halt), einmal nur mit Industriestrompreise (Ind), einmal mit Haushalt und Industriestrompreise (H-halt und Ind) zusammen, einmal Haushalt und Industriepreise in der Woche (Spl-W), einmal Haushalt und Industriepreise am Wochenende (Spl-Wo) und einmal Haushalt und Industriepreise (Spl-zusamm) in der Woche und am Wochenende manuell in Excel. SVM wird fünfmal mit unterschiedlichen Parametern getestet: dreimal $\gamma = 0,01$ und $C = 1000$ mit Haushaltstrompreis, Industriestrompreis und Haushalt- und Industriestrompreise zusammen und zwei Mal $\gamma = 0,001$ und $C = 1000$ mit Industriestrompreis und Haushalt- und Industriestrompreise zusammen.

Die folgenden drei Durchläufe beziehen sich auf die jeweils besten Durchläufe des eines jeden Algorithmus. Die Berechnung mit der multiplen linearen Regression von (Durchlauf H-Halt) ergab die folgenden Fehlerkennzahlen: R-Squared= 0,55 , MAE= 7387,86 und RMSE= 8948,69.

Die Werte der Fehlerkennzahlen der exponentiellen Regression von (Durchlauf Hhalt) lauten R-Squared = 0,54, MAE = 8042,53 und RMSE = 9727,77 .

Die Werte der Fehlerkennzahlen der Support Vector Machine mit den Anpassungsfaktoren $C = 1000$ und $\gamma = 0,01$ (Durchlauf HH) betragen. hier R-Squared = 0,71, MAE = 9704,17 und RMSE = 11283,55.

Aus den verschiedenen Durchläufen liefern die oben genannten drei die besten Ergebnisse. Im Vergleich der drei Ergebnisse ergibt sich, dass die Multiple Linearen Regression besten Ergebnisse erzielt.

12.7 Evaluation der Ergebnisse

Dieses Kapitel beschäftigt sich mit der Evaluation des Projektes. Hierzu wird wie folgt vorgegangen: Zunächst erfolgt die Durchführung eines Rankings der in den vorigen Kapiteln vorgestellten Durchläufe anhand der Methode Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE). Mit diesen Rankings kann der insgesamt beste Durchlauf aller Versuche bestimmt werden. Ebenfalls kann hiermit verifiziert werden, in wie weit die jeweiligen Unterhypothesen die Haupthypothese unterstützen. Außerdem wird der betriebswirtschaftliche Mehrwert beschrieben. Im Anschluss wird die Gesamtproblematik nochmals aufgegriffen und es erfolgt die Unterbreitung eines Vorschlags, wie die Prognosen weiter verbessert werden können. Hiernach erfolgt die Gesamtevaluation des Verlaufes der Projektgruppe anhand der in Kapitel 6 definierten Anforderungen.

12.8 Evaluation der Hypothesen

Das PROMETHEE bietet wie andere Outranking Methoden auch einen paarweisen Vergleich zwischen Alternativen [ZT11, S. 409]. Insgesamt gibt es laut Brans und Mareschal sechs verschiedene PROMETHEE Verfahren [BM05, S. 164]. In diesem Kontext wird PROMETHEE II (complete ranking) als Verfahren genutzt. Das Ranking erfolgt anhand der Fehlerkennzahlen R-Squared, MAE und RMSE. Dabei werden die Kennzahlen R-Squared als maximal, MAE und RMSE als minimal interpretiert. Bei den Kriterien handelt es sich um Kriterien mit linearer Präferenz (criterion with linear preference). Dieser Kriterientyp wird gewählt, da eine linear steigende Präferenz des Entscheidungsträgers mit steigender Differenz der Werte von Alternativen abgebildet werden soll [BV85, Vgl. S. 650ff.]. Das bedeutet beispielsweise, dass eine hohe Differenz im MAE zwischen zwei Alternativen sich höher im Ranking widerspiegeln soll als eine niedrige. Allen Kennzahlen ist die Gewichtung 1 zugeteilt, somit sind die Kennzahlen alle gleich gewichtet. Der F-Wert (auch net outranking flow genannt) definiert, wie hoch die Alternative anhand der gesetzten Kriterien bewertet wird [BV85, Vgl. S. 654].

Ranking	Alternative	F	F+	F-
1	**LSDNA-Woe-14**	**0,75999**	**0,77645**	**0,01646**
2	**LSDNA-Zusammen**	**0,73300**	**0,74979**	**0,01679**
3	**LSDNA-Wo-14**	**0,71950**	**0,73670**	**0,01720**
4	**MLR-H2-Spl-Woe**	**0,66995**	**0,68534**	**0,01539**
5	**MLR-H2-Jan09-13**	**0,62584**	**0,66658**	**0,04074**
6	**Exp.R-H2-Spl-Woe**	**0,60073**	**0,65024**	**0,04951**
7	**MLR-H3-Spl-We**	**0,57942**	**0,64007**	**0,06066**
8	**MLR-H2-Spl-Zusam**	**0,55810**	**0,63007**	**0,07198**
9	**MLR-H3-H-Halt**	**0,53489**	**0,61732**	**0,08243**

10	**Exp.R-H3-Spl-We**	**0,52666**	**0,61320**	**0,08654**
11	**MLR-H2-Temp-Fe-Wo**	**0,51843**	**0,60909**	**0,09066**
12	**MLR-H1-Fe-Wo**	**0,49859**	**0,59905**	**0,10045**
13	**MLR-H3-Spl-Zusam**	**0,44287**	**0,57246**	**0,12959**
14	**MLR-H2-Spl-Wo**	**0,44205**	**0,57300**	**0,13095**
15	**LSDNA-Jan-14**	**0,43044**	**0,56773**	**0,13728**
16	**Exp.R-H2-Slit-Zusam**	**0,40979**	**0,55394**	**0,14416**
17	**MLR-H1-Fe**	**0,38444**	**0,54419**	**0,15975**
18	**Exp.R-H3-H-Halt**	**0,38337**	**0,54143**	**0,15807**
19	**MLR-H1-Wo**	**0,37933**	**0,54205**	**0,16272**
20	**MLR-H3-H-Halt-Ind**	**0,36863**	**0,53337**	**0,16473**
21	**Exp.R-H2-Temp-Fe-Wo**	**0,34386**	**0,52086**	**0,17700**
22	**MLR-H3-Ind**	**0,32913**	**0,51444**	**0,18531**
23	**SVM-H1-We**	**0,30337**	**0,50728**	**0,20391**
24	**MLR-H2-Jan2013**	**0,29472**	**0,49839**	**0,20366**
25	**MLR-H2-Temp**	**0,24221**	**0,48012**	**0,23790**
26	**MLR-09-13**	**0,22575**	**0,47188**	**0,24613**
27	**MLR-H3-Spl-Wo**	**0,22468**	**0,46485**	**0,24016**
28	**SVM-09-13/11**	**0,20419**	**0,45312**	**0,24893**
29	**MLR-H2-2013**	**0,20081**	**0,45077**	**0,24996**
30	**Exp.R-H2-Spl-Wo**	**0,18501**	**0,44649**	**0,26148**
31	**Exp.R-H3-Spl-Zusam**	**0,18435**	**0,44254**	**0,25819**
32	**Exp.R-H1-Fe-Wo**	**0,16114**	**0,43032**	**0,26918**
33	**SVM-09-13/13**	**0,15184**	**0,43053**	**0,27868**
34	**Exp.R-H3-H-Halt-Ind**	**0,12337**	**0,41156**	**0,28819**
35	**Exp.R-H3-Ind**	**0,08707**	**0,39328**	**0,30621**
36	**SVM-09-13/12**	**0,07653**	**0,39077**	**0,31424**
37	**SVM-H2-Temp/3**	**0,06937**	**0,39370**	**0,32432**
38	**SVM-H2-Temp/5**	**0,04995**	**0,37987**	**0,32992**
39	**SVM-09-13/9**	**0,03176**	**0,37147**	**0,33971**
40	**SVM-09-13/7**	**-0,01104**	**0,35032**	**0,36136**
41	**Exp.R-H2-Temp**	**-0,02791**	**0,34793**	**0,37584**
42	**Exp.R-H1-Fe**	**-0,03869**	**0,33316**	**0,37185**
43	**Exp.R-H3-Spl-Wo**	**-0,04544**	**0,33127**	**0,37671**
44	**SVM-H2-Temp-Wo-We/10**	**-0,06404**	**0,31983**	**0,38387**
45	**Exp.R-H1-Wo**	**-0,07013**	**0,31892**	**0,38905**
46	**Exp.R-H1-2013Fe-Wo**	**-0,10709**	**0,29633**	**0,40342**
47	**SVM-H3-H-Halt/1**	**-0,12717**	**0,30275**	**0,42992**

48	**SVM-H1-Wd**	**-0,12881**	**0,28600**	**0,41481**
49	**SVM-09-13/5**	**-0,13153**	**0,29464**	**0,42617**
50	**SVM-H2-Temp-Wo-We/9**	**-0,13853**	**0,28057**	**0,41909**
51	**SVM-H2-Temp/2**	**-0,16108**	**0,27847**	**0,43955**
52	**Exp.R.Jni-Dez13**	**-0,16635**	**0,28184**	**0,44819**
53	**SVM-H3-Ind/5**	**-0,18844**	**0,25534**	**0,44379**
54	**SVM-H2-Temp/4**	**-0,22355**	**0,24612**	**0,46967**
55	**SVM-H2-Temp-Wo-We/8**	**-0,25227**	**0,22427**	**0,47654**
56	**SVM-H3-H-Halt-Ind/3**	**-0,27844**	**0,21061**	**0,48905**
57	**Exp.R.-09-13**	**-0,29128**	**0,21625**	**0,50753**
58	**MLR-H2-Dez2013**	**-0,31182**	**0,23131**	**0,54313**
59	**Exp.R-H1-2013Fe**	**-0,31515**	**0,19440**	**0,50955**
60 – 61	**SVM-H1-Wd**	**-0,31960**	**0,18592**	**0,50551**
60 – 61	**SVM-H1-Wd-We**	**-0,31960**	**0,18592**	**0,50551**
62	**Exp.R-H1-2013Wo**	**-0,32190**	**0,19226**	**0,51416**
63	**SVM-09-13/2**	**-0,32849**	**0,19464**	**0,52313**
64	**SVM-H2-Temp-Wo-We/7**	**-0,36602**	**0,16814**	**0,53416**
65	**SVM-H1-Wd-We**	**-0,42347**	**0,15460**	**0,57807**
66	**SVM-09-13/1**	**-0,43268**	**0,14267**	**0,57535**
5	**MLR-H2-Jan09-13**	**0,62584**	**0,66658**	**0,04074**
67	**SVM-09-13/6**	**-0,44248**	**0,15884**	**0,60132**
68	**SVM-H2-Temp/1**	**-0,44914**	**0,13444**	**0,58358**
69	**SVM-H1-Wd**	**-0,45005**	**0,13744**	**0,58749**
70	**SVM-H2-Temp-Wo-We/11**	**-0,45301**	**0,14131**	**0,59432**
71	**SVM-09-13/4**	**-0,46560**	**0,12621**	**0,59181**
72	**MLR-2013**	**-0,49515**	**0,11032**	**0,60547**
73	**SVM-H3-H-Halt-Ind/2**	**-0,50577**	**0,11345**	**0,61922**
74	**Exp.R.-2013**	**-0,51836**	**0,10123**	**0,61959**
75	**SVM-H3-Ind/4**	**-0,53869**	**0,09699**	**0,63568**
76	**SVM-H2-Temp6**	**-0,56610**	**0,06880**	**0,63490**
77	**SVM-09-13/15**	**-0,58215**	**0,06559**	**0,64774**
78	**SVM-09-13/16**	**-0,58717**	**0,06226**	**0,64942**
79	**SVM-09-13/10**	**-0,59145**	**0,06929**	**0,66074**
80	**SVM-09-13/14**	**-0,59186**	**0,06295**	**0,65481**
81	**SVM-09-13/3**	**-0,61300**	**0,11523**	**0,72822**
82	**SVM-09-13/8**	**-0,61466**	**0,05514**	**0,66979**

Tabelle 109: Promethee-Ranking zu den Durchläufen

Wie in Tabelle 109 ersichtlich wird, belegen die drei Durchläufe des Algorithmus Lineare Regression mit gedämpftem Trend und saisonaler Anpassung der Datenbasis die ersten drei Plätze des Rankings. Hierbei ist zu beachten, dass der Durchlauf „LSDNA-Woe-14“ die Prognosen für die Wochenenden des Januar 2014 erstellt und der Durchlauf „LSDNA-Wo-14“ die Prognosen für die Arbeitstage des Januar 2014 berechnet. Der Durchlauf „LSDNA-Zusammen“ fügt beide Prognosen zu einer ganzheitlichen Prognose zusammen. Dieser Versuch wird in Abbildung 146 an dieser Stelle nochmals grafisch illustriert. Von allen getesteten Algorithmen ist die lineare Regression mit gedämpftem Trend und saisonaler Anpassung der einzige Algorithmus, welcher in der Lage ist, die Stromverbrauchsschwankungen innerhalb eines Tages mit hoher Annäherung prognostiziert (siehe auch Abschnitt 12.3.1). Ebenfalls werden mit dem Versuch „LSDNA-Zusammen“ die Stromverbrauchsschwankungen zwischen Arbeitstagen und Wochenenden mit hoher Annäherung prognostiziert. Die von diesem Durchlauf produzierte Prognose erzielt folgende Fehlerkennzahlen: Der R-Squared liegt bei 0,82, der MAE liegt bei 6404,69 und der RMSE liegt bei 7097,30. Dies sind die besten Werte aller durchgeführten Durchläufe über alle Hypothesen. Insbesondere wird anhand des R-Squared die gute Anpassung des prognostizierten Stromverbrauches im Vergleich zum tatsächlichen Stromverbrauch sichtbar. Wie in Kapitel 12.3.1 bereits angedeutet, fehlt dieser Prognose jedoch ein Sockelbetrag von ca. 5000-6000 Megawatt (MW), welche die Prognosequalität weiter verbessern könnte.

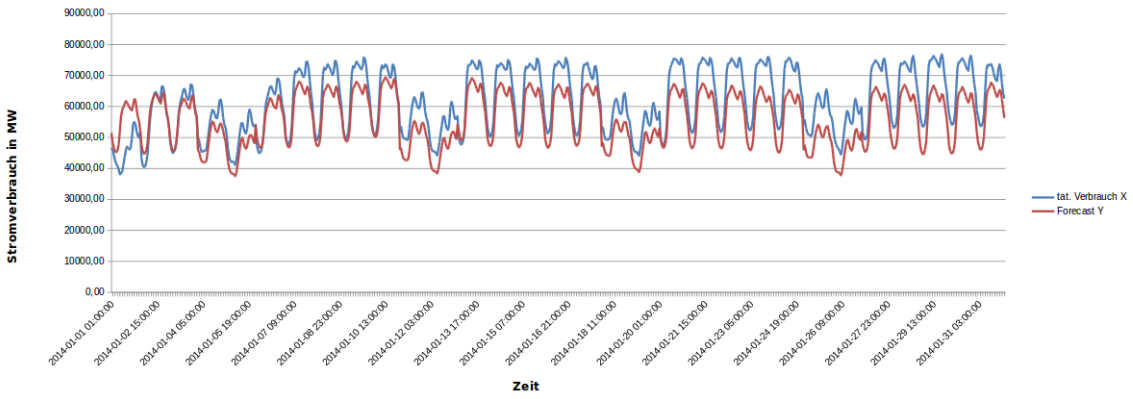


Abbildung 146: Vergleich von tatsächlichen Stromverbrauch und vorhergesagten Verbrauch. Versuch: „LSDNA-Zusammen“

Der Algorithmus erlaubt lediglich die Prognose anhand eines Parameters, was der historische Stromverbrauch selbst ist. Zusätzliche unabhängige Variablen, zum Beispiel der Einfluss des Wetters oder Angaben zu Feiertagen oder Strompreisen können bei der Modellberechnung nicht berücksichtigt werden, so dass im Anschluss eine separate Betrachtung der Haupthypothese und der Unterhypothesen (siehe 3.2) mit den Algorithmen multiple lineare Regression, exponentielle Regression sowie der Support Vector Machine erfolgt. Dies wird durch die Tabelle 109 unterstützt. Hierbei ist zu beachten, dass keine Beachtung der geteilten Durchläufe (Split von Wochenende und Arbeitswoche) stattfindet - hier

werden nur die zusammengefügten Ergebnisse betrachtet. Zusätzlich gilt: Die Menge an verarbeiteten Daten wird pro Hypothese erhöht. Demnach enthält die Datenbasis die geringste Datenmenge und die dritte Hypothese enthält die größte Datenmenge. Dies wird über die verschiedenen Features der jeweiligen Hypothesen realisiert.

Multiple Lineare Regression Im ersten Versuch der Datenbasis kann mit Hilfe der multiplen linearen Regression keine signifikanten Ergebnisse erzielt werden. Erst im zweiten Versuch der Datenbasis, in dem der Verlauf der Zeit differenzierter für den Algorithmus modelliert wird, ist es möglich mit Hilfe der multiplen linearen Regression erste brauchbare Ergebnisse zu erzielen. Die relevanten Ergebnisse aus allen Hypothesen - bezogen auf die multiple lineare Regression - ist in Tabelle 110 ersichtlich.

Durchlauf	**R-Squared**	**MAE**	**RMSE**	**CV_MAE**	**CV_RMSE**
2009-2013 (Datenbasis)	0,39	8229,84	10049,13	0,13	0,16
Fe_Wo (1. Hypothese)	0,54	7585,87	9165,8	0,12	0,14
Split-zusammen (2. Hypothese)	0,57	7268,98	8804,53	0,11	0,14
H-Halt (3. Hypothese)	0,55	7387,86	8948,69	0,11	0,14

Tabelle 110: Fehlerkennzahlen der besten Prognosen der multiplen lineare Regression für alle Hypothesen

Wie in der Tabelle ersichtlich wird, steigt je Hypothese der R-Squared Wert, was einer steigenden Anpassung der von den Algorithmus produzierten Vorhersagen an den tatsächlichen Verlauf des Stromverlaufes entspricht. Ebenso sinkt der MAE- und RMSE Wert je Hypothese, was einer höheren Prognosegenauigkeit der Werte zum tatsächlichen Stromverbrauch entspricht. An dieser Stelle wird die Haupthypothese bis auf die dritte Hypothese bestätigt. Dies ist wahrscheinlich der Granularität der Strompreisdaten geschuldet: Hier liegen die Strompreisdaten für Haushalt und Industrie lediglich in jährlicher Form vor.

Exponentielle Regression Analog zur multiplen linearen Regression produziert die exponentielle Regression erst mit der differenzierten Darstellung der Zeit ab dem zweiten Versuch der Datenbasis brauchbare Ergebnisse. Eine Auswahl der besten Ergebnisse über alle Hypothesen ist in Tabelle 111 ersichtlich.

Wie bereits bei der multiplen linearen Regression festgestellt, wird auch bei der exponentiellen Regression der R-Squared mit steigender Anzahl von Daten höher, was einer steigenden Anpassung der von den Algorithmus produzierten Vorhersagekurve an die tatsächliche Stromverbrauchskurve entspricht. Ebenso sinkt der MAE- und RMSE Wert bei steigender Anzahl von Daten. An dieser Stelle wird die Haupthypothese bis auf die dritte Hypothese

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
Juni2013 (Datenbasis)	0,35	9335,74	11272,23	0,15	0,18
Fe_Wo 2009-2013 (1. Hypothese)	0,54	8810,64	10540,96	0,14	0,17
Spl_zusammen (2. Hypothese)	0,55	8005,92	9664,11	0,13	0,16
H_Halt (3. Hypothese)	0,54	8041,53	9727,77	0,13	0,16

Tabelle 111: Fehlerkennzahlen der besten Prognosen der exponentiellen Regression für alle Hypothesen

bestätigt. Auch hier lässt sich dieses Verhalten wahrscheinlich auf die bereits im vorigen Abschnitt genannte jährliche Granularität der Strompreisdaten zurückführen.

Support Vector Machine Bei der multiplen linearen Regression und der exponentiellen Regression ist ein Trend erkennbar: Es verbessern sich die Fehlerkennzahlen bei steigender Datenmenge. Dies ist bei der Support Vector Machine nicht der Fall.

Durchlauf	R-Squared	MAE	RMSE	CV_MAE	CV_RMSE
11 (Datenbasis)	0,57	8814,5	10230,36	0,14	0,17
WE (1. Hypothese)	0,62	8299,08	9815,41	0,13	0,16
T (2. Hypothese)	0,39	8825,68	10634,47	0,14	0,17
(3. Hypothese)	0,71	9704,17	11283,55	0,16	0,18

Tabelle 112: Fehlerkennzahlen der besten Prognosen der Support Vector Machine für alle Hypothesen

Gemessen an den Fehlerkennzahlen erzielt der Durchlauf „WE“der ersten Hypothese das beste Ergebnis aller getesteten Hypothesen. Die Ergebnisse der zweiten und dritten Hypothese sind allesamt schlechter als die Ergebnisse der ersten Hypothese. Insbesondere sind sogar die Ergebnisse der dritten Hypothese schlechter als die der zweiten Hypothese. Daher ist kein Trend zu erkennen und die Haupthypothese kann mit der Support Vector Machine nicht bestätigt werden.

Zentrale Schlussfolgerungen zu den verwendeten Algorithmen Basierend auf den Erkenntnissen der multiplen lineare Regression sowie der exponentiellen Regression ist ersichtlich, dass eine größere Trainingsdatenmenge zu besseren Prognoseergebnissen führt. Ausgenommen hiervon ist die dritte Hypothese. Daraus ist zu folgern, dass die Qualität und Granularität aber auch die Relevanz der Daten eine wichtige Rolle hinsichtlich der Prognosen spielt. Dennoch bleiben die produzierten Ergebnisse der multiplen linearen und exponentiellen Regression trotz steigender Datenmenge hinter den Erwartungen zurück.

Insbesondere kann mit den berechneten Werten keine verlässliche Stromverbrauchsprognose für den Betrachtungszeitraum erstellt werden. In diesem Kontext hat zwar eine Einteilung der Modellbildung in Arbeitstagen und Wochenenden eine leichte Verbesserung der Prognosewerte erzielt, jedoch konnte auch mit diesem Vorgehen keine verlässliche Prognose erstellt werden. Ebenfalls fällt auf, dass bei den verwendeten Algorithmen oftmals die Stromverbrauchshöhen der jeweiligen Tage nicht korrekt modelliert werden. Hervorzuheben sind an dieser Stelle die Beschreibungen der Durchläufe in Kapitel 12.3, 12.4, 12.5 und 12.6. Wie bereits festgestellt, kann mit der Support Vector Machine die Haupthypothese nicht bestätigt werden. Insbesondere bessert sich nicht die Prognosegenauigkeit bei steigender Datenmenge. Der Durchlauf „LSDNA-Zusammen“ erzielt mit dem Algorithmus lineare Regression mit gedämpftem Trend und saisonaler Anpassung das beste Ergebnis aller Versuche. Da dieser Algorithmus jedoch keine N unabhängigen Variablen zur Berechnung von Prognosen zulässt, erfolgt an dieser Stelle zusätzlich die Benennung des besten Durchlaufes des Algorithmus, der N Variablen in die Prognoseberechnungen zulässt. Dies ist der Versuch „MLR-H2-Jan09-13“ der multiplen linearen Regression, welcher sich im Ranking in Tabelle 109 auf Platz 5 befindet. Generell befinden sich die zusammengefügten Prognosen der multiplen linearen Regression auf den vorderen Plätzen des Rankings (Platz 7, 8 und 9). Hieraus lässt sich schließen, dass die multiple lineare Regression unter den Algorithmen, die N Variablen für die Prognoseberechnung einbeziehen, die besten Prognosen im Vergleich zur exponentiellen Regression und der Support Vector Machine erstellt. Auf den nachfolgenden Plätzen (16, 18, 21 ...) des Rankings befinden sich Durchläufe der exponentiellen Regression. Daraus lässt sich schließen, dass die exponentielle Regression unter den Algorithmen, die N Variablen für die Prognoseberechnung einbeziehen, vor der Support Vector Maschine die zweitbesten Prognosen erzeugt. Die Support Vector Maschine erzeugt als verbleibender Algorithmus mit N Variablen für die Prognoseberechnung die drittbesten Prognosen.

12.9 Betriebswirtschaftlicher Mehrwert

In Kapitel 2.4 ist definiert, dass im Anschluss evaluiert werden kann, ob betriebswirtschaftliche Vorteile unter der Verwendung von SAP HANA im Energiesektor entstehen und ob dies zu einer Optimierung des Planungsprozesses führt. SAP HANA ist im Rahmen dieser Projektarbeit ein Werkzeug zur Durchführung der Versuche. Aus technischer Sicht ist es grundsätzlich denkbar die Prognosen auch mit Hilfe eines anderen DBMS durchzuführen. SAP HANA bietet an dieser Stelle den Vorteil, dass Energieversorger die betriebswirtschaftlichen Funktionsbereiche sowie die Prognosemodelle in einem Gesamtsystem integrieren können. Dies unterstützt auch die Transparenz im Planungsprozess. Außerdem stellt sich die Frage, ob diese Lösung betriebswirtschaftliche Vorteile generiert. Im Rahmen dieser Projektarbeit sind prototypische Umsetzungen erfolgt. Für eine in der Praxis nutzbare Prognose sollte die Prognosegenauigkeit durch weitere Forschung verbessert wer-

den. Außerdem gilt es im Anschluss praktisch Nutzbare Module für einen Endanwender zu entwickeln. Ein Beispiel für ein solches Modul ist eine grafische Benutzeroberfläche. Sind diese Voraussetzungen erfüllt, ist eine praktische Nutzung dieser Forschungsergebnisse denkbar.

12.10 Ausblick

In diesem Abschnitt soll dargestellt werden, wie die Prognosen mit den behandelten Algorithmen und dem verwendeten System weiter verbessert werden könnten.

Differenziertere Modellbildung Die Ergebnisse der einzelnen Durchläufe des linearen Regression mit gedämpften Trend und saisonaler Anpassung, multiple lineare Regression und exponentielle Regression haben gezeigt, dass die Durchläufe, in denen Wochenenden und Arbeitswochen getrennt betrachtet - und anschließend zu einer ganzheitlichen Prognose zusammengefügt werden - zu besseren Ergebnissen als die ganzheitliche Betrachtung der gesamten Trainingsdatenmenge führen. Dieser Ansatz könnte weiter verfolgt werden: Beispielsweise wäre eine isolierte Betrachtung einzelner Wochentage denkbar. In diesem Fall wird für jeden einzelnen Wochentag ein Modell aus den Trainingsdaten gebildet. Jedes Modell erstellt anschließend Prognosen für den entsprechenden Wochentag des Prognosezeitraumes. Diese einzelnen Prognosen werden anschließend wieder zu einer ganzheitlichen Prognose des Betrachtungszeitraumes zusammengefügt. Dieses Vorgehen kann sogar auf einzelne Stunden eines Wochentages erweitert werden, was den Aufwand zur Modell- und Prognosebildung und dem anschließenden Zusammenfügen der jeweiligen Prognosen jedoch erheblich erhöht.

Hinzufügen weiterer Features Eine weitere Verbesserung kann durch das Hinzufügen weiterer Features realisiert werden. Beispielsweise könnte eine Betrachtung weiterer kalendarischer (zum Beispiel Weihnachts- und Sommerferienindex) und meteorologischer Daten (zum Beispiel die Windgeschwindigkeit) zu einer Verbesserung der Prognosequalität führen. Auch ist an dieser Stelle eine tiefergehende differenziertere Darstellung der Zeit im Vergleich zur ersten Hypothese (Siehe Abschnitt 12.4) denkbar. Die Daten eines Tages könnten zum Beispiel mit folgenden Angaben ausgestattet werden:

- Der Zeitraum des höchsten Arbeitsaufkommens.
- Der Zeitraum an dem am häufigsten Mittagspause gemacht wird.
- Wann schläft der größte Teil der Bevölkerung.
- Wann sind die produktivsten Stunden eines Tages, zum Beispiel in der Industrie.
- Wann ist die unproduktivsten Stunden eines Tages, zum Beispiel in der Industrie.

Hierbei ergibt sich jedoch das Problem, dass geeignete, verlässliche Datenquellen für solche Informationen beschafft werden müssen. Solche Daten müssten im Idealfall stündlich, geografisch differenziert und für jeden Wochentag gesondert vorliegen. Eine Alternative hierzu ist das Schätzen solcher Daten.

verkürzte Prognose Insbesondere bei der linearen Regression mit gedämpften Trend und saisonaler Anpassung wird für die ersten 72 Stunden des Prognosezeitraumes eine im Vergleich zu anderen Prognosen gute Anpassung der prognostizierten Werte an den tatsächlichen Verbrauch sichtbar (Siehe Grafik 147). Hiermit wird ein R-Squared von 0,85, ein MAE von 6575,53 und ein RMSE von 7586,53 erreicht.

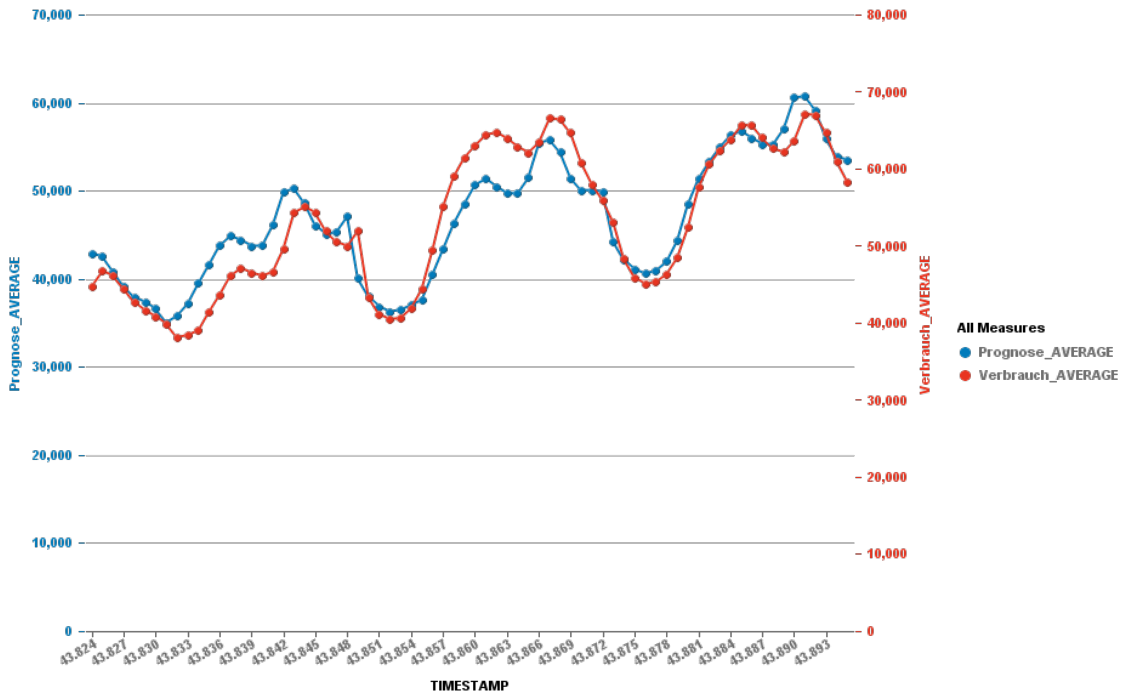


Abbildung 147: Vergleich zwischen tatsächlichen und prognostizierten Stromverbrauch mit der lineare Regression mit gedämpften Trend und saisonaler Anpassung

Ein möglicher Ansatz die Prognosen für den gesamten Trainingszeitraum zu verbessern ist der folgende: Zunächst wird ein Modell auf Basis der gesamten Trainingsdaten (2009-2013) gebildet. Mit diesen Modell werden die ersten 72 Stunden des Betrachtungszeitraumes vorhergesagt. Die von dem Modell erstellten Vorhersagen gehen anschließend in die Modellberechnung für die nächsten 72 Stunden ein. Diese Vorgehensweise wird anschließend wiederholt, bis das Ende des Betrachtungszeitraumes erreicht ist.

Einschränkung des geografischen Betrachtungsraumes Die Projektgruppe fokussiert die Vorhersage von Stromprognosen für den geografischen Bereich Deutschland. Da dieser Bereich sehr groß ist und aus diesem Grund davon auszugehen ist, dass die verwendeten

Daten unvollständig sind, beziehungsweise gemittelt werden[13], kann dies zur Verzerrung der Prognosen führen. Aus diesem Grund könnte auch die Betrachtung eines wesentlich kleineren geografischen Bereiches (zum Beispiel eine Stadt oder eine Gemeinde) zu genaueren Prognosen führen. Möglicherweise verbessert sich die Prognosegenauigkeit, wenn auf Daten zugegriffen wird, die allgemein eine höhere Datenqualität aufweisen und im Speziellen geografisch differenziert sind.

Hiermit ist die Evaluation der Ergebnisse und der Ausblick über mögliche Verbesserungen der Prognosequalität abgeschlossen. Im folgenden erfolgt eine Gegenüberstellung der im Kapitel 6 definierten Anforderungen zu den tatsächlich umgesetzten Anforderungen.

12.11 Evaluation der Anforderungen

Die folgende Übersicht zeigt die Gegenüberstellung der in Kapitel 6 gestellten Kriterien und Qualitätsanforderungen mit der tatsächlich erfolgten Umsetzung.

Legende:

Symbol	**Bedeutung**
X	Anforderung erfüllt
0	Anforderung nicht erfüllt

Tabelle 113: Legende

Wie in Tabelle 114 ersichtlich, werden sämtliche Musskriterien erfüllt.

Das Wunschkriterium 2.1 *Automatisierung des gesamten ETL-Prozesses* aus Tabelle 115 kann nicht umgesetzt werden, da die notwendige Hardware-Infrastruktur nicht bereitgestellt wurde.

Tabelle 116 zeigt, dass sämtliche an die Qualität der Umsetzung gestellten Anforderungen eingehalten werden.

[13] Beispiel: Die Wetterdaten des DWD stammen von vielen einzelnen Stationen, die über ganz Deutschland verteilt sind. Für die Stromverbrauchsprognosen wird das arithmetische Mittel dieser Aufzeichnungen verwendet (Siehe Kapitel 12.5).

Musskriterien		
Nummer	**Kriterium**	**Umsetzung**
1.1	Identifikation relevanter Daten für die Prognoseerstelltung.	X
1.2	Lokalisierung potentieller (heterogener) Datenquellen.	X
1.3	Die Prognosen basieren auf historischen Daten und die Prognose selbst wird ebenfalls für einen bereits vergangenen Zeitraum durchgeführt.	X
1.4	Vorverarbeitung der Daten für den Datenimport in SAP HANA.	X
1.5	Identifikation geeigneter Algorithmen für die Prognoseerstellung.	X
1.6	Transformation der Daten für die Anwendung von Algorithmen in SAP HANA.	X
1.7	Durchführung der Prognosen in SAP HANA.	X
1.8	Messung der Zuverlässigkeit/Genauigkeit der Prognose.	X
1.9	Visualisierung/Vergleich der Ergebnisse mit Realdaten.	X

Tabelle 114: Umsetzung der Musskriterien

Wunschkriterien		
Nummer	**Kriterium**	**Umsetzung**
2.1	Automatisierung des gesamten ETL-Prozesses.	0
2.2	Export von Auswertungen nach Excel.	X
2.3	Automatisierte Visualisierung der Ergebnisse.	X

Tabelle 115: Umsetzung der Wunschkriterien

Qualitätsanforderungen		
Nummer	**Kriterium**	**Umsetzung**
4.1	Die Prädiktionsgenauigkeit der Algorithmen wird anhand von festgelegten Fehlerkennzahlen evaluiert.	X
4.2	Der korrekte Datenimport der Rohdaten in das System muss sichergestellt werden.	X
4.3	Es muss sichergestellt werden, dass die Daten vor der Anwendung der Algorithmen bereinigt werden.	X

Tabelle 116: Umsetzung der Qualitätsanforderungen

13 SAP HANA Verbesserungen

Die genannten Fehler und Verbesserungsvorschläge beziehen sich auf SAP HANA in der Version SP08. Zuerst wird darauf eingegangen, dass in AFM fehlerhafte Modelle erzeugt werden. Dann werden Vorschläge für die Verbesserung der PAL-Dokumentation erbracht. Es ist wünschenswert, die Algorithmen in der Anwendung zu vereinfachen. Anschließend wird auf eine Limitierung der Resultsets von Queries eingegangen. Es erfolgt die Transparentmachung von Ungenauigkeiten in Fehlerdateien. Zuletzt wird aufgezeigt, dass komplexe Operationen zu Fehlerverhalten führen können.

13.1 Fehlerhafte Modelle AFM

Mit Hilfe des Application Function Modelers (AFM) können einfach und intuitiv Modelle durch einen WYSIWYG-Editor erzeugt werden, die in die Predictive Analysis Library (PAL) übersetzt werden. In der praktischen Anwendung zeigt es sich, dass die vom AFM automatisch generierten Quellcodes nicht immer valide sind und sich von den tatsächlichen Implementierungen der PAL unterscheiden. Das bedeutet, dass die in AFM erstellten Modelle nicht immer praktisch nutzbar sind. Die generierten PAL-Quelltexte sind jetzt manuell zu prüfen und zu korrigieren. Aufgrund dieser Tatsache wird der AFM weiterhin nicht mehr innerhalb der Projektgruppe genutzt und direkt die Funktionalität der PAL angewandt. Für zukünftige Releases von SAP HANA ist eine erhöhte Qualitätssicherung des AFM wünschenswert, damit besagte Vorteile nutzbar sind.

13.2 Dokumentation PAL vertiefen

Zu der Predictive Analysis Library (PAL) existiert ein Dokument, das die Anwendung der verschiedenen Algorithmen und deren theoretische Grundlagen beschreibt. Das Dokument ist in der Form gegliedert, dass erst die Parameter des jeweiligen Algorithmus definiert werden und anschließend ein Beispielaufruf gegeben ist. Die Einstiegshürde in die Thematik des maschinellen Lernens ist für Anfänger sehr hoch. Für zukünftige Versionen der PAL-Dokumentation ist es wünschenswert, mehr praktische Beispiele mit echten Daten anzugeben. Außerdem wird an einigen Stellen zu wenig auf den mathematischen Hintergrund der Algorithmen eingegangen. Ein Beispiel dazu ist die mathematische Beschreibung der neuronalen Netze. Vereinzelt ist es selbst durch weitergehende Literaturrecherche nicht möglich herauszufinden, welche Bedeutung einige Parameter der Algorithmen haben. Auch ist es manchmal nicht möglich zu erkennen, um welche Variante des jeweiligen Algorithmus es sich bei der Implementierung handelt. Eine weitergehende Literaturrecherche ist somit nicht möglich. Hier sind umfassende Informationen aus erster Hand wünschenswert.

13.3 Vereinfachung Anwendung Algorithmen

Einige Algorithmen benötigen eine synthetisch generierte, fortlaufende Identifikationsnummer. Dadurch wird die Stationarität der Daten sichergestellt. An dieser Stelle ist eine Vereinfachung wünschenswert, die die Nutzung der Algorithmen ohne diese Identifikationsnummer erlaubt.

13.4 Limitierung der Resultsets

Die Resultsets (Ergebnismengen) von SQL Queries sind in SAP HANA Studio standardmäßig auf 5000 Zeilen limitiert. Manchmal ist es jedoch zwecks Datenanalyse notwendig, auch mehr Datensätze auszugeben. Hier ist eine einfache und schnelle Möglichkeit sinnvoll, um die Resultset-Größe temporär zu erhöhen.

13.5 Fehlerdatei CSV-Import ungenau

Der SAP HANA CSV-Import mittels SFTP und Control Files bietet die Möglichkeit, schnell und einfach große Datenmengen in SAP HANA zu importieren. Im Falle eines Fehlers wird eine Fehlerdatei im Importverzeichnis erzeugt, die Auskunft über die Ursache gibt. Leider werden nicht die Zeilennummern der fehlgeschlagenen Importe angegeben. Dadurch ist die Fehlersuche umständlich. An dieser Stelle ist eine Angabe der fehlerhaften Zeilennummern bei fehlgeschlagenem Import wünschenswert.

13.6 Assistent Control Files

Für den CSV-Datenimport in SAP HANA müssen sogenannte Control Files erzeugt werden, die Metadaten zum Importvorgang bereitstellen. In den Control Files werden beispielsweise das Zielschema, die HANA Instanz und Formatangaben zu CSV (Delimiter, Quote, ...) angegeben. Für die einfache Erstellung dieser Control Files ist es wünschenswert, einen Assistenten in SAP HANA Studio zur Erstellung von Control Files zur Verfügung zu stellen.

13.7 Abbrechen komplexer Operationen

Werden im SAP HANA Studio komplexe Datenbankoperationen gestartet, so wird die Möglichkeit eines Abbruchs angegeben. Nach Initiierung des Abbruchs reagiert das SAP HANA Studio oftmals nicht. Die Folge darauf ist entweder, dass bis zum Ende der Operation gewartet werden muss, oder dass das SAP HANA Studio abstürzt. Hier ist eine funktionierende Abbruchfunktion wünschenswert. Gerade bei einer umfassenden Modellbildung - wie zum Beispiel bei der Support Vector Machine mit mehr als drei Eingabeprädikatoren - ist manchmal ein Abbruch bei versehentlicher Ausführung notwendig.

13.8 Einfrieren bei komplexen Operationen

Bei komplexen Datenbankoperationen, die über eine Stunde Ausführungszeit beanspruchen, friert das SAP HANA Studio regelmäßig ein. Es ist an der Stelle wünschenswert, mehr Stabilität bei Operationen mit langer Ausführungszeit bereitzustellen.

14 Fazit

In diesem Fazit wird auf einige Kernthemen eingegangen, eine Wertung vorgenommen und Empfehlungen für die Zukunft gegeben.

Rahmenbedingungen Der Start der Projektgruppe war am 01.04.2014. Nach einer Seminar- und Schulungsphase endete am 13.08.2014 unerwartet die Zusammenarbeit mit der eXin AG. Im Oktober 2014 begann die Projektgruppe mit der Findung einer Vision. Als neuer externer Betreuer wurde Dr. Joachim Kurzhöfer von der AS Inpro GmbH gewonnen, der sich im weiteren Verlauf als Richtungsweiser für den Erfolg des Projekts herausstellte. Am 19.10.2014 ist die Vision ermittelt und es begannen erste Recherche- und Implementierungsarbeiten. Bis zum 02.03.2015 erfolgte die Durchführung neuer Implementierungen, anschließend fokussierte sich die Projektgruppe auf den CeBIT-Auftritt vom 16.03.2015 bis zum 20.03.2015 und den Abschluss der Dokumentation zum 31.03.2015. Es resultiert daraus, dass die Kernarbeitszeit für die Erstellung dieses Dokuments und der damit verbundenen Artefakte etwa fünf Monate betrug.

Umsetzung der Anforderungen Wie in Kapitel 12.11 beschrieben, konnten alle zuvor definierten Musskriterien sowie Qualitätsanforderungen erfüllt werden. Lediglich bei den Wunschkriterien konnte die Automatisierung des gesamten ETL-Prozesses nicht umgesetzt werden, da die notwendige Hardwareinfrastruktur nicht zeitnah bereitgestellt wurde. An dieser Stelle wäre es wünschenswert, wenn die Bereitstellung von weiteren Hardwareressourcen für Projektgruppen flexibler erfolgen könnte.

Ergebnis der Gesamtevaluation Im Kapitel 12.7 ist die Gesamtevaluation im Detail nachzulesen. Es konnte ein Modell unter Verwendung des Algorithmus *lineare Regression mit gedämpftem Trend und saisonaler Anpassung* erzeugt werden, das eine relativ hohe Prognosegenauigkeit mit genannten Prämissen darbietet. Außerdem konnte die Haupthypothese zumindest teilweise bestätigt werden. Retrospektiv betrachtet ist die Erzeugung von passenden Modellen eine zeitaufwändige Prozedur, bei der viele Variablen berücksichtigt werden müssen. Bezüglich der Prognosegenauigkeit besteht auch weiterhin Verbesserungsbedarf, der möglicherweise im Rahmen einer weiteren Ausarbeitung abgedeckt werden kann. Es ist hervorzuheben, dass das Service Pack 09 der SAP HANA Instanz etwas spät eingespielt wurde, sodass keine Tests mit neuronalen Netzen durchgeführt werden konnten.

Die Arbeit mit SAP HANA Aus technischer Sicht war die Arbeit mit SAP HANA sehr interessant und lehrreich, da im Team bisher nur wenige Erfahrungen in Verbindung mit der In Memory Technologie vorlagen. Es wurde seitens der SAP ausreichend Lernmaterial

zur Verfügung gestellt, sodass sich schnell eine steile Lernkurve für das Team ergab. Als besondere Herausforderung stellten sich die bereits in Kapitel 13 genannten Fehler und Unzulänglichkeiten, die erfolgreich bewältigt wurden. Im Verlauf der Entwicklung wurden seitens des HPI regelmäßig neue Service Packs für SAP HANA eingespielt. Dies sorgte zwar dafür, dass neue Funktionalitäten zur Verfügung gestellt wurden – aber jedoch im Gegenzug auch dafür, dass bisher entwickelte Scripts nicht mehr valide waren. Die Scripts mussten entsprechend regelmäßig migriert werden.

In Memory und SAP HANA Eine in dieser Ausarbeitung nicht weiter betrachtete Fragestellung ist, in wieweit die In Memory Technologie und SAP HANA diese Aufgabenstellung unterstützt haben. Diese wurde nicht betrachtet, da die Nutzung von SAP HANA eine gegebene Anforderung ist (siehe auch Kapitel 6). An dieser Stelle ist jedoch zu erwähnen, dass ein Vergleich zwischen SAP HANA und alternativen Open Source Lösungen wie zum Beispiel Apache™ Hadoop®! [Fou15] für die Planung und Prognose ebenfalls eine sehr interessante weitergehende Fragestellung wäre.

Hypothesenbildung Eine besondere Herausforderung war die Hypothesenbildung (siehe auch Kapitel 3). Es stellte sich heraus, dass Hypothesen einerseits Top-Down und andererseits auch Bottom-Up gebildet werden können. Top-Down bedeutet, dass erst die Hypothesen gebildet werden und anschließend deren Machbarkeit geprüft wird. Bottom-Up bedeutet, dass anhand der Machbarkeit und Datenverfügbarkeit die Hypothesen gebildet und gewählt werden. In der Praxis war es ein iterativ-inkrementeller Prozess der Diskussion und Recherche, bei dem beide Verfahren zum Einsatz kamen, um machbare und vielversprechende Hypothesen zu entwickeln. Diese Vorgehensweise basiert auf der Erkenntnis, dass keine Hypothese Top-Down festgelegt werden kann, ohne dass fundierte Daten verfügbar sind. An dieser Stelle wird auch die Wichtigkeit der Verfügbarkeit von qualitativ hochwertigen und zweckgebundenen Daten hervorgehoben.

Algorithmen und Mathematik Die hohe Anzahl der möglichen Algorithmen sowie deren umfassende mathematische Hintergründe stellten die größte fachliche Herausforderung dar. Die theoretischen Grundlagen sowie deren Anwendung in SAP HANA wurden weitestgehend selbstständig erarbeitet. Trotz des großen Lerneffekts wäre es für künftige Projektgruppen in dem Themengebiet möglicherweise noch besser, ein interdisziplinäres Team bestehend aus Informatikern und Mathematikern aufzustellen. Es wäre an der Stelle auch eine Zusammenarbeit mit einer anderen Fakultät (beispielsweise der Fakultät V der Universität Oldenburg) denkbar.

Kommunikation Das Projektteam kam zu der Erkenntnis, dass bei verteilter Zusammenarbeit (örtlich und zeitlich) der Kommunikationsaufwand im Vergleich zur direkten

Kommunikation zunimmt. Um die Kommunikation im Projekt zu verbessern, wurden folgende Maßnahmen getroffen:

- Einführung einer wöchentlichen, moderierten Sitzung mit externen Betreuern.
- Umfangreiche Protokollierung der Sitzungen und der wichtigen Entscheidungen.
- Festlegung von Kernarbeitstagen und -zeiten.
- Einrichtung eines externen und internen E-Mail-Verteilers sowie eines Instant Messenger und eines Systems zur Telekonferenz.
- Durchführung von Teambildungsmaßnahmen/Events.
- Festlegung eines Kommunikationsbeauftragten.

Bei der wöchentlichen Zusammenarbeit in kleineren Teams hat sich als besonders kommunikationsfördernd erwiesen, die Teamzusammensetzung regelmäßig zu variieren. So erhält jedes Teammitglied die Chance mit jedem zusammenzuarbeiten. Dies ist auch als Wissensmanagementmaßnahme zu verstehen.

Scrum Als Vorgehensmodell für die Projektdurchführung wurde Scrum gewählt (siehe auch Kapitel 4.2). Die Sprint-Länge wurde auf eine Woche festgelegt. Zu Anfang erfolgte eine Einteilung in zwei Teams (PA und ETL), um den Koordinationsaufwand innerhalb der Teams gering zu halten. Schnell hat sich herausgestellt, dass jedoch die Koordination der Teams untereinander nicht funktionierte. Die Empfehlung für zukünftige Projektvorhaben lautet, möglichst keine Einteilung in feste Teams vorzunehmen und außerdem Maßnahmen einzuleiten, um die teamübergreifende Koordination zu verbessern (beispielsweise Scrum of scrums).

Die Abgabequalität von einigen Aufgaben war nach den ersten Sprints nicht wie erwartet. Deshalb wurde schnell eine Definition of Done (DoD) vereinbart, die besagt, dass Tasks erst als abgeschlossen markiert werden dürfen, sobald sie implementiert und dokumentiert wurden (wie in Kapitel 4.2 beschrieben). Damit konnten die Qualitätsprobleme eingedämmt werden.

Interkulturelle Zusammenarbeit Eine besonders spannende Komponente in der Teamzusammensetzung ist die interkulturelle. Im Rahmen dieser Projektgruppe wurde die Chance gegeben mit Menschen mit verschiedenen kulturellen Hintergründen zusammenzuarbeiten. Dadurch wurden Kompetenzen aus verschiedensten Fachbereichen mit unterschiedlichen Sprachkenntnissen mit eingebracht. Einerseits stellt diese Zusammensetzung eine Herausforderung dar, die sich beispielsweise in Kommunikationsschwierigkeiten äußern kann – im Fall dieser Projektgruppe kann retrospektiv gesagt werden, dass diese Herausforderung bewältigt wurde. Zudem hat die Zusammenarbeit Spaß bereitet und jeder konnte sich einbringen.

Exkursionen Zwei Exkursionen haben im Rahmen der Projektgruppe OliMP stattgefunden. Dies waren der „New Business Generation Day“ in Ettlingen bei Karlsruhe und der „Future SOC Lab Day“ am Hasso Plattner Institut (HPI) in Potsdam. An der ersten Exkursion haben vier Projektmitglieder teilgenommen. Neben interessanten Vorträgen von verschiedenen Unternehmen haben die Projektmitglieder an einem Workshop zum Thema „Design Thinking“ teilgenommen. Das praktische Umsetzen des Konzepts haben die Teilnehmer durch einen Wettbewerb kennengelernt. Aufgrund fehlender Erfahrung wurde das Konzept im Rahmen der Projektgruppe nicht umgesetzt, aber trotzdem war die Erfahrung sehr positiv, da die Teilnehmer die Chance hatten, die neuesten Ideen im Bereich „Design Thinking“ kennenzulernen und mit anderen Studenten und Fachleuten ins Gespräch zu kommen.

An der zweiten Exkursion haben sieben Projektmitglieder teilgenommen. Beim HPI fand am Dienstag, den 28. Oktober 2014, das HPI Cloud Symposium und am nächsten Tag der Future SOC - Lab Day statt. Verschiedene interessante Themen von IT-Sicherheit bis zur Performance von SAP HANA wurden vorgestellt. Ein interessantes Projekt mit einem der Projektgruppe OliMP ähnlichen Thema wurde von Wissenschaftlern der Universität Posen vorgestellt. Die Projektmitglieder sind mit den Wissenschaftlern der Universität Posen ins Gespräch gekommen. Dabei wurden wichtige Herausforderungen und Empfehlungen besprochen, die unserer Projektgruppe nützliche Erkenntnisse gebracht haben.

Beide Exkursionen hatten einen positiven Einfluss auf den Teamgeist, die fachlichen Kenntnisse und die Präsentationserfahrung der Teilnehmer. Dies war bei der Weiterarbeit am Projekt sehr hilfreich.

15 Veranstaltungen

In diesem Kapitel erfolgt die Beschreibung der Veranstaltungen, an denen die Projektgruppe im Verlauf der Projektdurchführung teilnahm. Zuerst werden die Erfahrungen innerhalb eines Design Thinking Workshops beschrieben. Anschließend erfolgt die Beschreibung eines Boule Turniers. Es wurde eine Schulung in SAP PA und SAP BPC mit der eXin AG durchgeführt. Zuletzt werden das HPI Cloud Symposium und der Future SOC Lab Day beschrieben.

15.1 Design Thinking Workshop

Am 10. Juli 2014 haben wir zu viert am Fachkongress „New Business Generation“ in Ettlingen (bei Karlsruhe) teilgenommen. Organisator des Events war die anthesis GmbH, ein IT Entwicklungs- und Beratungshaus.

Unter dem Motto „Unternehmen Zukunft“ wurden mehrere Vorträge über die aktuellen Entwicklungen und Trends im Bereich Business gehalten. Im Rahmen einer Firmenausstellung präsentierten sich verschiedene Unternehmen aus der Region und führten ihre Lösungsansätze und Anwendungsszenarien vor.

Oliver Kempkens, Co-Founder & Chairman des Unternehmens ADAPT OR DIE CONSULTING, hielt einen Vortrag zum Thema „DESIGN THINKING – Die Methode für erfolgreiche Innovationen“ȦDAPT OR DIE ist eine international tätige Innovationsberatung mit Büros in Heidelberg, Essen und London. Das Beratungsunternehmen ist ein Premium-Partner der SAP AG für die Implementierung des Innovationsprozesses Design Thinking. Diese Methode bringt Menschen unterschiedlicher Disziplinen in einem kreativen Arbeitsumfeld zusammen, um systematisch nutzerzentrierte Innovationen zu entwickeln.

Die Methode des Design Thinkings wurde im Rahmen der Seminarphase unserer Projektgruppe als möglicherweise sinnvolles Verfahren zur Erarbeitung einer Produktidee beziehungsweise einer Vision diskutiert. Nur aufgrund der theoretischen Kenntnisse und ohne praktische Erfahrung fiel uns eine Beurteilung der Methode jedoch schwer. Der im Rahmen des Kongresses veranstaltete Design Thinking Contest stellte daher für uns eine gute Gelegenheit dar, das Verfahren praktisch auszuprobieren.

Der als Workshop gestaltete Contest wurde geleitet von Oliver Kempkens und Sascha Wolf, Co-Founder & Managing Director von ADAPT OR DIE. Der Workshop befasste sich unter dem Schlagwort „Informationsflut“ mit der Problemstellung, wie die individuellen Informationsbedürfnisse von Menschen besser befriedigt werden können. Zusammen mit den anderen Teilnehmern des Workshops haben wir, aufgeteilt in zwei Teams à sechs Personen, die verschiedenen Phasen des 6-stufigen Design Thinking Prozesses durchlaufen. Unter Anleitung der Coaches haben wir so konkrete Lösungsideen entwickelt und anschließend prototypisch als Mock-ups auf Flipcharts umgesetzt. Nachfolgend wurden

die Ergebnisse den Teilnehmern des Fachkongresses präsentiert. Die Zuhörer haben anschließend eine der beiden Gruppen zum Gewinner des Contests gewählt.

Durch die Teilnahme am Workshop konnten wir die Methode des Design Thinkings auf praktische und realitätsnahe Weise kennenlernen und anwenden. Die im Seminar aufgestellte These, dass mit Hilfe von Design Thinking verhältnismäßig leicht und schnell innovative Ideen erzeugt werden können, hat sich für uns im Verlauf des Workshops bestätigt. Dabei waren die einzelnen Prozessphasen sogar anschaulicher und die Ergebnisse greifbarer, als wir es uns vorgestellt hatten. Wie von den Prinzipien des Design Thinkings vorausgesetzt, stammten die Teilnehmer des Workshops aus unterschiedlichen Fachbereichen. Das sich daraus ergebende Potential zur Innovationsfindung war trotz der kurzen Dauer des Designprozesses zumindest im Ansatz zu erkennen. Die durch die Teilnahme am Workshop gesammelten Erfahrungen könnten uns nun dabei helfen, den Prozess des Design Thinkings in der Projektgruppe zu implementieren.

Im Anschluss an den Contest haben wir gemeinsam einen Vortrag zum Thema „Feel Good Management“ gehört, der von Anka Hansen, einer Expertin für gehirngerechtes Denken & Handeln, gehalten wurde.

Als Tagesausklang fand ein GetTogether zum gegenseitigen Austausch statt. Dies bot uns die Gelegenheit, mit den Fachleuten vor Ort und anderen Studierenden ins Gespräch zu kommen.

15.2 Boule Turnier

Am 23. Juli fand das diesjährige Boule-Turnier der Projektgruppen des Department für Informatik statt. Auf der Boule-Anlage des Oldenburger Turnerbund (OTB) am Osterkampsweg traten insgesamt 12 Teams gegeneinander an. OliMP hat mit einem 5er-Team teilgenommen und einen hervorragenden 3. Platz belegt.

15.3 Schulung in SAP PA und SAP BPC mit eXin AG

Am 18. und 19. September wurde die Projektgruppe von der eXin AG im Umgang mit den Planungs- und Prognosewerkzeugen „SAP Predictive Analysis“ und „SAP BPC“ geschult. Die Projektgruppenmitglieder lernten die Softwarelösungen anhand von praktischen Übungsaufgaben kennen und besitzen nun das nötige Grundwissen, um eines der beiden Werkzeuge im Rahmen des Projekts einzusetzen. Für die erworbenen Kenntnisse erhielten die Projektgruppenmitglieder ein Zertifikat, in dem die Inhalte der Schulung beschrieben wurden:

SAP PA 1.17

- Architektur und Einsatzmöglichkeiten von SAP Predictive Analysis
- Erstellen von Datensätzen auf Flat-File Datenprovider
- Vorbereitung von Daten für Data-Mining, Reporting und Analyse
- Verwendung von Data-Mining Modellen für prädiktive Analysen
- Erstellen und Arbeiten mit den Berichtsbestandteilen Tabellen und Diagrammen
- Filtern von Daten im Datenprovider und im Bericht
- Exportieren von gefilterten Datensätzen

SAP BO BPC 10.0

- Architektur und Einsatzmöglichkeiten von SAP Business Planning und Consolidation 10.0
- Konzeption einer Planung-Lösung
- Erstellen von Environment, Dimensionen und Modellen mit dem Administration Client
- Erstellen/ Bearbeiten von Strukturen und Stammdaten
- Sicherheitskonzept umsetzen, Erstellen von Berechtigungen für Datenzugriff und Funktionen
- Erstellen von Eingabemasken mit BPC EPM Excel
- Arbeiten mit BPC EPM Excel Plan-Funktionen
- Verwendung von Daten-Paketen, um Daten zu kopieren und löschen
- Erstellung von zentralen Filter-Objekten

15.4 HPI Cloud Symposium und Future SOC Lab Day

Am Dienstag, den 28. Oktober 2014, haben wir als Gruppe von sieben Studenten aus der PG OliMP uns um vier Uhr morgens vor der Uni getroffen, um nach Potsdam zum HPI zu fahren. Dort fand am Dienstag das HPI Cloud Symposium und am nächsten Tag der Future SOC – Lab Day statt.

Am ersten Veranstaltungtag konnten wir um 9:30 Uhr die für uns bereitliegenden Namensschildchen bei der Anmeldung abholen. Kurz darauf begannen die verschiedenen Vorträge rund um das Thema Operating the Cloud. Einführend wurde das Hasso-Plattner-Institut von Prof. Dr. Christoph Meinel vorgestellt. Als erster geladener Redner sprach Thorsten Höhnke von Fujitsu Technology Solutions über unterschiedliche Sicherheitsherausforderungen, denen die heutige IT-Infrastruktur gegenübersteht. Dabei verwies er auf die Lösung

diverser bekannter Angriffe auf IT-Systeme durch eine baldige – jedoch nicht näher spezifizierte – Produkteinführung von Fujitsu. Anschließend stellte Peter Kirchner, ein externer Doktorand beim HPI, der bei Microsoft arbeitet, die Möglichkeiten von Microsoft Azure vor. Sascha Bosse von der Universität Magdeburg hielt darauffolgend einen Vortrag über die Möglichkeiten von Unternehmen bei der Wahl ihres IT Service Designs und wie diese verglichen werden können. Hierbei wurden insbesondere die Parameter Verfügbarkeit, Antwortzeiten und Kosten betrachtet. Resümierend wurde betont, dass die Unternehmen sich entsprechend ihrer Schwerpunkte für eine Alternative entscheiden sollten.

Nach dem Mittagessen hielt Hendrik Müller, ebenfalls von der Universität Magdeburg, einen Vortrag darüber, wie große Daten mit einer besseren Performance in SAP HANA importiert werden können. Er kam zu dem Ergebnis, dass die Standard-Einstellungen in HANA keine optimale Import-Performance erzielen. So arbeite die Datenbank mit den von SAP empfohlenen Einstellungen etwa fünf mal performanter. Noch bessere Ergebnisse ließen sich nur erzielen, wenn die Einstellungen manuell auf die zu importierenden Daten abgestimmt werden. Es folgte ein Vortrag von Tomasz Szepieniec, der den FitSM Standard erläuterte. Hierbei handelt es sich um eine vereinfachte Version von ITIL die den Anspruch hat in der Praxis besser einsetzbar zu sein. Danach stellte Mohamed Elsaid das Thema seiner Doktorarbeit vor: Er untersucht wie eine virtuelle Maschine, die unter Vollast arbeitet, im laufenden Betrieb zu einem anderen Wirtssystem migriert, sodass dieser Prozess in Zukunft eventuell noch effizienter ablaufen kann. Nach der Kaffeepause sprach Christian Neuhaus über den Vergleich unterschiedlicher Cloud-Lösungen miteinander. Dazu erläuterte er die verschiedenen Attribute, die hierfür von Bedeutung seien. Eine große Herausforderung stelle hierbei insbesondere die schwierige Messbarkeit unterschiedlicher Eigenschaften dar. Anschließend thematisierte Prof. Dr. Andreas Thor die Verwaltung von Hot Spot Data Objekten in NoSQL-Datenbanken. Den letzten Vortrag des Tages hielt Christian Frank von HP zu den Einsatzmöglichkeiten von OpenStack beziehungsweise HP Helios OpenStack für HP selbst, als auch für dessen Kunden. Schließlich wurde das HPI Cloud Symposium durch Prof. Dr. Meinel beendet.

Um halb Zehn am nächsten Morgen begann der Future SOC – Lab Day mit der Anmeldung. Die anschließende Begrüßung wurde außerplanmäßig von Bernhard Rabe in Vertretung für Prof. Dr. Andreas Polze übernommen. Der erste Vortrag an diesem Tag wurde von Bernd Winkelsträter über die Zukunft von nichtflüchtigen Datenspeichern in Rechenzentren gehalten. Anschließend wurde von Wissenschaftlern der Universität Posen ein Projekt vorgestellt, das die Ressourcen des HPI zur Datenanalyse nutzt, um rationale Entscheidungen im Energiesektor besser unterstützen zu können. Dafür wurden Daten von Windfarmen, Solaranlagen, Meinungsplatformen zu Stromanbietern und Smartmetern verarbeitet und anschließend graphisch dargestellt. Eine wichtige Erkenntnis des Projektes war, dass insbesondere die schnelle Verarbeitung der Daten zu einer höheren Zufriedenheit bei den Energieendekunden führen kann. Es folgte ein Vortrag von David Schwalb zu

seiner Doktorarbeit, die sich damit beschäftigt wie In-Memory Datenbanken einen Crash unbeschadet überstehen können.

In der folgenden Mittagspause konnten wir einigen Interessierten unser Forschungsposter präsentierten. Beim Mittagessen kam es auch zu einem anregenden Gespräch mit Prof. Witold Abramowicz von der Universität Posen bezüglich der Entwicklung unseres Projekts.

Nach dem Mittag begann ein Vortrag zu den unterschiedlichen Facetten der Business Process Modellierung. Es wurden drei konkrete Szenarien beschrieben, die die Frage klären sollten, wie eine High Performance IT-Infrastruktur genutzt werden kann, um damit zusammenhängende Probleme zu lösen. Darauf folgte die Vorstellung der Doktorarbeit von Arvid Heise, die die Verarbeitung und Auswertung von Open Government-Daten mit Stratosphere zum Thema hat. Dabei wurden insbesondere finanzielle Investitionen von verschiedenen Regierungen analysiert. So soll es langfristig möglich sein Korruption mit diesen Daten automatisch aufzudecken, da scheinbar zufällige Zusammenarbeiten verschiedener Personen entdeckt werden können.

Nach der Kaffeepause folgte ein Vortrag über ein Projekt, das sich mit Data Mining mit SAP HANA beschäftigt. Dabei wurden in erster Linie die verschiedenen Performance- und Genauigkeitseigenschaften unterschiedlicher Algorithmen auf R, PAL und C++ im Zusammenhang mit SAP HANA betrachtet. Der letzte Vortrag der Tagung betrachtete potentielle Vorteile von In-Memory Technologien im operativen Tagesgeschäft mithilfe von EUS in Echtzeit.

Zum Schluss verabschiedete Prof. Dr. Christoph Meinel alle Anwesenden und wünschte eine gute Heimreise. Diese traten wir dann auch sogleich an, um am Abend erschöpft aber um einige Erfahrungen reicher in Oldenburg anzukommen.

16 CeBIT

Dieses Kapitel handelt von dem CeBIT Auftritt der Projektgruppe. Im ersten Abschnitt wird zunächst das Kick-Off-Ausstellertreffen vom 26.11.2014 beschrieben. Im nächsten Unterkapitel befindet sich ein kurzer Bericht, der nach der CeBIT angefertigt wurde.

16.1 Kick-Off Ausstellertreffen am 26.11.2014

Am 26. November 2014 fand um zehn Uhr das Kick off Ausstellertreffen der CeBIT 2015 im Haus der Wirtschaftsförderung in Hannover statt. Eingeladen hatte die mit der Organisation des Standes des Landes Niedersachsen beauftragte Agentur „KRISPIN Marketing Management“. Zentrale Punkte des Treffens waren die Präsentation des Standkonzeptes, eine kurze Vorstellung aller 18 Aussteller und die wichtigsten Daten zur Vorbereitung auf die CeBIT.

Nach einer kurzen Begrüßung durch Frau Ulrike Rom von der KRISPIN-Agentur, stellte Frau Diana Schreiber von der Deutsche Messe AG das Konzept der CeBIT 2015 vor: Demnach lautet das „Top-Thema“ der CeBIT 2015 „d!conomy“ und soll die völlig neuen und veränderten Geschäftsmodelle beleuchten, die durch IT und Digitalisierung entstehen. Hierbei wurde auch der Geschäftskunden-Fokus der Messe betont. Es folgten einige Daten, die das Wachstum und die Relevanz der CeBIT verdeutlichten.

Anschließend wurden einige Personen kurz vorgestellt, die für den reibungslosen Ablauf seitens des Landes Niedersachsens mitverantwortlich sind. Erste Ansprechpartnerin für Fragen organisatorischer Natur ist für alle Aussteller auf dem Niedersachsen-Stand jedoch Frau Ulrike Rom.

Annelies Bruhne stellte das Enterprise Europe Network (EEN) vor, eine internationale Kooperationsbörse für Unternehmen und Forschungseinrichtungen aus dem IKT-Bereich. Das EEN wird etwa ein viertel der Fläche des Messestandes einnehmen und hat seit 1997 das Ziel, internationale Kooperationen zu fördern. Für alle Aussteller des Landes Niedersachsen ist es kostenlos möglich Kooperationsprofile unter `www.futurematch.cebit.de` mit Kontaktdaten, Beschreibung der Einrichtung sowie eines Kooperationsangebots bzw. -gesuchs zu erstellen. Dabei können alle Profile der teilnehmenden Institutionen eingesehen und Termine vereinbart werden. Der Code für die kostenlose Teilnahme kann den Folien des Ausstellertreffens entnommen werden.

Anschließend wurde das Raumkonzept des Standes des Landes vorgestellt. Für die Planung von Interesse sind insbesondere die Arbeitsplätze, die in zwei unterschiedlichen Varianten existieren:

Standard Arbeitsplatz PC- Counter; elliptische Arbeitsplatte; Maße: Länge 1,35 m x Breite 0,73 m x Höhe 1,15 m; Dreifachsteckdose; Unterschrank, abschließbar für Rech-

nereinheiten. Hinweis: Technisches Equipment wie Laptop und Bildschirme bringt jeder Aussteller selber mit.

Präsentationsplatz mit Monitorhaltung Standard Arbeitsplatz mit Befestigungsmöglichkeit für Großbildschirm mittels Adapter[14]

An den Arbeitspätzen ist ein Anschluss an das Wissenschaftsnetz des DFN Vereins über eine leistungsfähige Standleitung vorhanden. Es besteht darüber hinaus die Möglichkeit einen eigenen WLAN AccessPoint mitzubringen, der aber gegen eine Gebühr von 50 Euro zzgl. MwSt. angemeldet werden muss. Die Messe AG prüft genau, ob dies geschehen ist.

Es folgte eine kurze mündliche Vorstellung aller Aussteller in folgender Reihenfolge:

- Hochschule Hannover / Fak. II, IVEK
- Leibniz Universität Hannover / L3S
- Leibniz Universität Hannover / Institut für Mensch-Maschine-Kommunikation
- Leuphana Universität Lüneburg / Gründungsservice
- Carl von Ossietzky Universität Oldenburg / Department für Informatik
- Lowo Tec (Ausgründung der Carl von Ossietzky Universität Oldenburg / Department für Informatik)
- Universität Osnabrück / Institut für Kognitionswissenschaft
- Technische Universität Clausthal / Institut für Informatik
- Technische Universität Clausthal / Institut für Maschinenwesen
- c4c Engineering GmbH Braunschweig
- HAWK Hildesheim/ Holzminden/ Göttingen / Fakultät Gestaltung
- Hochschule Emden/Leer / I2AR
- Hochschule Emden/Leer / FB Technik, E&I
- Jade Hochschule Oldenburg / Inst. f. Hörtechnik und Audiologie
- Jade Hochschule Wilhelmshaven / Institut für Wirtschaftsinformatik
- Ostfalia Hochschule für angewandte Wissenschaften
- OFFIS – Institut für Informatik
- ikn2020 - Das digitale Niedersachsen

Der letzte Tagesordnungspunkt betrachtete die Eckpunkte des Projekt- und Zeitplans für die CeBIT-Vorbereitung. Es wurden in diesem Zusammenhang auch die Leistungen erläutert, die die Aussteller durch das MWK bzw. die Messe AG erhalten:

[14] Passend für alle Monitore mit VESA 100 x 200 bzw. 200 x 200 mm Halterung. Eigene mitgebrachte Bildschirme von 22 Zoll bis max. 42 Zoll können befestigt werden; ausreichend langes Monitorkabel (5,0m DVI bzw. HDMI) mitbringen

- Fünf Karten für die Eröffnungsfeier der CeBIT
- 300 kostenfreie registrierungspflichtige Fachbesucher-Tickets (print oder elektronisch)
- Zwei kostenfreie Ausstellerausweise (berechtigen die Hallen im Zeitfenster von 7:00 und 19:00 Uhr zu betreten, übertragbar), jeder weitere Ausstellerausweis kostet 64 Euro
- Auftritt im Produktgruppenkatalog
- Aufnahme in das gedruckte Ausstellerverzeichnis
- Werbemittel (Poster, Flyer, Einladungskarten) können kostenfrei bestellt werden.
- Presseservice
- Standbroschüre

Der Messestand wird am Freitag, 13. März 2015 aufgebaut.

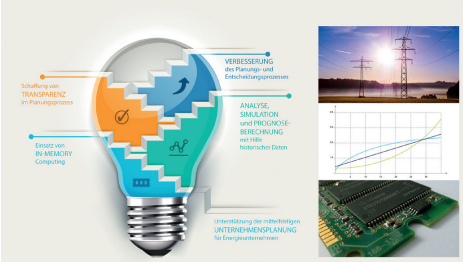
16.2 Bericht des Messeauftritts aus Sicht des OliMP Projekts

Am 16. März startete die CeBIT 2015. Der Messestand der Universität Oldenburg mit dem Projekt OliMP war dabei in den Stand des Landes Niedersachsens integriert. In diesem Rahmen haben wir auch am Future Match Programm des Enterprise Europe Network teilgenommen. Das Future Match Programm unterstützt kleine und mittlere Unternehmen dabei ihre Geschäftschancen zu verbessern und vernetzt dabei die etwa 600 angehörigen Organisationen aus Industrie, Technologiezentren, Unternehmen und Universitäten [Com15]. So konnten neue Partnerschaften auf europäischer Ebene geknüpft werden. Es konnten viele Erfolge während der Messezeit verbucht werden. Der Projektstand konnte viele Besucher vom Nutzen der IT-gestützen Energieverbrauchsvorhersagen überzeugen. Insbesondere die Möglichkeit der automatische Datenanalyse lockte das Interesse und die Begeisterung des Messepublikums. So wurden internationale wie auch nationale Kontakte geknüpft und potentielle zukünftige Kooperationen eingeleitet. Interessant waren auch die Ideen, die sich aus den Impulsen der vielfältigen Gespräche am Messestand ergeben haben. Auch der Austausch mit den anderen Ausstellern auf dem Niedersachsenstand und darüber hinaus hat uns sehr gefallen. Insbesondere die Besucher aus wissenschaftlichem Kontext konnten die Ideen des Projekts und unsere Passion widerspiegeln. Die diversen Hintergründe der Standbesucher konnten mit umfangreichen Feedback dem Projekt zusätzlichen Schwung geben. Die entstandenen Synergieeffekte werden den Messeerfolg der Universität auch nachhaltig spürbar machen.

CARL VON OSSIETZKY universität OLDENBURG

Olimp: Energieverbrauchsdaten zuverlässig vorhersagen

Olimp: Reliable Redictions of Energy Consumption



Die Energie von morgen schon heute planen.
Plan future energy-demand today.

Das Projekt Olimp (Oldenburg In-Memory-Planung mit SAP HANA) beschäftigt sich mit dem Einsatz von In-Memory-Planungs- und Prognosewerkzeugen. Das langfristige Ziel ist, Unternehmen bei Planungsprozessen durch vorhersagbare Datenanalysen zu unterstützen. Es werden unterschiedliche Algorithmen eingesetzt und untersucht, die verschiedene Eingangsparameter betrachten und heterogene Zukunftsvorhersagen ermöglichen.

Ein spannendes Einsatzgebiet ist dabei der Energiesektor: In-Memory-Technologien könnten Verbrauchsdaten besser vorhersagen als konventionelle Methoden. Insbesondere die Echtzeitfähigkeit durch effizient eingesetzte Prognosemethoden machen die Betrachtung vielseitiger Datenquellen wirtschaftlich attraktiv und damit erst praktisch möglich. Auf wissenschaftlicher Basis ist es eine Herausforderung zu klären, wie viele relevante Datenquellen und Parameter die Prognosegenauigkeit signifikant verbessern. Eine vielfältige Datenbasis wird zum Beispiel vom deutschen Wetterdienst zur Verfügung gestellt. Praktisch wird die Datenbasis des Projekts in einer In-Memory-Datenbank abgebildet. Das erlaubt es, große Datenmengen um ein vielfaches schneller als mit herkömmlichen Datenbanken zu verarbeiten. Die Anwendbarkeit wird laufend in einer SAP-HANA-Instanz verifiziert und damit fortlaufend auf ihre Praxistauglichkeit hin überprüft.

The research project "In-Memory Planning with SAP HANA" is focusing on operational planning and optimization. It is our agenda to improve the process of business planning with predictive data analysis. This can be done by using In-Memory-Technology. This new technology is able to create much faster prediction when dealing with many data sources which is especially interesting for enterprises. In our special case of research, we deal with the estimation and calculation of energy consumption. Although this can be done by just using historical data, a more accurate result can be achieved by including further data. Concerning the scientific side of the project, it needs to be clarified how the predictions change with involvement of different combinations of data. This is done by using the SAP HANA-In-Memory-Database with SAP AFM.

Fakultät II
Department für Informatik
Abt. Wirtschaftsinformatik / VLBA
Ansprechpartner: Prof. Dr. Jorge Marx Gómez
info@ol-imp.de
http://ol-imp.de

VERY LARGE BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

OLIMP

10

Abbildung 148: Projektseite des OliMP Projekts in der CeBIT Broschüre des Landes Niedersachsen

Literatur

[Ban03] Banschbach, Volker: *Einflussgrößen des Energieverbrauchs : Eine empirische Analyse für Deutschland ; Diplomarbeit*, Universität Heidelberg, Diplomarbeit, 2003

[Ber08] Bergin, Joseph: *Coding at the Lowest Level Coding Patterns for Java Beginners*. http://csis.pace.edu/~bergin/patterns/codingpatterns.html#avo. Version: 2008

[BFG04] Briegel, Ramon ; Filzek, Dirk ; GmbH), Peter Ritter (CUBE E.: *Stromlastprognose für RegModHarz*. http://www.regmodharz.de/fileadmin/user_upload/bilder/Service/Arbeitspakete/AP-Bericht-Stromlastprognose_AP2.4_CUBE.pdf. Version: 2004

[BM05] Brans, Jean-Pierre ; Mareschal, Bertrand: Promethee Methods. In: *Multiple Criteria Decision Analysis: State of the Art Surveys* Bd. 78. Springer New York, 2005. – ISBN 978–0–387–23067–2, S. 163–186

[Bus14] BusinessObjects, S. A.: *Leitfaden für SAP Predictive Analysis*. http://help.sap.com/businessobject/product_guides/SAPpa10/de/pa1_0_8_user_de.pdf. Version: 2014

[BV85] Brans, J. P. ; Vincke, Ph.: *A Preference Ranking Organization Method The PROMOTHEE method for Multiple Criteria Decision-Making Pages 647–656*. http://www.lamsade.dauphine.fr/~mousseau/pmwiki-2.1.5/uploads/Research/Brans1985.pdf. Version: 01.06.1985

[CD14] Chai, T. ; Draxler, R. R.: *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance*. http://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.pdf/. Version: 2014

[Com15] Commission, European: *Enterprise Europe Network*. http://een.ec.europa.eu. Version: 2015

[Far06] Farkisch, Kiumars: *Data-Warehouse-Systeme kompakt*. Springer Heidelberg, 2006. – 1–127 S. – ISBN 978–3–642–21532–2

[Fou15] Foundation, The Apache S.: *Welcome to Apache^{TM} Hadoop®!* https://hadoop.apache.org/. Version: 2015

[GHJV15] Gamma, Erich ; Helm, Richard ; Johnson, Ralph ; Vlissides, John: *Design Patterns - Entwurfsmuster als Elemente wiederverwendbarer objektorientierter Software*. 2014. Aufl. MITP Verlags GmbH, 2015. – ISBN 978–3–826–69700–5

[Glo13] Gloger, Boris: *Scrum - Produkte zuverlässig und schnell entwickeln*. Wien : Hanser, 2013

[GM05] Grace-Martin, Karen: *Assessing the Fit of Regression Models*. http://www.cscu.cornell.edu/news/statnews/stnews68.pdf. Version: 2005

[Gra13] Graf, Benjamin: *SAP Predictive Analysis — CIO Online*. http://www.sap-cio.de/neue-beitrage/losungen/analytics/sap-predictive-analysis/, 04 2013. – (Zuletzt aufgerufen am 11.03.2015)

[Hai00] Haider, Günter: *DESKRIPTIVE STATISTIK ANALYSE UND DARSTELLUNG VON DATEN I*. https://www.sbg.ac.at/erz/people/paschon/Internet_Kassel/MODUL%2012%20Korrelation.pdf. Version: 2000

[HF14] Hyndman, Rob J. ; Fan, Shu: *Monash Electricity Forecasting Model*. Monash University, 2014

[JWHT13] James, Gareth ; Witten, Daniela ; Hastie, Trevor ; Tibshirani, Robert: *An Introduction to Statistical Learning - with Applications in R*. 1. Aufl. Berlin Heidelberg : Springer Science and Business Media, 2013. – ISBN 978–1–461–47138–7

[Lei13] Leicht, Caroline: *Analyse und Optimierung von Algorithmen des Maschinellen Lernens in der Virtuellen Messtechnik*, Universität Leipzig, Diss., 2013

[Mic99] Microsystems, Sun: *Code Conventions for the Java Programming Language*. http://www.oracle.com/technetwork/java/codeconvtoc-136057.html. Version: 01.01.1999

[NRW15] NRW, Energieagentur: *Wissenswertes zum Thema Heizkosten*. http://www.energieagentur.nrw.de/kraftwerkstechnik/wissenswertes-zum-thema-heizkosten-12148.asp. Version: 2015, Abruf: 16.11.2014

[RH06] Reussner, Ralf ; Hasselbring, Wilhelm: *Handbuch der Software-Architektur*. dpunkt.verlag, 2006. – 1–557 S. – ISBN 978–3898643726

[SAP14a] SAP: *SAP HANA Developer Guide*. http://help.sap.com/hana/SAP_HANA_Developer_Guide_en.pdf. Version: 2014

[SAP14b] SAP: *Using Predictive Analysis Library*. https://help.hana.ondemand.com/help/frameset.htm?793823233aab4420aeff88ab9118d59c.html. Version: 2014, Abruf: 28.11.2014

[SAP15] SAP: *SAP HANA Predictive Analysis Library (PAL)*. http://help.sap.com/hana/SAP_HANA_Predictive_Analysis_Library_PAL_en.pdf. Version: 2015

[Sch07] Schweitzer, Udo: *Statistik mit Microsoft Excel -*. 1. Aufl. Witten : W3l GmbH, 2007. – ISBN 978–3–937–13784–1

[SH11] Starke, Gernot ; Hruschka, Peter: *Software-Architektur kompakt - angemessen und zielorientiert*. Spektrum, 2011. – 1–118 S. – ISBN 978–3–8274–2834–9

[Som12] Sommerville, Ian: *Software Engineering.* Pearson Studium - IT, 2012. – 1–409 S. – ISBN 978–3868940992

[Sta11] Starke, Gernot: *Effektive Software-Architekturen, ein praktischer Leitfaden.* Hanser, 2011. – 1–409 S. – ISBN 978–3446436145

[Sta14] Statista: *Korrelation.* `http://de.statista.com/statistik/lexikon/definition/77/korrelation/`. Version: 2014

[Ste04] Steinberger, Thomas: *Kurzfristige Prognose des Stromverbrauchs in Vorarlberg auf Stunden- und Viertelstundenbasis.* Fachhochschule Voralberg, 2004

[Wal13] Walker, Mark: *SAP HANA Starter: SAP HANA Integration with Microsoft Excel.* `http://www.packtpub.com/article/sap-hana-integration-with-microsoft-excel` Zuletzt aufgerufen: 18.05.2014. `http://www.packtpub.com/article/sap-hana-integration-with-microsoft-excel`. Version: 01 2013, Abruf: 18.05.2014

[We13] Wasserwirtschaft e.V., Bundesverband der Energie-und: *Stromverbrauch im Haushalt.* Berlin : BDEW, 2013

[Wet14] Wetterdienst, Deutscher: *Wetterrekorde - Lufttemperatur.* `http://www.dwd.de/rekorde`. Version: 2014, Abruf: 16.12.2014

[WM05] Willmott, Cort J. ; Matsuura, Kenji: *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* `http://climate.geog.udel.edu/~climate/publication_html/Pdf/WM_CR_05.pdf/`. Version: 19.12.2005

[ZT11] Zavadskas, Edmundas K. ; Turskis, Zenonas: Multiple criteria decision making (MCDM) methods in economics: an overview. In: *Technological and Economic Development of Economy* 17 (2011), Nr. 2

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Planungsprozesse im Unternehmen aus fachlicher und organisatorischer Sicht und deren Ziele

Seminararbeit
im Rahmen der Projektgruppe VLBA inMemory Planung mit SAP HANA

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dirk Peters

Vorgelegt von:	Igor Perelman Eichenstraße 105, Zi.510 26131 Oldenburg Telefonnummer: 0163-1739525 igor.perelman@uni-oldenburg.de

Abgabetermin:	01.08.2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

1 Einführung **4**
1.1 Beschreibung der Thematik 4
1.2 Zielsetzung 4
1.3 Aufbau der Arbeit 4

2 Planung **5**
2.1 Planungsbegriff 5
2.2 Planarten 6
2.3 Grundsätze der Planung 6

3 Planungsprozess **7**
3.1 Der globale Planungsprozess 7
3.1.1 Phase der Informationsbeschaffung 8
3.1.2 Phase der Situations- und Zukunftsanalyse 8
3.1.3 Phase der Zielplanung 9
3.1.4 Phase der Strategieplanung 10
3.1.5 Phase der Maßnahmenplanung 11

4 Planungsorganisation **11**
4.1 Begriff der Organisation 12
4.2 Aufbauorganisation 12
4.2.1 Planungsträgerschaft 13
4.2.2 Aufgaben- und Größenabhängigkeit der Struktur 13
4.2.3 Wichtige Organisationsstrukturen 13
4.3 Ablauforganisation 14
4.3.1 Arbeitsinhalt 14
4.3.2 Arbeitszeit 15
4.3.3 Arbeitsort 15
4.3.4 Arbeitszuordnung 15

5 Praxisbeispiele des SAP-Einsatzes bei den Planungsprozessen **15**

6 Fazit **17**

Literaturverzeichnis **18**

Abbildungsverzeichnis

1 Managementfunktionen im Regelkreislauf-Modell, [Mag95]S.2 5
2 Phasen des globalen Planungsprozesses, [Ehr97] S.33 7
3 Unternehmensziele, [**?**] S.33 10
4 Strategien und inhaltliche Beschreibung, [Ehr97] S.200 11
5 Maßnahmen für die Strategien, [Ehr97] S.221 11
6 Organisationsvariablen nach Leavitt, [Mag95] S.120 12
7 Planung in Linienstellen, [Mag95] S.123 14
8 Planung durch den Stab, [Mag95] S.123 14

1 Einführung

Dieses Kapitel stellt die Beschreibung der Thematik (1.1), die Zielsetzung (1.2) und den Aufbau (1.3) der vorliegenden Hausarbeit dar.

1.1 Beschreibung der Thematik

Ein amerikanischer Top- Manager sagte einmal: „ Niemand plant zu versagen, aber die meisten versagen beim Planen“. Um dieses „Versagen“ zu vermeiden, gibt es in der Betriebswirtschaftslehre die Unternehmensplanung, die als Managementkonzept zur Unterstützung der Unternehmensführung gesehen wird. Die heutigen Unternehmen sind durch die sogenannten Megatrends herausgefordert, so zu handeln und zu agieren, dass auf langfristige Sicht Erfolg auf den relevanten Märkten gesichert wird. Die Innovation wird als zentraler Wachstumstreiber und Wettbewerbsfaktor angesehen. Das Wort Mobilität spiegelt sich nicht nur in der ständig steigenden Flexibilität der Mitarbeiter des Unternehmens wider, sondern auch in der IT, z.B. durch die Zusammenführung von Verkehrs- und Warenströmen. Der Wandel der Märkte zwingt die Unternehmen, die Kundenbindung durch Verstärkung produktionsbegleitender Service-Dienstleistungen zu realisieren. Um den oben genannten Megatrends erfolgreich gegenüberzutreten, ist von Seiten der Planungsträger eine korrekte und erfolgreiche Unternehmensplanung erforderlich. Unternehmerische Aktivitäten benötigen eine Planung, damit ihre möglichen Folgen überschaubar und ihr zukünftiger Erfolg so weit als möglich sichtbar gemacht werden kann.

1.2 Zielsetzung

Im Rahmen diese Seminararbeit sollen Grundlagen des Planungsprozesses sowie deren Umsetzung in der Aufbau- als auch Ablauforganisation erklärt werden. Anhand von Praxisbeispielen wird gezeigt, wie IT bei der Unternehmensplanung eingesetzt werden könnte. Das Ziel ist es, die grundlegende Komplexität des Planungsprozesses und die damit verbundenen Schritte darzustellen. Es soll gezeigt werden, aus welchen Schritten der Planungsprozess besteht und wie dieser organisatorisch umgesetzt wird.

1.3 Aufbau der Arbeit

Um die Zielsetzung der Arbeit zu erreichen, soll mit den Grundlagen der Planung angefangen werden. Es werden der Begriff Planung erläutert, die Planungsarten vorgestellt sowie erklärt, wer im Unternehmen als Planungsträger in Frage kommen kann. Abgeschlossen wird das Kapitel 2 mit den Anforderungen. Kapitel 3 befasst sich mit dem

Planungsprozess. Hier geht um die jeweiligen Schritte des Planungsprozesses, diese werden jeweils erläutert und anhand von Beispielen vervollständigt. Im Kapitel 4 findet sich eine Einführung in das Thema Planungsorganisation. Es werden 2 Begriffe, Aufbau- und Ablauforganisation, eingeführt. Neben allgemeinen Definitionen werden auch Bestandteile vorgestellt. Kapitel 5 beinhaltet eine kurze Darstellung des IT-Werkzeug-Einsatzes. Es werden 2 Success-Stories vorgestellt, mit denen gezeigt wird, wie Unternehmen durch den Einsatz von SAP-Software Verbesserungen in der Unternehmensplanung erreicht haben. Abschließend gibt es im Kapitel 6 eine Zusammenfassung dieser Seminararbeit und eine Schlussbetrachtung.

2 Planung

In den ersten Kapiteln wird der Begriff Planung erklärt, Regelkreismodell gezeigt Planarten, Planungsträger, sowie Grundsätze der Planung vorgestellt.

2.1 Planungsbegriff

Um sich dem Begriff des Planungsprozesses zu nähern, ist eine der Definition angebracht. PPlanung ist der Entwurf einer Ordnung, nach der sich das betriebliche Geschehen in der Zukunft vollziehen soll, sie ist das gedankliche, systematische gestalten des zukünftigen Handels “ ([Ehr97], S.19). „[...] in erster Linie „vorausschauendes Denken“, eine geistige Tätigkeit also, die darauf ausgerichtet ist, künftige Gegebenheiten, Wirkungen und Wechselbeziehungen überschaubar zu machen.“([Ham11],S. 108). „Planung kann als Bindeglied zwischen Zielsetzung und kalkuliertem Handeln gesehen werden“. ([WD08],S. 81) Aus diesen Begriffen folgt, dass die Planung zum einen zukunftsorientiert, zum anderen zielorientiert ist. Sie dient dazu, Handlungsalternativen zu finden, die einem Unternehmen helfen, die definierten Ziele durch bestimmte Handlungsweisen zu erreichen. Die Bedeutung der Planung kann man unter anderem anhand des Regelkreis-Modells nachvollziehen:

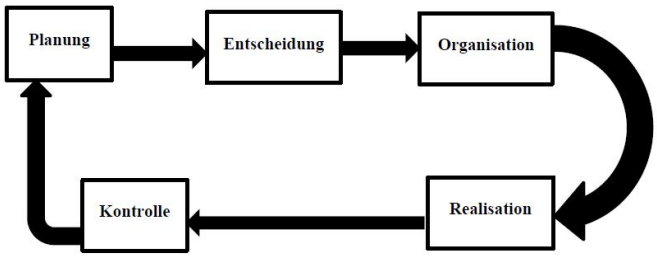


Abbildung 1: Managementfunktionen im Regelkreislauf-Modell, [Mag95]S.2

2.2 Planarten

In den Unternehmen gibt es verschieden Arten der Planung. Die Planung kann sich auf Zeiträume beziehen(Vgl. [DS13],S.327) :

- langfristige Planung für Zeiträume über fünf Jahre
- mittelfristige Planung für Zeiträume von ca. zwei bis fünf Jahren
- kurzfristige Planung für Zeiträume bis zu einem Jahr

Planung nach dem hierarchischen Überordnungsverhältnis der Planungsstufen kann man in eine strategische und eine operative Form aufteilen. Die strategische Planung befasst sich mit globalen Erfolgsfaktoren, Aktivitäten des Unternehmens für die nächsten fünf bis 10 Jahre. Die operative Planung meint die Umsetzung der strategischen Planung. Ein klassisches Unternehmen besteht aus verschiedenen Abteilungen, somit ergibt sich auch eine Möglichkeit der Planung in diversen Bereichen: Absatzplanung, Finanzplanung, Beschaffungsplanung Werden die Ziele definiert und ebenso die benötigten Handlungsmaßnahmen, so spricht man von der Planung auf der Grundlage des Inhalts (Vgl. [Mag95] S. 41).

2.3 Grundsätze der Planung

In der Zeiten der Globalisierung sind Märkte sehr dynamisch, das Kundenverhalten ändert sich schnell, d.h., Unternehmen sollten in der Lage sein, sich zum einem schnell an die Rahmenbedingungen anzupassen und zum anderen bei der Unternehmensplanung wichtige Grundsätze, die Mindestanforderungen darstellen, im Zuge dieser Fähigkeit zur Anpassung beachten (Vgl.[Ehr97],S.29):

- Langfristigkeit der Planung: Die Planung sollte permanent und nicht gelegentlich erfolgen. Langfristigkeit bezieht ich dabei nicht auf den Zeitraum, sondern auf das fortlaufende Planen.
- Vollständigkeit der Planung: Bei der Planung sollten die Interdependenzen zwischen relevanten Bereichen (Beschaffungsplanung, Absatzplanung, Investitionsplanung usw.) berücksichtigt werden.
- Anpassungsfähigkeit der Planung: Falls es Situationsänderungen im Unternehmen oder der Umwelt kommt, sollte die Planung auf diese reagieren können. Zum einen betrifft das die Reaktionsgeschwindigkeit und zum anderen die Handlungsalternativen.

- Stabilität der Planung: Werden die nötigen Anpassungsmaßnahmen durchgeführt, muss die Planung so stabil sein, dass Störungen die gesamte Planung nicht zerstören.
- Verbindlichkeit der Planung: Diese Anforderung betrifft die Mitarbeiter und ihre Verpflichtung, die Planvorgaben möglichst zu erreichen. Aufgabe des Unternehmens ist es dabei, die sachlichen Voraussetzungen zu schaffen.
- Kontrollierbarkeit der Planung: Die Planung ist kontrollierbar, wenn gewährleistet ist, dass alle Teilpläne nach den gleichen Planungsmethoden erstellt sind, die Planungen sind aufeinander abgestimmt und enthalten nur quantifizierbare Größen.
- Realisierbarkeit der Planungsvorgaben: Die geplanten Vorgaben müssen realisierbar und daraus folgt, dass sie realistisch sein sollten.

3 Planungsprozess

Die Unternehmensplanung besteht aus einer Menge Einzelaktivitäten. Diese müssen aufeinander abgestimmt, strukturiert und koordiniert werden. Das wird durch die Festlegung des Planungsprozesses erreicht, welcher die Gliederung und organisatorische Gestaltung des Planungsablaufs beinhaltet. (Vgl.[Kre97], S.37). Wichtig ist, anzumerken, dass der Planungsprozess sowohl an die strategische als auch an die operative Planung gerichtet ist.

3.1 Der globale Planungsprozess

Zur Verständnisvereinfachung wird dieses Kapitel mit einer graphischen Darstellung des globalen Prozesses der Planung, bestehend aus folgenden Phasen, eingeführt:

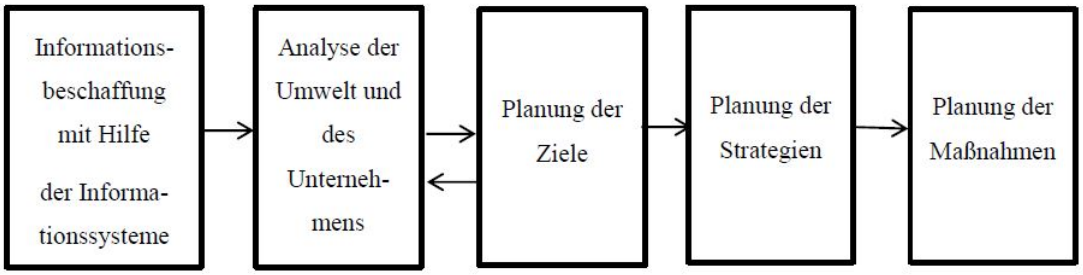


Abbildung 2: Phasen des globalen Planungsprozesses, [Ehr97] S.33

In den folgenden Abschnitten wird jede Phase mit ihren Aufgaben detailliert beschrieben.

3.1.1 Phase der Informationsbeschaffung

Informationen bilden die Grundlage der Planung. In Unternehmen herrscht eine hohe Informationsflut, welche viele nicht relevante Daten beinhaltet, d.h., diese könnten nicht aktuell, nicht präzise, schlecht formuliert und nicht zuverlässig sein. Aus diesem Grund gilt als Mindestanforderung an die Informationsbeschaffung, dass diese durch gezieltes, systematisches Handeln dem Planungsträger den Zugang zu den richtigen Informationen in der richtigen Form und zum richtigen Zeitpunkt ermöglicht (Vgl.[Hor06], S.323). Bei der Planung könnte interne und/oder externe Informationsquellen genutzt werden. Zu den internen Informationsquellen zählen das Unternehmen selbst bzw. die relevanten Fachabteilungen. Von besonderer Bedeutung ist das allgemeine Rechnungswesen, welches beispielweise die Informationen über die Umsätze, Verbindlichkeiten und Lagerbestände liefert. Die Abteilung der Kostenrechnung repräsentiert ebenso eine wichtige Informationsquelle, die alle möglichen Vorgänge bei der Leistungserstellung umfasst. Wird in ein Unternehmen eine eigene Marktforschung integriert, so liefert diese aktuelle Informationen zu den Themen Marktanalyse und Marktprognose (Vgl.[Ehr97], S.46). Die externen Informationsquellen befinden sich logischerweise außerhalb des Unternehmens und sind durch das Unternehmen nicht beeinflussbar, dazu zählen Behörden, statistische Ämter, Industrie – und Handelskammer, wirtschaftliche Institute, Fachbücher.

3.1.2 Phase der Situations- und Zukunftsanalyse

In dieser Phase wird eine strategische Analyse von Umwelt und Unternehmen durchgeführt. Um die strategische Planung von Zielen und deren Realisation sinnvoll zu gestalten, sind Kenntnisse über die Einzelteile des Marktes, Wissen über die Reaktionen der Konkurrenz, Kenntnisse über die Bedingungen der Umwelt auf einer Seite und auf der anderen Seite Erkenntnisse über die eigenen Möglichkeiten, Stärken und Schwäche des Unternehmens notwendig (Vgl.[Kre97], S.40). Die Analyse des politischen Umfelds umfasst die politische Entwicklung in Deutschland und außerhalb, um die unternehmensinternen Risiken und Chancen richtig einzuschätzen. Eine Analyse der gesetzlichen Umweltbedingungen ist notwendig, da es um die Aktivitäten des Staates und der Körperschaften geht, deren Verfolgung für ein Unternehmen sehr wichtig ist, z.B. Steuergesetzgebung, Außenhandelsgesetzgebung, Wettbewerbsordnung usw. (Vgl. [Ehr97], S.117). Die Marktanalyse gehört zu dem Bereich der Marktforschung und besteht aus drei Aspekten, der Marktanalyse, deren Aufgabe darin besteht, den Markt einmalig oder zu bestimmten Zeitpunkten zu analysieren, der Marktbeobachtung mit dem Ziel der permanenten Lieferung von Fakten und der Marktprognose, welche zukünftige Entwicklungen auf dem Markt behandelt

(Vgl. [Ehr97], S.117). Von großer Bedeutung für Entscheidungen im Prozess der Planung sind Informationen und Daten über die Konkurrenten. Als Resultat der Auswertung der Konkurrenzanalyse erhält die Führung des Unternehmens einen Überblick über das Produktportfolio und die relevanten Aktivitäten der Wettbewerber und kann anhand dieser eigene Möglichkeiten auf dem jeweiligen Markt einschätzen. Wichtig für den Leser ist, zu verstehen, dass die Hauptaufgabe der Konkurrentenanalyse im Vergleich der Möglichkeiten des eigenen Unternehmens mit denen der Konkurrenten besteht, daraus folgt, dass es notwendig ist, für die funktionierende vollständige Ausführung eine Unternehmensanalyse bzw. eine Stärken-/Schwächenanalyse durchzuführen (Vgl.[Ehr97], S.117). Die Konkurrentenanalyse kann sich auf folgende Bereiche beziehen (Vgl.[Voi93], S.98): Anzahl der Konkurrenten, Marketing-Instrumentarium, Absatzgebiete, Kundenstruktur, Innovation. Die wichtigen Quellen von Informationen sind folgende(Vgl. [Ehr97], S.120): veröffentlichte Jahresabschlüsse, Unternehmensberichte aus Fachzeitungen und –Zeitschriften, Pressekonferenzen, Verbandsmitteilungen. Die Unternehmensanalyse gibt Informationen über die Leistungsfähigkeit des eigenen Unternehmens. Die Stärken-/Schwächenanalyse beschäftigt sich mit der Bewertung von Ressourcen im Unternehmen. Dadurch, dass Stärken und Schwächen sich in den unterschiedenen Fachbereichen offenbaren, ist es empfehlenswert, diese Analyse von Spezialisten aus dem jeweils relevanten Bereich durchführen zu lassen. Wie bekannt, agiert das Unternehmen nicht allein auf dem Markt, somit sind bei der Analyse auch die Stärken/Schwächen der bedeutendsten Konkurrenten mit einzubeziehen. Wurden eine Umweltanalyse und eine Ermittlung von Stärken/Schwächen durchgeführt, fasst man diese zusammen und gewinnt aus diesen Erkenntnissen einen Überblick über Chancen für das Unternehmen und Risiken hinsichtlich der Erreichung des Unternehmensziele (Vgl. [BB11], S.129).

3.1.3 Phase der Zielplanung

Ein Unternehmen benötigt eine Beurteilung dahingehend, ob seine Aktivitäten gut oder schlecht sind, dafür werden auch Ziele eingeplant, diese dienen der Messung von Handlungen. Grundsätzlich werden die Ziele in drei Dimensionen festgelegt: Die Zielvorschrift oder auch das Zielausmaß besagt, was mit der Zielgröße geschehen soll, sie beschäftigt sich mit der Extremierung, d.h. Minimalisierung (ZMin) und/oder Maximierung (ZMax) des angestrebten Ausmaßes der Ziele. Die Zieldauer bestimmt den Zeitraum, in dem der angestrebte Zustand erreicht werden soll, sowie die Zeitperiode, innerhalb derer der Zielinhalt gelten soll. Der Zielinhalt wird in den wirtschaftlichen, den sachlichen und den personellen unterschieden. Die wirtschaftlichen Ziele sind für das „Überleben“ des Unter-

nehmens wichtig, z.B. eine Vergrößerung der Marktanteile. Die sachlichen Ziele beziehen sich auf die Dienstleistungen oder Produktion des Unternehmens und dienen somit der Realisierung von wirtschaftlichen Zielen. Der Bereich von personellen oder auch sozialen Zielen beschäftigt sich mit persönlichen Motiven von Einzelpersonen oder Personengruppen (Vgl. [Mag95], S.48). Die folgende Tabelle zeigt, in welchen Bereiche Ziele formuliert werden können:

1. Marktleistungsziele	**5. Rentabilitätsziele**
- Produktqualität	- Gewinn
- Kundenservice	- Rentabilität des Eisenkapitals
2. Marktstellungsziele	**3. Finanzwirtschaftliche Ziele**
- Umsatz	- Kreditwürdigkeit
- Marktanteile	- Liquidität

Abbildung 3: Unternehmensziele, [**?**] S.33

Wichtig ist, anzumerken, dass bei der Zielplanung die Ergebnisse der Unternehmensanalyse mit einbezogen werden sollten, da die Ziele darauf ausgerichtet werden sollten, die Stärken des Unternehmens auszubauen und Chancen zu nutzen.

3.1.4 Phase der Strategieplanung

Der Prozess der Entwicklung von Strategien zur Erreichung der zuvor definierten Ziele erfordert von den Verantwortlichen ein hohes Maß an Kreativität und Sachkenntnissen. Grob wird diese Phase in zwei Schritte aufgeteilt:

1. Suche durchführbaren Möglichkeiten und deren Herleitung (Alternativen)
2. Bewertung und Analyse der Durchführbarkeit

Zu 1): Hier wird zuerst die laufende Strategie überprüft und gegebenenfalls eine neue entwickelt. Für diesen Schritt kämen folgende Techniken in Frage: Brainstorming, Szenario-Technik, Portfolio-Technik.

Zu 2): Es ist durchaus möglich, dass nach dem Suchprozess mehrere Alternativen zur Verfügung stehen, somit wird ein Bewertungsvergleich durchgeführt, um die richtige, also die Strategie, mittels derer die Zeile am besten erreicht werden können, zu ermitteln.

An dieser Stelle besteht für die Entwickler die Aufgabe, Bewertungskriterien aufzustellen und eine Bewertungsmethode auszuwählen. Die gängigsten Methoden aus der Praxis sind: Kosten-Nutzen-Analyse, Nutzwertanalyse usw. (Vgl. [Ehr97], S.182).

Strategie/Bereich	Hinweis für die inhaltliche Beschreibung der Strategien
Marketing→ Technologieorientierung	- Führerschaft - Imitation
Fertigung→ Kapazität	- Abbau von überschüssigen Kapazitäten - Fremdbezug
Beschaffung→ Lieferantenauswahl	- Wirtschaftliche/technische Leistungsfähigkeit der Lieferanten - geringe/große Anzahl

Abbildung 4: Strategien und inhaltliche Beschreibung, [Ehr97] S.200

3.1.5 Phase der Maßnahmenplanung

Beim letzten Schritt im Planungsprozess geht es um die zur optimalen Zielerreichung benötigten Maßnahmen, das heißt, die zuvor formulierten Strategien werden konkretisiert. Die Aktivitäten werden oft über die gleichen Instanzen festgelegt, die sich auch mit der Strategieplanung auseinandergesetzt haben. Als Beispiel seien einige Maßnahmen für zwei Strategien aufgeführt:

Marktdurchdringung	Marktentwicklung
- Aufbau eines neuen Vertriebsweges(z.B. Online-Handel)	- Erweiterung der Außendienstorganisation
- Veränderung der Werbeträger (z.B. Radio-Spots statt Anzeigen in Zeitschriften)	- Preissenkung zur Überbrückung der Markteintrittsbarrieren

Abbildung 5: Maßnahmen für die Strategien, [Ehr97] S.221

4 Planungsorganisation

Bisher wurde konkret der Planungsprozess beschrieben. Dieser wird von den Planungsverantwortlichen, die eine bestimmte Rolle in der (Unternehmens-)Organisation vertreten, beeinflusst. In den nächsten Kapiteln werden die Optionen zur organisatorischen Gestaltung des Planungsprozesses detaillierter erläutert.

4.1 Begriff der Organisation

„Eine Organisation liegt dann vor, wenn mehrere Personen zur gemeinsamen Aufgabenerfüllung harte und weiche Technologien(Computer, Modelle, Methoden und Verfahren) einsetzen und dazu eine Kommunikationsstruktur aufbauen“ ([Mag95], S.120).

Die Abbildung repräsentiert die Überlegung von Leavitts, der zur Beschreibung der Organisation 4 Variablen einführt: Task (Aufgabe), People (Personen), Technology(Computer, Methoden) und Structure (Struktur):

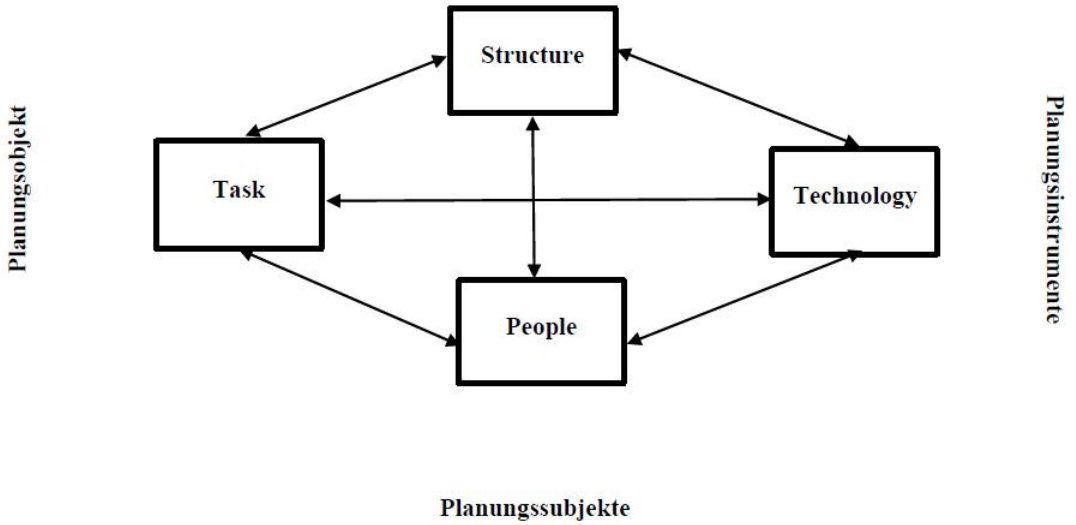


Abbildung 6: Organisationsvariablen nach Leavitt, [Mag95] S.120

4.2 Aufbauorganisation

Die Aufbauorganisation teilt die Aufgaben eines Unternehmens in Aufgabenbereiche und bestimmt die Stellen und Abteilungen, die diese bearbeiten sollen. Übertragen auf den Planungsprozess meint dies, wie die Aufteilung der Aufgaben auf die Verantwortlichen

für die Planung erfolgt und welche Personen mit entsprechenden Kompetenzen zu den Aufgabenträgern werden. (Vgl.[Kre97], S.190).

4.2.1 Planungsträgerschaft

Die bekanntesten Funktionsträger, denen man planungsaufgabenerfüllende Fähigkeiten theoretisch zutraut und Planungen praktisch überträgt, können folgende sein (Vgl.[Kre97], S.121): Instanzen des Linienmanagements, Zentrale Planungsstäbe und –abteilungen, Typische Querschnittsregler (z.B. Project Manager), Besondere Organisationseinheiten (z.B. Controller), Übergeordnete Einheiten (z.B. Konzernzentrale).

4.2.2 Aufgaben- und Größenabhängigkeit der Struktur

Es gibt keine Aufbaustruktur für die Planung im Unternehmen, die für alle Unternehmen fortwährend gilt. Die zwei Faktoren, die dabei eine wichtige Rolle spielen, sind zum einem die Art der Planungsaufgabe und zum anderen die Unternehmensgröße. Die Art der Planungsaufgabe sagt etwas darüber aus, wie umfangreich die Planungsaufgabe ist, außerdem über den Planungshorizont und letztlich über die Durchführungshäufigkeit (permanent, gelegentlich oder einmalig). Bereits durch diese Merkmalen könnten die ersten Schwierigkeiten hervorgerufen werden, da z.B. für die einmalige Planung höchstwahrscheinlich gar keine Aufbaustruktur notwendig, sondern ein Planungsmeeting mit Interessierten veranstaltet wird, welches sich nach der Findung der Lösung der Planungsaufgabe sofort wieder auflöst. Die Unternehmensgröße beeinflusst die Art der Strukturierung ebenfalls. In kleineren Unternehmen gibt es oft gar keine Probleme mit Strukturierung, da dort oft die Planungsverantwortlichen und Planungsträger dieselben Personen sind, während bei mittleren und größeren Unternehmungen die Planungsobjekte umfangreicher sind und somit an die spezielle Stelleninhaber bzw. Planungsträger abgegeben werden. (Vgl. [Mag95], S. 122)

4.2.3 Wichtige Organisationsstrukturen

Der häufigste Strukturtyp zur Durchführung der Planungsaufgaben in der mittleren Unternehmung erfolgt in Linienstellen.

Wie in der Abbildung 7 Links, ersichtlich, wird die Planung direkt von den Leitern der zweiten Ebene (z.B. Produktionsleiter) erfüllt. Die Abbildung 7 Rechts, sieht in der zweiten Ebene eine Abteilung „Planung“ vor, dadurch können eine Spezialisierung und Vorkoordination der Teilpläne erfolgen, allerdings setzt das einen regen Informationsaustausch

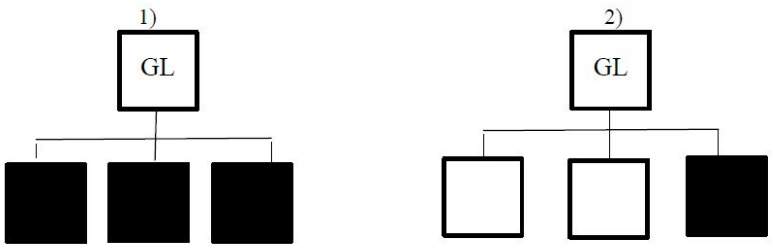


Abbildung 7: Planung in Linienstellen, [Mag95] S.123

zwischen den Fachabteilungen der gleichen Ebene voraus. Als klassische Organisationstruktur gibt es die Planung durch die Stäbe. Der Stab in Abbildung 8 ist der Geschäftsleitung untergeordnet, somit ist er als Generalstab für langfristige Unternehmensplanung verantwortlich; auch zählen zu seinen Aufgaben die Entwicklung von Zielen und Strategien, er ist aber zudem aus unternehmerischer Sicht für die Koordination aller anderen planerischen Aktivitäten zuständig. (Vgl. [Mag95], S.123)

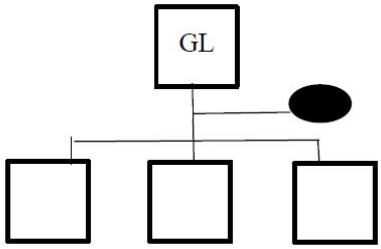


Abbildung 8: Planung durch den Stab, [Mag95] S.123

4.3 Ablauforganisation

„Aufgabe der Ablauforganisation ist die Gestaltung der Arbeitsabläufe, die raum-zeitliche Strukturierung dieser Verrichtungen und Verrichtungsfolgen. Der Arbeitsablauf muss in verschiedener Hinsicht geordnet werden. Man unterscheidet die Ordnung des Arbeitsinhalts, der Arbeitszeit, des Arbeitsraums und der Arbeitszuordnung“ ([WD08] Wöhe,S. 121). Es geht hier um die Durchführung von Planungsprozessen.

4.3.1 Arbeitsinhalt

Der Arbeitsinhalt eines Planungsprozesses ergibt sich aus dem Planungsobjekt (Aufgabe). In der Aufbauorganisation geht es um die Festlegung von Aufgaben, nicht aber um die Zusammenhänge zwischen diesen bzw. wie diese konkret ablaufen müssen. Es werden die Aufgaben analysiert und dadurch die Teilaufgaben gebildet, die verrichtet werden müssen. Beachten wir den Planungsprozess genauer, so könnte man für die Bildung der Teilaufgaben die Phasenschemata der Planung verwenden und in dieser Ordnung die Teilaufgaben abarbeiten: Zielanalyse, Problemanalyse, Alternativanalyse, Prognose, Bewertung und Entscheidung (Vgl. [Mag95], S.126).

4.3.2 Arbeitszeit

Diese Kategorie kann man auch als Teilphase der Planungszeit bezeichnen. Es wird zum einem festgelegt, wie lange die Ausführung der jeweiligen Teilaufgabe dauern darf, zum anderen geht es um die Festlegung des Anfangs- und Endtermins des Planungsprozesses. Als Beispiel könnte es in der Praxis so aussehen, dass zum 01.08. die Teilplanungen abgeschlossen werden sollen, um den Unternehmungsgesamtplan bis zum 01.11. des gleichen Jahres erstellen zu können. Gibt es Abhängigkeiten zwischen den Teilplanungen und sind diese nicht parallel durchführbar, wäre es sinnvoll, die kritischen Zwischentermine zu integrieren, da somit keine Gefahren bzgl. des Endtermins bestehen (Vgl. [Zel11] S.50).

4.3.3 Arbeitsort

Der Arbeitsort bezeichnet den räumlichen Aspekt der Ablauforganisation. Sind mehrere Personen am Planungsprozess beteiligt, die sich nicht in einer räumlichen Organisationeinheit befinden, so gilt es als besonders wichtig, eine ausgebaute Kommunikationsstruktur zu schaffen, die festlegt, wo die Informationen gespeichert sind und wie deren Transportwege sich gestalten (Vgl. [Mag95], S.126).

4.3.4 Arbeitszuordnung

Die Arbeitsordnung beschäftigt sich mit der Festlegung der notwendigen Kommunikation zwischen Organisationseinheiten bei der Planungsdurchführung. Wichtiger Aspekt dieser Komponenten ist es, dass hier die Rollenverteilung zwischen den Beteiligten (personellen Planungsträgern) im arbeitsteiligen Prozess stattfindet. (Vgl. [Zel11] S.50)

5 Praxisbeispiele des SAP-Einsatzes bei den Planungsprozessen

„Gut geplant ist halb gewonnen“, diesem Motto folgen moderne große Unternehmen in verschiedenen Branchen, die wettbewerbsfähig bleiben wollen. SAP AG ist ein führender Anbieter von Unternehmenssoftware aus Deutschland. In den folgenden Absätzen werden Beispiele des erfolgreichen Einsatzes von SAP-Anwendungen bei den Planungsprozessen aufgeführt. 1) „Universal Music Group: Transparente Planungsprozesse stärken die globale Spitzenposition. Heute führt die Finanzabteilung bei Universal Music mit der Anwendung SAP Business Planning and Consolidation Planungsprozesse und Berichtswesen nahtlos zusammen. „Wir wissen heute jederzeit, wo wir stehen, und können bessere Entscheidungen treffen“, sagt von Wiedebach. Die Integration reduziert Fehler und schafft Transparenz, das Unternehmen kann die aktuelle Finanzleistung besser mit den Zielwerten vergleichen.“ [AGb]. 2) „Die Rittal GmbH & Co. KG. Der führende Anbieter von Schaltschränken und Lösungen zur Stromverteilung, Klimatisierung und für IT-Infrastrukturen beschäftigt weltweit 10.000 Mitarbeiter in über 60 Tochtergesellschaften und 40 Vertretungen. Diese dezentralisierte Struktur schlug sich bis vor Kurzem auch in den Planungsprozessen nieder. So nutzten die Gesellschaften bislang zu Planungszwecken Excel-Dokumente. Die auf diese Weise erfassten Zahlen gingen an die Konzernzentrale ins hessische Herborn – zur manuellen Konsolidierung. Das kostete Zeit und war fehlerträchtig, Tagesaktualität somit so gut wie gar nicht zu verwirklichen. Mit dem Umstieg auf die Anwendung SAP BusinessObjects Planning and Consolidation wollte Rittal diese Prozesse nachhaltig verbessern.“ [AGa]

6 Fazit

In der modernen Zeit der Globalisierung, des ständig wechselnden Kundenverhaltens und starken Wettbewerbsdrucks ist eine kontinuierliche Kontrolle des eigenen Unternehmens als auch der Umwelt sehr wichtig. Diese Seminararbeit verdeutlicht die Notwendigkeit und Wichtigkeit des Planungsprozesses im Unternehmen. Unternehmensplanung heißt, unternehmerische Ziele zu formulieren, die man mit dem Unternehmen im kurzfristigen und langfristigen Zeitverlauf erreichen möchte. Bei kurzfristigen Zeiträumen geht es insbesondere darum, den Gewinn zu erhöhen. Bei langfristigen Zeiträumen ist die primäre Aufgabe die Sicherung der Existenz des Unternehmens. Nur wenn eine Zielsetzung in Unternehmen vorgenommen wird, wissen sie, wie ihre vorhandenen Ressourcen optimal einzusetzen sind, da bekannt ist. Findet keine Planung statt, so besteht die Gefahr, sich zu verzetteln, da die Ausrichtung auf einen Punkt fehlt. Eine Gefahr, die dabei entstehen kann: durch fehlendes systematisches Beobachten der Veränderungen an den relevanten Märkten werden Trends möglicherweise nicht so früh oder gar nicht erkannt, dadurch ist eine angemessene Veränderung und, was wichtig ist, zeitgerechte Reaktion gar nicht möglich. Unter anderem entgehendem Unternehmen möglicherweise Chancen, die dem Betrieb neue Einkünfte einbringen würden. Eine fundierte Unternehmensplanung ist daher für alle Unternehmen, egal ob klein oder Mittelstand, ob Modeindustrie oder Maschinenbau, ein Muss.

Literatur

[AGa] SAP AG.

[AGb] SAP AG. `http://de.news-sap.com/2013/02/18/universal-music-group-transparente-planungsprozesse-staerken-die-globale-spitzenp` Zugriff Juli 2014, year = 2014.

[BB11] Rainer Bergmann and Michael Bungert. *Strategische Unternehmensführung, 1.* 2011.

[DS13] Ralf Dillerup and Roman Stoi. *Unternehmensführung.* Vahlen, 2013.

[Ehr97] Harald Ehrmann. *Unternehmensplanung.* Kiehl, 1997.

[Ham11] Richard Hammer. *Planung und Führung.* Oldenbourg Verlag, 2011.

[Hor06] Péter Horváth. *Controlling.* Vahlen, 2006.

[Kre97] Hartmut Kreikebaum. *Strategische Unternehmensplanung.* Kohlhammer, 1997.

[Mag95] Wolfgang Mag. *Unternehmungsplanung.* Vahlen, 1995.

[Voi93] Kai-Ingo Voigt. *Strategische Unternehmensplanung.* 1993.

[WD08] Günter Wöhe and Ulrich Döring. *Allgemeine Betriebswirtschaftslehre, Verlag Franz Vahlen, München, 23.* Auflage, 2008.

[Zel11] Helmut Zell. *Die Grundlagen der Organisation: Lernen und Lehren.* Books on Demand, 2011.

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Chancen und Herausforderungen in der klassischen Unternehmensplanung

Seminararbeit
Im Rahmen des Moduls Projektgruppe

Betreuer:	Jens Siewert
Vorgelegt von:	Steffen Scheer steffen.scheer@informatik.uni-oldenburg.de
Abgabetermin:	2. August 2014

Inhaltsverzeichnis

1 Motivation **1**

2 Grundlagen **1**

3 Chancen **3**
3.1 Prognose 4
3.2 Zukunftsplanung 6
3.3 Funktionen der Unternehmensplanung 6

4 Herausforderungen **9**
4.1 Dynamik der Umwelt 9
4.2 Interessenkonflikte 11
4.3 Begrenzte kognitive Fähigkeiten 12
4.4 Gesetzmäßigkeiten des unternehmerischen Handelns 13

5 Fazit **14**

Literaturverzeichnis **16**

Abbildungsverzeichnis

1 Management-Regelkreis 2
2 Zeitdimensionen der Unternehmensführung 3
3 Ablauf der Delphi-Methode am Beispiel eines Projekts 6
4 Denkansätze zu Ursache-Wirkungs-Zusammenhängen 13

1 Motivation

Im Unternehmen laufen diverse Prozesse ab, die als ganzes betrachtet von äußerster Komplexität sind. Die Unternehmensplanung ist schon lange eine der wichtigsten Aufgaben im strategischen Management und bleibt stets Betrachtungsgegenstand zahlreicher Publikationen, wie Recherchen gezeigt haben. Die Unternehmensplanung lässt Chancen entstehen, die ohne sie nicht nutzbar gewesen wären. Die Prognose kann dabei als Grundvoraussetzung der Planung angesehen werden. Das Koordinieren und Entscheiden ist ohne die Planung ausgeschlossen. Entscheidungen „aus dem Bauch“ heraus liegen nicht immer richtig und können im Zweifelsfall schwerwiegende Folgen für das Unternehmen und seine Stakeholder haben. Eine Herausforderung ergibt sich dabei unter anderem aus den kognitiven Grenzen der entscheidenden Individuen und Instanzen. Im Wesentlichen entscheiden aber diverse Faktoren über den nachhaltigen Erfolg einer adäquaten Unternehmensplanung. Daher forciert diese Ausarbeitung – wie der Titel vermuten lässt – die Erörterung verschiedener Chancen und Herausforderungen der klassischen Unternehmensplanung im Rahmen des bestehenden wissenschaftlichen Diskurs. Anschließend an die einleitende Motivation, erfolgt zunächst die Klärung einiger Grundlagen der Planung. Drauffolgend werden die verschiedenen Chancen der Planung mit ihren Funktionen im Unternehmen beleuchtet. Äquivalent dazu widmet sich die Arbeit dann den Herausforderungen der Planung und betrachtet insbesondere die Wissenskomponente, die in der Planung eine Rolle spielt.

2 Grundlagen

In diesem Kapitel sollen die Grundlagen der klassischen Unternehmensplanung geklärt werden. Nach Gutenberg ist Planung „der Entwurf einer Ordnung, nach der sich das betriebliche Geschehen in der Zukunft vollziehen soll“ [[Gut83] zitiert nach [Ehr06, S. 23]]. Sie ist das gedankliche, systematisch Gestalten des zukünftigen Handelns [Ehr06, S. 23]. Dabei muss beachtet werden, dass dem Entwerfen dieser Ordnung ganz unterschiedliche Ausgangsszenarien zugrunde liegen. Die Planung darf nicht mit der Prognose verwechselt werden, da diese versucht vorauszusagen wie hoch die Wahrscheinlichkeit unterschiedlicher Ereignisse ist. Die Planung hingegen soll klären welche Entscheidungen zu treffen sind, damit künftige Ereignisse eintreten. Der Planung und Prognose gemein ist, dass sich beide Vorgehensweisen mit Ereignissen der Zukunft beschäftigen [Ehr06, S. 47]. Eine Planung ist ohne die Prognose nicht möglich, weshalb die Unternehmensplanung auch die Prognose zum Untersuchungsgegenstand macht.

Planungsprozess

> Der Planungsprozess erstreckt sich auf alle Planungsbereiche und Planungshandlungen in einem Unternehmen, er ist Ausdruck der Gliederung, Ordnung und arbeitsteiligen Gestaltung des Planungsablaufs und der Phasenbildung [Ham98, S. 68].

Die Planung durchdringt alle Bereiche des unternehmerischen Handelns und ist damit eine Querschnittsfunktion. Wie auf Abbildung 1, dem Management-Regelkeis nach Schön, zu sehen, hat das *Entscheiden* und das *Planen und Gestalten*, also das Ermitteln von Plan-Werten und das *Setzen von Zielen*, also Zielgrößen, direkten Einfluss auf den Planungsprozess. Jene Zielgrößen sind mithilfe von Abweichungen zu *kontrollieren*, *vergleichen*, *analysieren* und *berichten*, um Ist-Werte zu *realisieren*, die bestenfalls den zuvor gesetzten Zielen entsprechen. All diesen Prozessen zugrunde liegt die *Information*, die *Kommunikation* und die *Koordination*. Der Planungsprozess wird maßgeblich vom be-

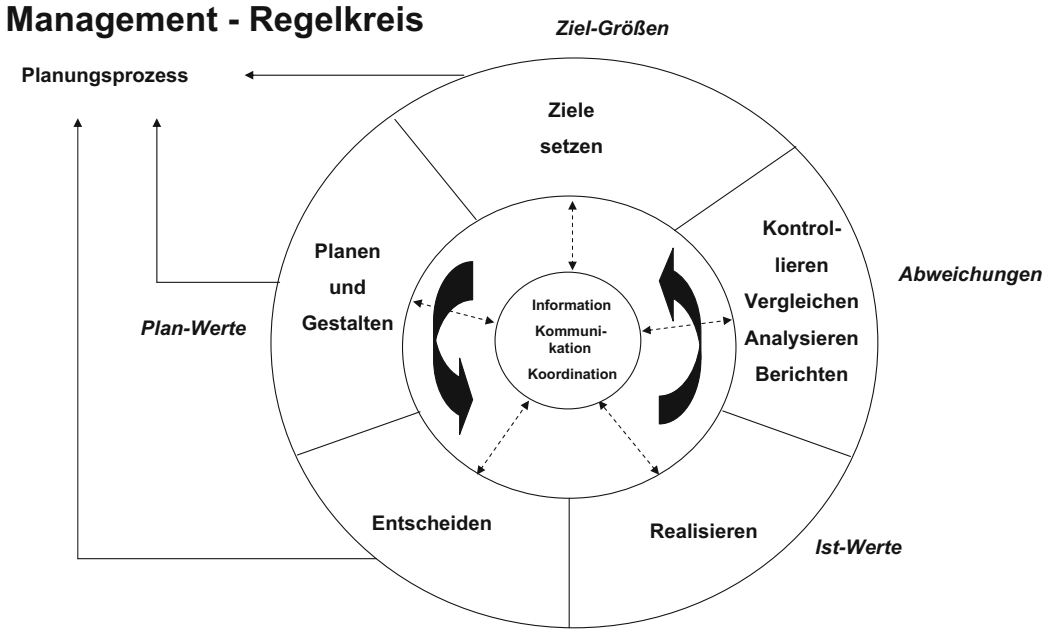


Abbildung 1: Management-Regelkreis nach Schön [Sch12, S. 18]

treffenden Planungszeitraum, also der Zeitdimension der Planung, beeinflusst (vgl. Inhalt der Planung auf Abbildung 2). Hier unterscheidet sich auch die operative von der strategischen Planung. Dabei müssen „strategische, taktische und operative Aufgaben [...] miteinander verzahnt werden“ [Sch12, S, 19]. Eine Möglichkeit dies zu realisieren ist die Top-Down-Planung wie in Abbildung 2 zu sehen. Hierbei erfolgt zuerst eine generelle

Zielplanung mit einem Zeithorizont von mehr als fünf Jahren. Diese bezieht sich auf die Unternehmenskultur, die Unternehmensphilosophie sowie Grundsatzziele [Ehr06, S. 50]. Die strategische Planung betrifft unterschiedliche Geschäftsfelder und Regionen eines Unternehmens für etwa fünf Jahre. Dabei werden laut Ehrmann [Ehr06, S. 50] strategische Ziele und Strategien festgelegt. Die Mittelfristplanung umfasst einzelne Maßnahmen und Projekte innerhalb eines Unternehmens. Hier kann von einem Planungshorizont von zwei bis drei Jahren ausgegangen werden. Die operative Planung bezieht sich auf einzelne Prozesse und Ressourcen in einem Zeitrahmen von unter einem Jahr, wobei operative Ziele und Maßnahmen [Ehr06, S. 50] definiert werden. Die operative Planung wiederum reicht die einzelnen Ziele mit der Steuerung und Kontrolle direkt an die durchführenden Instanzen weiter. Die Informationen laufen dabei in entgegengesetzter Richtung und erreichen je nach Priorität unterschiedliche Ebenen der Unternehmensplanung.

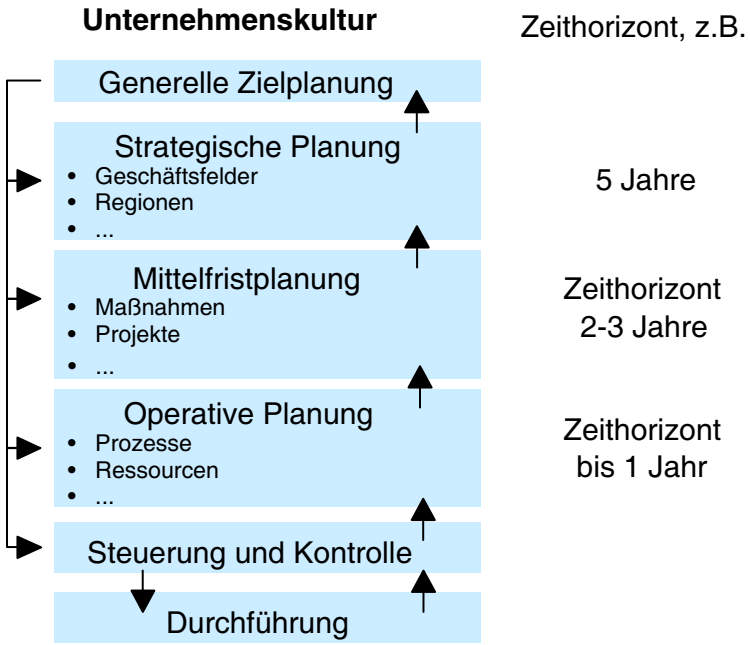


Abbildung 2: Zeitdimensionen der Unternehmensführung [Sch12, S. 19]

3 Chancen

Der dauerhafte und künftige Fortbestand eines Unternehmens zählt zu den elementaren betriebswirtschaftlichen Zielsetzungen. Erst auf Grundlage einer gesicherten Unterneh-

mensexistenz können Gewinne erwirtschaftet werden. Wesentlich hierfür ist eine adäquate Unternehmensplanung auf Basis derer begründete Entscheidungen getroffen werden können.

3.1 Prognose

Die Planung, also das bedingte Treffen von Entscheidungen, verbessert die Prognosemöglichkeit, welche wiederum das Planen erleichtert. Folglich existiert eine Wechselwirkung zwischen Prognose und Planung, die sich in der Unternehmensplanung über das Tauschen von Informationen manifestiert. Eine Prognose wäre genauer betrachtet ohne Planung nicht möglich. Es ergibt sich erst durch die Planung die Möglichkeit langfristig auf dynamische Marktveränderungen und sich verändernde Unternehmensbedingungen zu reagieren und negativen Entwicklungen entgegen zu wirken. Eine Prognose basiert dabei auf Daten oder Erfahrungswerten aus der Vergangenheit und ist stets mit Unsicherheit behaftet [BH09, S. 301]. Die Qualität einer Prognose hängt dabei nach Bea [Bea11, S. 788] vom

- Umfang der Bedingungen, unter denen die Prognoseaussage gelten soll,
- von der Allgemeinheit der Aussage oder des Prognosegegenstandes,
- von der Güte des Erklärungsmodells,
- von der Fehlerfreiheit der darin benutzten Daten und
- von der zeitlichen Stabilität des Erklärungsmodells,
- von der Berücksichtigung der Selbstzerstörung oder Selbstbestätigung der Prognose

ab. Eine Prognose ist für unterschiedliche Geschäftsfelder, Projekte, Prozesse und Ressourcen möglich. Es wird zwischen zwei unterschiedlichen Verfahrenstypen unterschieden:

- Systematische Verfahren
- Intuitive Verfahren

Unter systematischen Verfahren werden diskrete bzw. mathematische Vorgehensweisen verstanden, die versuchen einen Zusammenhang zwischen zu prognostizierenden und bekannten oder angenommenen Variablen herzustellen [Ehr06, S. 48]. Hierzu zählen etwa die Regressionsanalyse, Simulation, Trendextrapolation, das Aufstellen von Wachstumsfunktionen oder das Ermitteln von gleitenden Durchschnitten [Ehr06, S. 48]. Die intuitiven Verfahren basieren dabei auf Kreativtechniken, Systemanalysen oder Befragungen. Letztere ist eine der wichtigsten intuitiven Prognoseverfahren, weshalb viele unterschiedliche Befragungstechniken existieren. Unterschieden werden [BH09, S. 303f.]

- Repräsentativbefragungen und
- Expertenbefragungen.

Bei einer Repräsentativbefragung erfolgt eine Stichprobenziehung aus einer bestimmten Personengruppe (bspw. den Kunden eines Unternehmens), welche mit einem offenen oder geschlossenen Fragebogen befragt wird. Ein offener Fragebogen lässt die Möglichkeit von Freitextantworten, wobei ein geschlossener Fragebogen nur quantitative Antwortmöglichkeiten zulässt. Die Fragen zielen in der Regel auf das Verhalten des Befragen ab [BH09, S. 303], um künftiges Handeln und Reagieren der entsprechenden Personengruppe in hypothetischen Situationen antizipieren zu können. Problematisch daran ist der Schluss der Stichprobe auf die Grundgesamtheit [BH09, S. 303].

Die Expertenbefragung betrachtet nicht das Wissen oder Verhalten einer Menge von Individuen sondern zieht lediglich die Erfahrung und das Fachwissen einzelner Experten heran. Im Gegensatz zur Repräsentativbefragung besteht hier nicht das Problem vom Schluss auf die Grundgesamtheit. Vielmehr ist das Auswählen des geeigneten Experten die Hauptherausforderung. „Neben der einmaligen Expertenbefragung hat sich mit der Delphi-Methode ein Verfahren der mehrfachen Expertenbefragung etabliert“ [BH09, S. 304]. Die Dephi-Methode hat primär das Ziel die unterschiedlichen Expertenmeinungen zusammenzuführen [Mai14]. Die Befragung läuft dabei etwa wie in Abbildung 3 dargestellt ab und gliedert sich nach Drews und Hillebrand [DH07, S. 58] in vier Schritte:

1. **Schritt: Vorbereitung** Zunächst wird das Problemfeld beschrieben. Unterschiedliche Kriterien für die Bewertung werden definiert und geeignete Experten ausgewählt. Termine für die folgenden Schritte müssen festgelegt werden.

2. **Schritt: Befragung** Die eigentliche Befragung wird durchgeführt. Die unterschiedlichen Experten geben ihre Antworten unabhängig voneinander.

3. **Schritt: Auswertung** Die Fragebögen werden ausgewertet und stark abweichende Antworten müssen von den jeweiligen Experten begründet werden. Alle Experten werden über die Mittelwerte der Befragung informiert. Die Begründungen mit den starken Abweichungen werden ebenfalls an die Beteiligten weitergegeben. Diese müssen nun erneut eine Einschätzung abgeben.

4. **Schritt: Ergebnis** Nachdem die Schritte zwei und drei ungefähr zwei- bis dreimal wiederholt wurden und sich der Konsensprozess stabilisiert hat, haben sich die Expertenmeinungen auf die überzeugendsten Argumente und Prognosen reduziert.

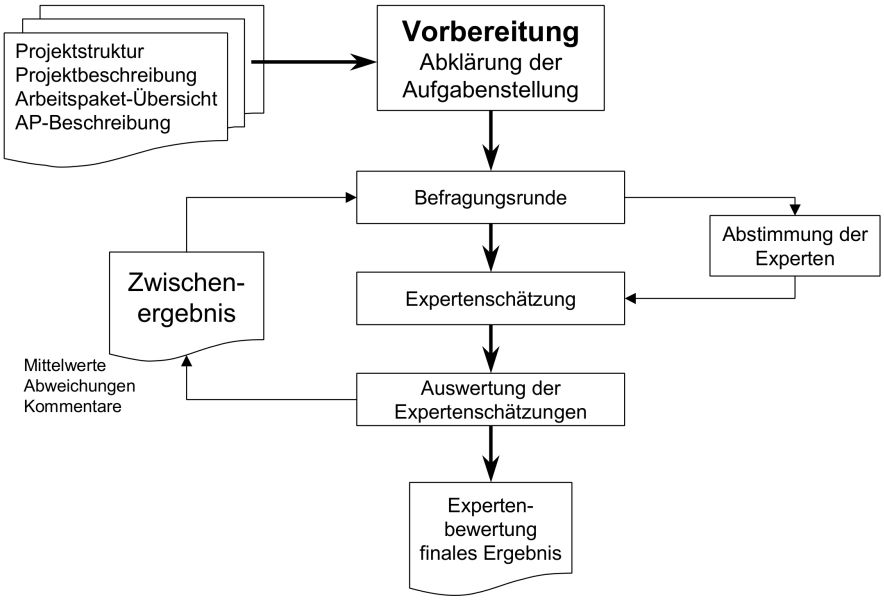


Abbildung 3: Ablauf der Delphi-Methode am Beispiel eines Projekts [DH07, S. 57]

3.2 Zukunftsplanung

„Der Wunsch des Menschen, in die Zukunft zu schauen, dürfte so alt sein wie die Menschheit selbst“ [MR12, S. 3]. Dabei ist es insbesondere in der heutigen, immer komplexer werdenden Welt mit globalen Einflüssen zunehmend von Bedeutung sichere Entwicklungsvorhersagen zu treffen. Entscheidungsträger stehen dabei vor einem Dilemma: Immer größere und umfangreichere und damit auch unüberschaubare Einflüssgrößen erfordern eine immer bessere Planung. Einerseits stehen umfangreichere Systeme zur Planung zur Verfügung, andererseits wird es aber auch schwieriger alle Input-Parameter richtig zu bewerten. Jedes Unternehmen hat dabei eine andere Position im Wettbewerb und auch unterschiedliche Ausgangssituationen, die nur mit Hilfe von umfangreichen Know-How richtig analysiert werden können. Trotz allem erwächst die Möglichkeit der Zukunftsplanung erst aus einer konsequent angewandten Unternehmensplanung.

3.3 Funktionen der Unternehmensplanung

„Die zentrale Aufgabe der Unternehmensplanung besteht darin, zukünftige, im Verlauf der betrieblichen Unternehmensprozesse entstehende Entwicklungen zumeist unter Unsicherheit zu quantifizieren“ [Sch12, S. 24]. Daraus resultiert, dass die Unternehmensplanung

verschiedene, Funktionen[1] erfüllt.

Entscheidungensfunktion

Das Entscheiden über Kriterien und das Auswählen von Alternativen ist die Grundlage einer jeden Strategie und Planung [BH09, S. 72]. Ohne eine Entscheidung kann auch kein Ziel erreicht werden. Daher bildet eine Entscheidung für ein Ziel mit einem abgeleiteten Zielsystem ist den Grundsatz einer jeden Planung. Zusammengefasst lässt sich konstatieren: Ohne Ziel kein Zielsystem, ohne Zielsystem keine Planung, ohne Planung keine Strategie. Ohne eine Strategie kann eine Unternehmung nicht existieren, weshalb der Entscheidungsfunktion eine große Bedeutung zukommt.

Koordinationsfunktion

„Ziele sind geeignet, Teilaktivitäten zu integrieren und auf eine Bezugsgröße, nämlich das Ziel, auszurichten“ [BH09, S. 72]. Durch die Planung lassen sich verschiedene Teile eines Unternehmens oder eines Unternehmensbereichs miteinander verzahnen. Dies geschieht beispielsweise mit dem Führungsmodell *Management by Objectives*[2]. Das gemeinsame Verfolgen eines Ziels richtet das Unternehmen als ganzes aus und ermöglicht die intrabetriebliche Zusammenarbeit (vgl. [BH09, S. 72]).

Informationsfunktion

Mitarbeiter und Stakeholder müssen über künftige Aktivitäten einer Unternehmung informiert werden. Diese Aufgabe wird durch die aus der Unternehmensplanung erwachsenen Ziele [BH09, S. 73] ursächlich abgedeckt und über unterschiedliche Informationskanäle weiter vermittelt. Insbesondere Investoren und Analysten sind unter den externen Informationsempfängern am bedeutendsten [BH09]. In großen Konzernen wird diese Funktion teilweise von eigenen Stabsstellen oder Abteilungen übernommen, die dann häufig unter dem Namen *Investor Relation* auftreten [BH09, S. 73].

[1] Es gilt zu beachten, dass die unterschiedlichen Funktionen je nach Quelle unterschiedlich abgegrenzt werden. Die hier erläuterten Funktionen richten sich nach Bea und Haas [BH09, S. 72].

[2] deutsch: *Führung durch Zielvereinbarungen*

Motivationsfunktion

Durch das Setzten von Zielen und das Kommunizieren von Absichten werden bei den Mitarbeitern einer Unternehmensentität[3] Anreize geschaffen diese zu erreichen [BH09, S. 73]. Auf der einen Seite entsteht dadurch eine zunehmende extrinsische Motivation im Bezug auf die gesetzte Zielgröße, weil diese erreicht werden soll [BH09, S. 73]. Auf der anderen Seite kann dies die intrisische Motivation der Mitarbeiter steigern, wenn es konkrete Ziele gibt, die sie erreichen können und auch wollen [BH09, S. 73]. Idealerweise kommt es bei den Mitarbeitern zu einer hohen Arbeitsmotivation im Zusammenhang mit der Koordinationsfunktion in einer besseren Zusammenarbeit im Unternehmen als Ganzem resultiert [BH09, S. 72].

Kontrollfunktion

Die Planung ermöglicht es Vergleichsmaßstäbe für die Analyse einer Entwicklung aufzustellen [Sch12, S. 24]. Daran können einzelne Instanzen eines Unternehmens gemessen werden und der Erfolg wird abbildbar (Performance Measurement). Vor allem werden damit auch Misserfolge sichtbar. „Dieser Effekt hält Manager nicht selten davon ab, Ziele konkret zu formulieren, um so der Gefahr des Versagens zu entgehen“ [BH09, S. 73].

Legitimationsfunktion

Mit der Unternehmensplanung lassen sich im Nachhinein oder während der Ausführung kritischer Prozesse Entscheidungen begründen, die somit die ausführenden Instanzen einer Unternehmung in ihrer Verantwortung entlasten. Dabei wirkt sich die Unternehmensumwelt direkt auf die Ziele aus, damit diese andere Entscheidungen legitimieren. Darunter fallen häufig Ziele wie der *Erhalt von Arbeitsplätzen* oder die *Verbesserung der Umweltverträglichkeit von Produkten und Verfahren* genannt [BH09, S. 73].

[3] Mit Unternehmensentität sind hier das Unternehmen als System selbst und alle sich darin befindlichen Subsysteme gemeint. Beispiele sind Unternehmensbereiche, Abteilungen, Projekte, Teams und Mitarbeiter.

4 Herausforderungen

„Die Unternehmensplanung stellt sich als komplexes und auch kompliziertes Geschehen dar, das nicht ohne die Beachtung wichtiger Grundsätze ablaufen darf. [...] Die folgenden Planungsgrundsätze können als die wichtigsten angesehen werden:“ [Ehr06]

- Langfristigkeit der Planung
- Vollständigkeit der Planung
- Anpassungsfähigkeit der Planung
- Stabilität der Planung
- Verbindlichkeit der Planung
- Kontrollierbarkeit der Planung
- Realisierbarkeit der Planungsvorgaben

Aus diesen Grundsätzen resultieren verschiedene Herausforderungen, die es bei einer ganzheitlichen Unternehmensplanung zu beachten gilt. So können unterschiedliche Grundsätze in Konflikt miteinander stehen. Beispielsweise kann die Anpassungsfähigkeit der Planung nicht immer mit der Stabilität vollständig kohärent sein. Genauso kann das Streben nach einer möglichst vollständigen Planung die langfristigen Aspekte beeinträchtigen. Es sind insbesondere die Grenzen der kognitiven Leistungsfähigkeit der unterschiedlichen Planungsbeteiligten (siehe Kapitel 4.3) zu nennen. Die Planungsbeteiligten können dabei jeweils andere Ziele verfolgen, die ggf. nicht mit den Zielen der Unternehmens-Shareholder übereinstimmen. Dies soll im Kapitel 4.2 näher ausgeführt werden. Eine weitere Frage werfen die Schlussfolgerungen der Planung mit den daraus abgeleiteten Gesetzmäßigkeiten auf. Dieser Problematik widmet sich Kapitel 4.4. Die bei weitem zentralste aller Herausforderungen der unternehmerischen Planung stellt die dynamische Umwelt und das dynamische Verhalten des Unternehmens dar, welche im folgenden Unterkapitel erläutert werden soll.

4.1 Dynamik der Umwelt

„Die Liberalisierung und Öffnung nationaler Märkte, der technologische Fortschritt und nachhaltige Veränderungen der Anforderungen geschäftlicher sowie der Lebensgewohnheiten und Verhaltensmuster privater Kunden haben mit hoher Dynamik neue Märkte entstehen lassen“ [BH09, S. 1]. Der Wandel in der Gesellschaft und die damit verbundenen Anforderungen stellen Unternehmen vor enorme Herausforderungen. Der nachhaltige Erfolg

eines Unternehmens ist deshalb auch davon abhängig, diese Herausforderungen zu meistern. Grundsätzlich ist es dabei wichtig, die Umweltsituation einer Unternehmung richtig zu analysieren und dies in die Planung mit einzubeziehen. Dies kann anhand von verschiedenen Betrachtungsgegenständen geschehen, die bewertet werden müssen. Die *Konjunkturabhängigkeit*[4] eines Unternehmens spielt hierbei eine besondere Rolle. Die oft unvorhersehbaren Entwicklungen der Branchen-Konjunktur müssen sofern möglich eingeplant werden. Um dies in umfangreicher Weise möglich zu machen, sollten auch unterschiedliche Indikatoren eines Wirtschaftsumschwungs als Eingangsparameter für die Prognose, auf deren Grundlage eine Planung ausgeführt wird, in Betracht gezogen werden. Auch sind die Auswirkungen der *Inflation* auf ein unternehmerisches Gebilde nicht zu vernachlässigen. Es sollten die sich verändernden Preisniveaus bei den Eingangsgütern sowie den Ausgangsgütern eingeplant werden, was ein großes Wissen über die Wertschöpfungsketten voraussetzt. Allerdings ist dieses Wissen gerade in kleineren, weniger einflussreichen Unternehmen selten gegeben [WMA04]. Die *Auswirkung der zunehmenden Schadstoffbelastung* [Ehr06, S. 176] sind eine nicht unwesentlicher Einflussgröße, die aber nur schwer gesteuert werden kann, da diese Problematik mit der gesamten Supply Chain in Verbindung steht.

Porter hat schon früh den Zusammenhang zwischen Umwelteinflüssen und dem Handeln eines Unternehmens identifiziert. Nach Porter [Por14, S. 25] müssen folgende „5-Kräfte“ in der Planung in Abhängigkeit zur Wettbewerbsposition in Betracht gezogen werden:

- Die Verhandlungsstärke der Lieferanten,
- die Verhandlungsmacht der Abnehmer,
- die Bedrohung durch Konkurrenten,
- die Bedrohung durch neue Konkurrenten und
- die Bedrohung durch Ersatzprodukte und –dienste.

Wichtig ist, diese unterschiedlichen *Kräfte* – sofern möglich – abzuschätzen und in die Planung mit einzubeziehen. Die Wettbewerber möchten aus strategischen Gründen die benötigten Informationen zur vollständigen Ermittlung der Bedrohungen logischerweise nicht preisgeben. Die Verhandlungsstärke der Lieferanten und die Verhandlungsmacht der Abnehmer lassen sich durch eine nachhaltige Unternehmensstrategie beeinflussen. Das Verhalten der Konkurrenten ist hingegen nicht steuerbar. Eine große Bedrohung stellen insbesondere die neuen Konkurrenten am Markt dar, die im Bezug auf Finanzstärke und

[4] Die Konjunktur ist die „mehr oder weniger regelmäßige Schwankung [...] des Auslastungsgrads des gesamtwirtschaftlichen Produktionspotenzials“ [lex13, S. 246] oder auch einer Branche.

Innovationskraft nur schwer einschätzbar sind. Das Entstehen von Ersatzprodukten oder –diensten ist ebenfalls nur schwer antizipierbar. Unvorhergesehene oder gar falsch eingeschätzte externe Innovationen haben das Potential leicht zu einem einschneidenden negativen Unternehmensereignis zu werden, weil sie aufgrund ihrer Unvorhersehbarkeit nicht in die Planung einbezogen werden können. In der Konsequenz können die vorhergegangenen unternehmensinternen Innovationen dann von anderen Wettbewerbern und ihren Produkten verdrängt werden.

Beispielhaft ist an dieser Stelle Kodak zu nennen. Kodak hat die Entwicklung hin zur digitalen Fotografie nicht früh genug als potentielles Geschäftsfeld erkannt und ist dadurch in große finanzielle Probleme geraten [Mat12].

4.2 Interessenkonflikte

Bei der Planung von Unternehmensstrategien und –prozessen sind verschiedene Personengruppen beteiligt. Wie schon in Abbildung 2 (Seite 3) zu sehen, wird die Unternehmensplanung mit unterschiedlichen Zeithorizonten bzw. Zeitdimensionen durchgeführt. In großen Unternehmen lassen sich diese ungleichen Zeithorizonte, allein auf Grund des Umfangs einer möglichst vollständigen Planung, nicht von einer Person alleine bewältigen. An der Unternehmensplanung sind also mehrere Individuen beteiligt. Dabei treten auf unterschiedlichen Ebenen eines Unternehmens und darüber hinaus Informationsasymmetrien auf, da beispielsweise der Anteilseigner[5] nicht alle Informationen über die internen Prozesse im Unternehmen besitzt, über die ein Manager entscheidet. Manager und Shareholder haben dabei grundsätzlich unterschiedliche Ziele, wie Alfred Rappaport 1986 in seinem Buch *Creating Shareholder Value. The New Standard for Business Performance* [Rap86] erstmals konstatierte. Seine zentrale Aussage darin lautet: „Critics of large corporations often allege that corporate managers have too much power and that they act on ways to benefit themselves at the expense of shareholders and other corporate constituencies“ [Rap86, S. 6]. Es findet also eine konsequente Benachteiligung der Anteilseigner statt. Damit einhergehend verschlechtern sich nachhaltig die Finanzierungsmöglichkeiten eines Unternehmens am Markt. Um diesem Problem zu begegnen hat sich vornehmlich in großen Aktiengesellschaften das sogenannte „Wertorientierte Management“ durchgesetzt, also die von Rappaport geforderte Orientierung des Managements am Shareholder Value [BH09, S. 82]. Der Ansatz nach dem Shareholder Value, mit dem eine Planung durch alle Unternehmensebenen hindurch gefördert werden soll, ist dabei eine Lösung im Sinne der Prinzipal-Agent-Theorie. Diese „untersucht Wirtschaftsbeziehungen, in denen

[5] engl. shareholder

ein Geschäftspartner [oder Mitarbeiter] Informationsvorsprünge gegenüber den anderen aufweist. Diese Informationsasymmetrien bewirken Ineffizienzen bei der Vertragsbildung oder Vertragsdurchführung“ [Erl14], welche sich auf allen Zeithorizonten negativ auf das nachhaltige Ziel einer Unternehmung und die daraus resultierenden Planungsdimension auf allen Zeithorizonten auswirken können.

4.3 Begrenzte kognitive Fähigkeiten

Wie in Kapitel 3.1 beschrieben, ist eine Unternehmensplanung ohne eine adäquate Prognose nicht möglich. Jede Prognose weißt hierbei Unsicherheiten auf, die aufgrund von Beschränkungen des relevanten Wissens und den wissensgenerierenden Fähigkeiten der Menschen entstehen [Pie13, S. 138]. Die Beschränkung des Wissens in Bezug auf einen Prognosegegenstand lässt sich nur schwer kompensieren, da sie eine Folge der Beschränkungen der wissensgenerierenden Fähigkeiten darstellt. Wäre das Generieren von Wissen gänzlich unbeschränkt, so könnte es auch keine Beschränkungen hinsichtlich des relevante Wissens geben, denn alles nötige Wissen könnte generiert werden. Die Unsicherheiten der Prognose rühren also ursächlich aus den mangelnden kognitiven Fähigkeiten der Wissensgenerierung im Bezug „auf das vorhandene Fakten-, Methoden- und Beziehungswissen de[r] Akteur[e]“ [Pie13, S. 138]. „Akteure sind aufgrund ihrer kognitiven Begrenztheit [...] darauf angewiesen, Situationen mit Hilfe von Heuristiken zu beurteilen. Heuristiken stellen vereinfachende kognitive Prozesse dar, die der möglichst schnellen Lösung komplexer Probleme dienen“ [Pie13, S. 138]. Heuristiken weisen allerdings nicht zu kompensierende Fehler bzw. kognitive Verzerrungen auf, die wiederum zu Fehlern in der Prognose und damit auch in der Planung führen. Eine Heuristik, die häufig zu Fehlern führt, ist beispielsweise die Verfügbarkeitsheuristik, die „besagt, dass Menschen Wahrscheinlichkeitsurteile eher auf der Basis der ihnen zur Verfügung stehenden Informationen als auf der Basis aller für eine optimale Entscheidung notwendigen Informationen fällen. [...] Je leichter die Erinnerung an ein Ereignis fällt, desto größer wird seine Wahrscheinlichkeit eingeschätzt“ [Pie13, S. 141]. Der kognitive Prozess der Prognose unterteilt sich in zwei zentrale Schritte: Die Informationsaufnahme und die Informationsverarbeitung [Pie13, S. 144]. Die Aufnahme der Information kann nach dem erweiterten Linsenmodell von Steward und Lusk [SL94, S. 584] niemals fehlerfrei sein. Genauso ist auch die Verarbeitung aufgrund der darin verwendeten Heuristiken – wie zuvor beschrieben – nicht fehlerfrei. Die Planung auf Basis der Prognose muss mit diesen Herausforderungen umgehen, kann sie aber nicht lösen.

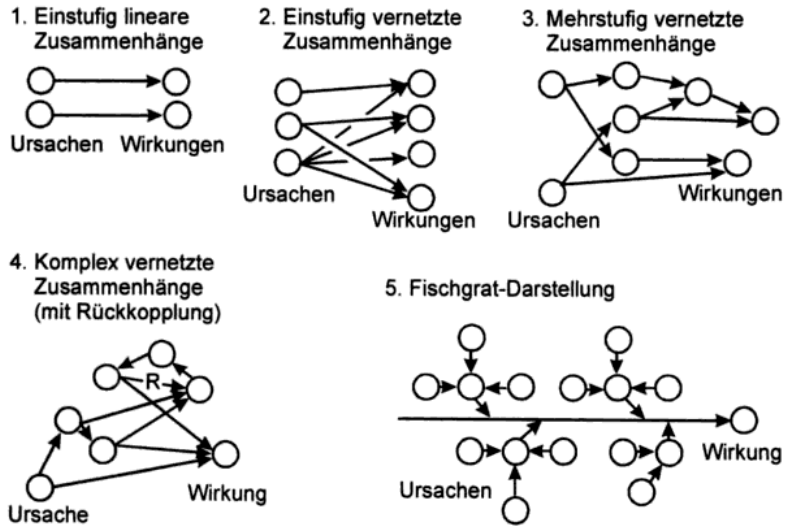


Abbildung 4: Denkansätze zu Ursache-Wirkungs-Zusammenhängen [HJ98, S. 93]

4.4 Gesetzmäßigkeiten des unternehmerischen Handelns

„Das Denken in linearen Ursache-Wirkungs-Zusammenhängen ist im abendländischen Kulturkeis die vorherrschende Denkweise“ [HJ98, S. 93]. Da die Planung eine Ordnung entwirft nach der sich das betriebliche Geschehen in der Zukunft vollziehen soll [Gut83, S. 148], müssen Wechselwirkungen von Ursachen und Wirkungen der vergangenen, heutigen und zukünftigen Entscheidungen miteinander verknüpft werden. Basierend auf ihrer gegebenen Komplexität stellt die Verknüpfung eine große Herausforderung dar, da es unbekannte Einflussfaktoren gibt, die bei der Planung nicht berücksichtigt werden können. Die Abhängigkeiten von Ursache und Wirkung können – wie in Abbildung 4 zu sehen – in fünf verschiedene Zusammenhangsklassen eingeordnet werden.

Bei einstufigen, linearen Zusammenhängen folgt aus einer Ursache auch nur eine Wirkung. Die einstufig vernetzten Zusammenhänge lassen aus unterschiedlichen Ursachen jeweils mögliche Wirkungen erwachsen, die auch gleich sein können. Ursachen mit mehrstufig vernetzten Zusammenhängen führen zu verschiedenen, aber auch gleichen Wirkungen, die wiederum weitere Folgen haben, die sich auch gegenseitig überschneiden können. Komplex vernetzte Zusammenhänge mit Rückkoppelungen berücksichtigen darüber hinaus auch die sich gegenseitig verstärkenden Auswirkungen einer Ursache. Der Denkansatz, der der Fischgrat-Darstellung zugrunde liegt, betrachtet nicht die unterschiedlichen Wirkungen, sondern primär eine Auswirkung und bringt die dazu führenden Ursachen in Beziehung.

Allein diese schematische Darstellung verschiedener Ursachen mit dazu in Beziehung

stehenden Auswirkungen zeigt, wie komplex und kompliziert das Planen einfacher Ziele mit unterschiedlichen Wechselwirkungen zueinander ist. Daraus resultiert, dass es sehr herausfordernd ist Gesetzmäßigkeiten zu finden, die die Planung unterstützen können. Eine Option stellen umfangreiche Software-Lösungen dar, die Informationen in einem größeren Umfang verarbeiten könnten als der Mensch. Somit lassen sich mehr Rahmenbedienungen in der Planung berücksichtigen. Allerdings ist es unmöglich alle relevanten Rahmenbedingungen zu erfassen, weil kein Informationssystem alle Informationen der Welt erfassen und verarbeiten kann. Daher bleibt fraglich, ob das vollständige Identifizieren von umfangreichen Gesetzmäßigkeiten für die Planung überhaupt möglich ist.

5 Fazit

Die vorangegangenen Ausführungen haben gezeigt, dass es für ein Unternehmen ohne Unternehmensplanung schwierig ist nachhaltig erfolgreich zu sein. Fehlen stetige Planungsprozesse ist die dauerhafte Handlungsfähigkeit des Unternehmens in Gefahr, da zentrale Voraussetzungen für Entscheidungen, Koordination, Motivation, Informationen und Kontrollen nicht existieren. Auch lässt sich das Agieren des Unternehmens gegenüber der Umwelt und den einzelnen Unternehmensentitäten kaum rechtfertigen. Die Planung erwächst in diesem Kontext aus der Prognose und weist starke Interdependenzen mit dieser auf: Die Planung kann ohne eine Prognose nicht durchgeführt werden. Genauso ist auch die Prognose ohne eine Planung zwecklos.

Es hat sich gezeigt, dass eine gute Unternehmensplanung zahlreichen Herausforderungen gegenübersteht. Insbesondere das mangelnde Wissen über verschiedene Abhängigkeitsbeziehungen und die daraus erwachsenden Gesetzmäßigkeiten sind nur mithilfe von Heuristiken zu bewältigen. Dies stellt aber lediglich einen Behelf dar, der kein optimales Ergebnis liefert. Da kein vollständiges Wissen vorhanden sein kann, ist eine perfekte Unternehmensplanung ausgeschlossen. Auch die zunehmende Dynamik der Umwelt bringt neue Herausforderungen mit sich, denn während „auf der einen Seite Unternehmen durch die zunehmende Komplexität eine längere Reaktionszeit benötigen, um sich angemessen auf eine Vielzahl an exogenen Veränderungen einzustellen, sinkt auf der anderen Seite die für sie verfügbare Reaktionszeit auf Grund steigender Dynamik“ [KM11, S. 428].

Zukünftig sind umfangreiche Trendanalysen in der Unternehmensplanung denkbar, um diese weiter zu verbessern. Die systematische Analyse des Unternehmensumfelds sollte hierbei insbesondere einen weiter wachsenden Stellenwert einnehmen, damit die Planungsprozesse zuverlässiger im Hinblick auf dynamische Umweltprozesse sind. Damit wird die Unternehmenszukunft besser vorhersehbar und geplante Investitionen können gewinnbrin-

gender eingesetzt werden. Abzuwarten bleibt, wie sich Marktteilnehmer mit unterschiedlicher Finanzkraft in diesem Planungsfeld voneinander unterscheiden werden. Vor allem das eingesetzte Know-How in den Unternehmen wird hierbei wahrscheinlich maßgeblich über den Erfolg einer Planung entscheiden.

Abschließend bleibt zu vermuten, dass ein Unternehmen ab einer bestimmten Größe die Unternehmensplanung nicht nachhaltig ohne wissensgestützte IT-Systeme durchführen kann. Die Unternehmen werden voraussichtlich Kapital daraus schlagen, wenn sie die Entwicklungen in diesem Bereich weiter ernst nehmen und frühzeitig über den Einsatz entsprechender IT-Systeme zur Planungsunterstützung entscheiden.

Literatur

[Bea11] Bea, Franz Xaver: *Allgemeine Betriebswirtschaftslehre: Bd. 2: Führung*. Lucius & Lucius, Stuttgart, 10., überarb. und erw. Aufl. Auflage, 2011.

[BH09] Bea, Franz Xaver und Jürgen Haas: *Strategisches Management*. Grundwissen der Ökonomik. UTB GmbH, 5., neu bearbeitete Auflage, 2009.

[DH07] Drews, Günter und Norbert Hillebrand: *Lexikon der Projektmanagement Methoden*. Haufe Verlag, München, 1. Auflage, 2007.

[Ehr06] Ehrmann, Harald: *Unternehmensplanung*. Kompendium der praktischen Betriebswirtschaft. Kiehl, 6., überarbeitete und aktualisierte Auflage, 2006.

[Erl14] Erlei, Mathias: *Prinzipal-Agent-Theorie*. Gabler Wirtschaftslexikon, 2014. Online unter `http://wirtschaftslexikon.gabler.de/Archiv/924/prinzipal-agent-theorie-v9.html`. Letzter Zugriff am 02.08.2014.

[Gut83] Gutenberg, Erich: *1. Die Produktion*. Springer, Berlin, 24., unveränd. Auflage, 1983.

[Ham98] Hammer, Richard: *Unternehmensplanung. Lehrbuch der Planung und strategischen Unternehmensführung.* München, 7. Auflage, 1998.

[HJ98] Hübner, Heinz und Stefan Jahnes: *Management-Technologie als strategischer Erfolgsfaktor: ein Kompendium von Instrumenten für Innovations-, Technologie- und Unternehmensplanung unter Berücksichtigung ökologischer Anforderungen*. de Gruyter, 1998.

[KM11] Klatt, Tobias und Klaus Möller: *Entscheidungsanomalien in der strategischen Unternehmensplanung*. Zeitschrift für Management, 6(4):427–449, 2011.

[lex13] *Gabler Kompakt-Lexikon Wirtschaft*. Springer, Wiesbaden, 11., akt. Auflage, 2013.

[Mai14] Maier, Günter W.: *Delphi-Technik*, 2014. Online unter `http://wirtschaftslexikon.gabler.de/Archiv/3268/delphi-technik-v8.html`. Letzter Zugriff am 02.08.2014.

[Mat12] Mattioli, Dana: *Letzte Blende: Kodak gibt das Geschäft mit Kameras auf*, 2012. Online unter `http://www.wsj.de/article/`

SB10001424052970203824904577213813232468818.html?mg=reno64-wsjde. Letzter Zugriff am 02.08.2014.

[MR12] Mertens, Peter und Susanne Rässler: *Prognoserechnung*. Physica-Verlag, Heidelberg, Siebte, wesentlich überarbeitete und erweiterte Auflage, 2012.

[Pie13] Pieroth, Guido: *Systematische Prognosefehler in der Unternehmensplanung, Eine ökonomisch-psychologische Analyse*, Band 47 der Reihe *Schriften des Center for Controlling & Management (CCM)*. Springer Gabler, Wiesbaden, 2013.

[Por14] Porter, Michael E.: *Wettbewerbsvorteile: Spitzenleistungen erreichen und behaupten. (Competitive advantage)*. Campus-Verag, Frankfurt, 8. Auflage, 2014.

[Rap86] Rappaport, Alfred: *Creating Shareholder Value. The New Standard for Business Performance*. Free Press, New York, London, 1986.

[Sch12] Schön, Dietmar: *Planung und Reporting im Mittelstand*. Springer, Heidelberg, 2012.

[SL94] Stewart, Thomas R. und Cynthia M. Lusk: *Seven Components of Judgmental Forecasting Skill: Implications for Research and the Improvement of Forecasts*. Journal of Forecasting, 13:579 – 599, Februar 1994.

[WMA04] Wöelffling, Peter, Horst Miethe und Peter Albrecht: *Wissensmanagement im Kontext von Unternehmenskultur und Wertschöpfung*, 2004. Online unter http://www.bibb.de/veroeffentlichungen/en/publication/download/id/888. Letzter Zugriff am 02.08.2014.

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Integrierte Unternehmensplanung

Seminararbeit
im Rahmen der Projektgruppe „inMemory Planung mit SAP HANA“

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dipl.-Inform. Dirk Peters
Vorgelegt von:	Farhad Murad Gavan El-Yazdin Ahlhorner Straße 30 27793 Wildeshausen 0173-3565137 farhad.gavan@uni-oldenburg.de
Abgabetermin:	02. August 2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

Tabellenverzeichnis **3**

1 Einleitung **4**

2 Grundlagen der klassischen Unternehmensplanung **4**
2.1 Definition der Unternehmensplanung 5
2.2 Ziele der Unternehmensplanung 6

3 Die integrierte Unternehmensplanung **7**
3.1 Isolierte Teilplanungen 8
3.1.1 Absatzplanung 8
3.1.2 Werbeplanung 9
3.1.3 Fertigungs- und Produktionsplanung 10
3.1.4 Beschaffungsplanung 10
3.1.5 Finanzierungsplanung 10
3.2 Integrierte Unternehmensplanung 12
3.2.1 Strategische Planung: 13
3.2.2 Taktische oder Bereichsplanung: 13
3.2.3 Operative Planung: 14
3.3 Werkzeuge in integrierte Unternehmenssteuerung 15
3.4 Vor-Nachteile einer integrierten Unternehmensplanung 17

4 Zusammenfassung **18**

Literaturverzeichnis **19**

Abbildungsverzeichnis

1 Unternehmensplanung(UPL) im Mittelpunkt des unternehmerischen Führungssystems[Fis96], S. 7 6
2 Von der klassischen zur strategischen UPL([Fis96], S. 8) 8
3 Werbetätigkeiten im Unternehmen([Fis96], S. 80) 9
4 Einflussfaktoren der Beschaffungsplanung([Rol01], S. 52) 10
5 Aufbau eines Finanzplans([Fis96], S. 139) 12
6 Strategische Planung ([Fis96], S. 40) 13
7 Taktische oder Bereichsplanung([Fis96], S. 41) 14
8 Operative Planung (Durchführungsplanung)[Fis96], S. 42 15
9 Top 10 Umsatz BI in Deutschland ([GB14]) 16

Tabellenverzeichnis

1 SAPs Planungslösungen im Vergleich([Com14]) 17

1 Einleitung

Die Unternehmensplanung ist der Vorgang der Planung in Wirtschaftsbetrieben. Wenn von einer integrierten Unternehmensplanung gesprochen wird, dann ist damit ein sehr umfassender, fortwährender Planungsprozess mit einer Vielzahl von Planungsverfahren und Planungsschritten gemeint.

Integrieren heißt Zusammenfügen zu einem übergeordneten Ganzen. Alle Teilplanungen zusammen bilden den Gesamtplan (Unternehmensplan), der jedoch – entsprechend dem Interdependenzgefüge des Unternehmensgeschehens – nicht als eine Addition von Teilplanungen, sondern als geordnetes System von aufeinander bezogenen Teilplanungen zu verstehen ist. Integration von Teilplanungen bedeutet daher Bildung von Planungssystemen durch methodisches und systematisches Zusammenfügen und Verknüpfen von Teilplanungen.

Die integrierte Unternehmensplanung gewinnt in Unternehmen immer mehr an Bedeutung, da die Unternehmen aufgrund steigenden internationalen Wettbewerbsdrucks und wachsender Globalisierung in der Lage sein müssen, schnell und flexibel zu reagieren bzw. zu handeln.

Fischer erklärt sich den Ursprung der Planung wie folgt:

„Ihr Sprung liegt wohl in der militärischen Planung von Kriegen und Eroberungszügen mit der notwendigen Versorgung der Soldaten mit Waffen, Wagen, Lebensmittel und Hilfsscharen. Schnell entstanden auch dadurch Finanzprobleme, was sich in „Finanzierungsplänen“ niederschlug und in der Summe Staatshaushalte entstehen ließ“([Fis96], S.1).

Daraus ist zu ersehen, dass der Zeitpunkt zur Entstehung der Planung unbekannt ist. Schon im Altertum gab es Planungsvorgänge, die von großer Bedeutung waren.

Meine Seminararbeit soll die Wichtigkeit der Planung im Unternehmen verständlich und realisierbar zeigen sowie die Notwendigkeit der BI in einer integrierten Unternehmensplanung darstellen. Um das Ziel zu erreichen, ist meine Seminararbeit ist in zwei Teilen gegliedert. In dem ersten Teil wird die Grundlage der klassischen Unternehmensplanung wie Definition und Ziele detailliert beschrieben. Im zweiten Teil wird spezifisch die integrierte Unternehmensplanung mit ihren Teilbereichen und den Werkzeugen dargestellt. Anschließend folgt eine kurze Zusammenfassung.

2 Grundlagen der klassischen Unternehmensplanung

Um Unternehmensplanung zu verstehen, muss zuerst der Begriff „Planung“ wie folgt definiert werden:

„Planung ist der Versuch, die Zukunft des Unternehmens im Voraus zu gestalten und damit in der gewünschten Zielrichtung die Entwicklung des Unternehmens zu beeinflussen“([Fis96], S. 4).

Damit ist es das zentrale Führungsinstrument für alle Tätigkeiten des unternehmerischen Tuns. Sie umfasst in der Erstellung und Wartung eines Plans eine grundlegende Eigenschaft intelligenten Verhaltens. Dieses Verhalten ist ein Denkprozess mit Fokus auf Erstellung und Verfeinerung eines Plans. Dadurch entsteht die Integration der Pläne untereinander. Diese verbinden Prognosen von Entwicklungen mit der Erstellung von Szenarien. Die Prognose ist ein wichtiger Aspekt in der Planung, der oft durchaus ignoriert wird. Prognose wird als Vorhersage der Zukunft beschrieben, d.h. es wird gezeigt, wie die Zukunft aussehen wird. Die Planung wiederum prognostiziert, wie die Zukunft aussehen soll.

Ehrmanns kennzeichnet die effektive Planung durch Sachlichkeit, Kompetenz, Kreativität, Problemlösungsorientierung und Zukunftsorientierung(Vgl.[Ehr13], S.23).
Es ist ein vereinfachtes, symbolisches Modell zukünftiger realer Systeme. Mit Plänen soll ein Kommittent geschaffen werden, welches innerhalb einer angegebenen Zeit bestimmte Systemzustände zu erreichen versucht(Vgl.[MS14]). Diese bauen Analysen des Istzustandes, der Zukunft des Unternehmens und der Umwelt auf. Sie verarbeitet eine Vielzahl von Informationen über Fakten, Entwicklungen, Trends und Verhaltensweisen und trägt unter Berücksichtigung der Ungewissheit des Eintritts von Ereignissen und der Realisierung von Annahmen zur Kursfixierung bei.

2.1 Definition der Unternehmensplanung

Unternehmensplanung ist Zielorientiert und Zukunftsbezogen und sollte die betriebliche Realität möglichst ganzheitlich erfassen. Diese werden den Wettbewerb und der Marktveränderung angepasst, die rechtzeitig erkannt werden und in der Unternehmensplanung berücksichtigt werden(Vgl.[Rol01],S. 5).

Durch das Lernen grundlegender Konzepte und Umsetzung können sowohl größere als auch kleinere Unternehmen einen umfassenden Plan erstellen, welches durch einen Prozess den Weg zur Umsatzwachstum und Gewinnsteigerung vereinfachen soll.
Die folgende Abbildung zeigt, dass die Unternehmensplanung das Kernstück eines Führungssystems sein kann. Voraussetzung für eine erfolgreiche Planung ist unteranderem die Unternehmensphilosophie. Unternehmensphilosophie ist eine einheitliche Grundaussage, ein einheitliches Konzept, das die Unternehmenspolitik beeinflusst und den Übergang von der strategischen Planung zur strategischen Führung ermöglicht(Vgl.[Fis96], S. 3).

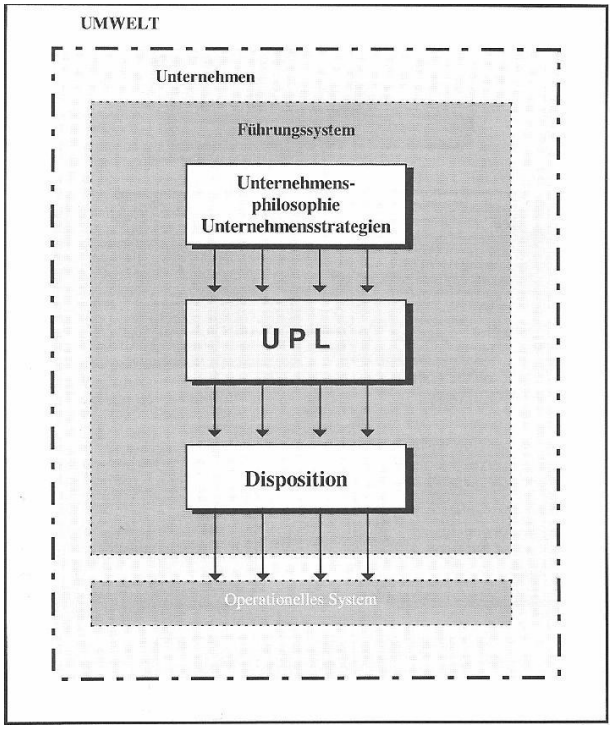


Abbildung 1: Unternehmensplanung(UPL) im Mittelpunkt des unternehmerischen Führungssystems[Fis96], S. 7

Koch sieht die Planung im Rahmen der Unternehmenspolitik als dominierend an. Er begründet es damit, dass selbst die Unternehmenspolitik Gegenstand der Planung ist, wodurch ihr Dynamik und Durchsetzungsvermögen verliehen werden(Vgl.[Koc82],S.7).

2.2 Ziele der Unternehmensplanung

Die integrierte Unternehmensplanung hat das Ziel, die Unternehmenszukunft zu sichern, sowie zur Gewährleistung der Erreichung der Unternehmensziele und zu Optimierung der Wirtschaftlichkeit beizutragen. Die Ziele des Unternehmens werden gesucht, analysiert und auf ihre Realitätsnähe überprüft, dadurch entsteht unter anderem das Ziel der Ordnungsfunktion im Unternehmen. Die erkannten Ziele werden entwickelt und variiert. Entstandene Konflikte im Unternehmen werden durch eine integrierte Unternehmensplanung schnell erkannt. Diese Konflikte werden so schnell wie möglich beseitigt und strukturiert. Prioritäten werden gesetzt, um zu erkennen und festzulegen, welches die wichtigsten Aufgaben des Unternehmens sind und diese vorrangig im Unternehmen zu behandeln.
Das Betriebsgeschehen wird durch die Abstimmung der Ziele und Maßnahmen der verschiedenen Unternehmensbereiche koordiniert. Durch die Koordinationsfunktion können gleichzeitig Synergieeffekte und Effizienzsteigerungen erzielt werden. Die integrierte Unternehmungsplanung ermöglicht durch Untersuchung des Soll-Ist-Vergleichs das Erkennen deutlicher Abweichungen und durch Untersuchung der Abweichungen deren Ursachen und schafft somit eine Kontrollmöglichkeit der Plannungsanpassung. Ein weiteres wichtiges Ziel ist die Bemühung um die Arbeitskräfte und das richtige Einsetzen der Arbeitskräfte im passenden Bereich. Dadurch

kann die Arbeitskraft ihre Fähigkeiten und Kenntnisse adäquat einsetzen, sich weiter entwickeln und die Motivation zur Rationalisierungsmöglichkeiten erhöhen. Das Resultat ist die Umsatzsteigerung des Unternehmens und die Wettbewerbsfähigkeit im Markt(Vgl.[MS14]).

3 Die integrierte Unternehmensplanung

Wie in den Grundlagen erwähnt, beinhaltet die integrierte Unternehmensplanung die Planung sämtlicher Unternehmensbereiche für einen Zeitraum unter gegenseitiger Abstimmung. Es wird für den Planungszeitraum ein Unternehmensgesamtplan aufgestellt. Damit geht die integrierte Unternehmensplanung weit über die isolierten Teilplanungen hinaus(Vgl.[Koc82], S. 9).

In früheren Zeiten wurde eher in isolierteren Teilplänen geplant, d.h. nur in bestimmten Abteilungen des Unternehmens. Dadurch stellte sich heraus, dass diese isolierten Teilpläne insbesondere zu Kapitalbindungen führten, die in guten Zeiten nicht besonders schwer wogen. Als sich die Märkte veränderten und auch internationalisierten, die Mengen sanken und Umsätze zu stagnieren begannen, kam die Einführung einer strategischen und integrierten Planung oft zu spät. Konkurswellen waren die Folge.
Durch negative Erfahrungen wurde auch in deutschen Bereichen die Planung professionalisiert und teilweise aus U.S.A kopiert. Die Amerikaner waren den deutschen Unternehmen in puncto Strategie und Planung oft bis zu 5 Jahre voraus. So wurde in deutschen Unternehmen aus den klassisch isolierten Teilplanungen allmählich eine strategisch, integrierte Planung entwickelt(Vgl.[Fis96], S. 8).

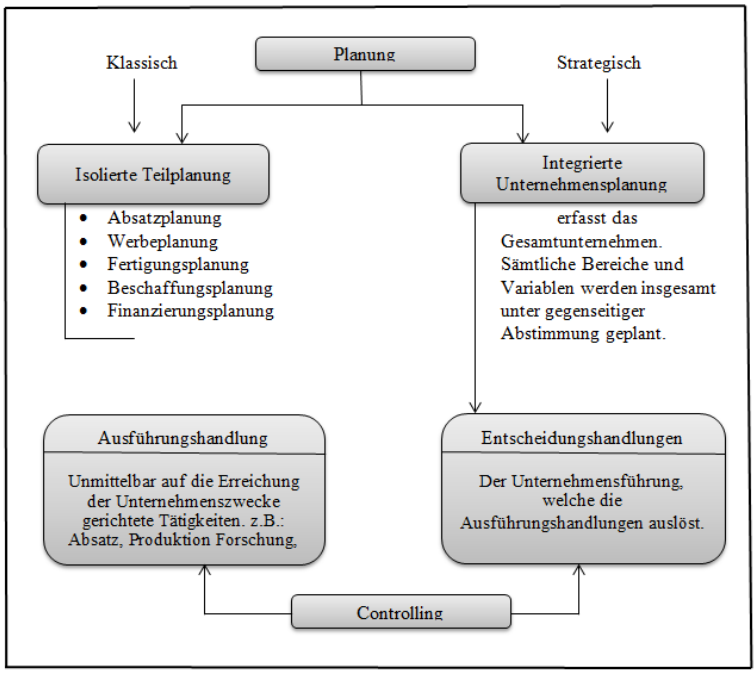


Abbildung 2: Von der klassischen zur strategischen UPL([Fis96], S. 8)

In Bezug auf das Unternehmen wird zwischen zwei Planungen unterteilt: der isolierten Teilplanung der verschiedenen Unternehmensbereiche im Unternehmen und der integrierten Unternehmensplanung des Unternehmens, die wie folgt beschrieben werden:

3.1 Isolierte Teilplanungen

Die isolierte Teilplanung beinhaltet einzelne Teilpläne des Unternehmensplans wie Absatzplanung, Werbeplanung, Fertigungs- und Produktionsplanung, Beschaffungsplanung und Finanzierungsplanung. Die einzelnen Teilpläne müssen aufeinander abgestimmt werden, um der gesamtbetrieblichen Zielsetzung gerecht zu werden.

3.1.1 Absatzplanung

Der Absatz beschreibt im wirtschaftlichen Sinne nichts anderes, als die Menge an Sach- und Dienstleistungen eines Unternehmens, die in einem bestimmten Zeitraum an den Verbrauchern erbrach werden und dadurch einen Gewinn erzielen. Es ist erforderlich, den Bedarf des Marktes an bestimmten materiellen bzw. immateriellen Gütern frühzeitig und zuverlässig abzuschätzen, um rechtzeitig bedarfsdeckende Maßnahmen einleiten zu können. Die Nachfrage nach Produkten hängt nicht allein von unbeeinflussbaren situativen Kontextfaktoren ab, sondern kann in gewissen Grenzen auch über entsprechende Handlungen, die die jeweiligen Unternehmen beeinflusst werden. Insofern ist es nicht nur möglich, bereits vorhandenen Bedarf zu decken, sondern auch latenten Bedarf zu wecken oder nach fehlendem Bedarf zu schaffen. Die Absatzplanung dient dem Unternehmen und den verantwortlichen Managern dazu, die Produktion so effizient wie nur möglich zu gestalten(Vgl.[Rol01], S. 33 f).

Ziel der Absatzplanung ist es, den Deckungsbeitrag des Unternehmens so weit wie möglich zu maximieren, wofür diese Planung den Grundstein legt(Vgl.[Edi14]).

3.1.2 Werbeplanung

Die Durchführung von Werbemaßnahmen ist sehr kostenintensiv. Aus diesem Grund sollten Werbemaßnahmen zielorientiert eingesetzt werden. Dem Unternehmen sollte klar sein, welches Ziel es damit erreichen will, welche Käufergruppe angesprochen werden soll, wann mit den Werbemaßnahmen begonnen werden kann, welchen Umfang die Werbemaßnahmen haben werden und welches Budget zur Verfügung steht(Vgl.[Sim14]).
Einige Werbemittel sind Werbespots, Anzeigen und Plakate. Für die Übermittlung dieser Werbemittel werden häufig Werbeträger wie Fernsehen und Rundfunk für Werbespots genutzt. Webseiten, Zeitungen, Zeitschriften für Anzeigen, Plakatwände und Litfaßsäulen stehen für Plakate zur Verfügung(Vgl.[Fis96], S. 79).
Die Werbung dient als Informationszweck für das Produkt bzw. die Produktgruppe. Durch den Aufbau eines Produkt- oder Markenimages soll das Vertrauen der Verbraucher gewonnen und das Nutzen des Produktes für den Nachfrager herausgestellt werden. Die Verbraucher werden auf das eigene Produkt bzw. die Produktgruppe aufmerksam und so entstehen unbekannte bzw. latente Bedürfnisse der Abnehmer.
In der Abb. 3 wird gezeigt wie der systematische Ablauf der Werbetätigkeit für das Unternehmen in einen strategischen, einen taktischen und einen operativen Bereich unter Einschluss des Werbe-Controlling unterteilt ist.

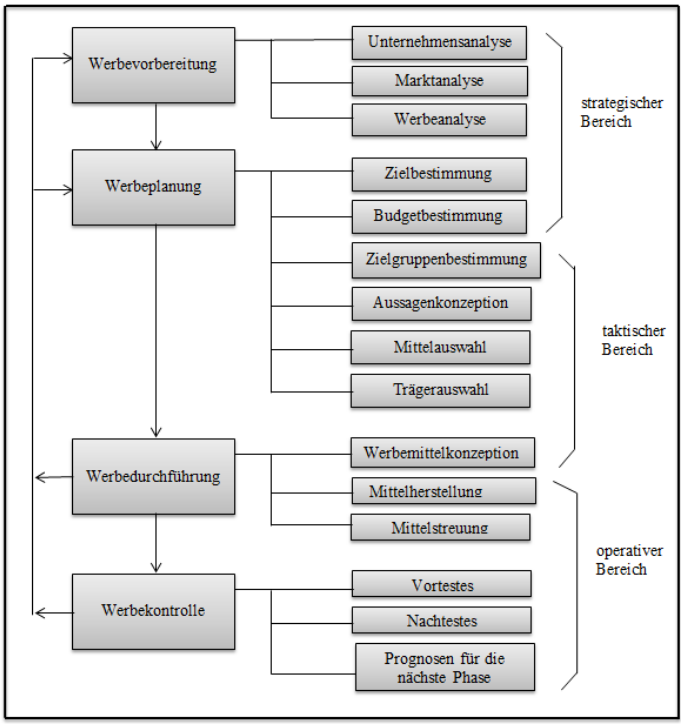


Abbildung 3: Werbetätigkeiten im Unternehmen([Fis96], S. 80)

3.1.3 Fertigungs- und Produktionsplanung

Die Fertigungsplanung ist ein Teilbereich der betrieblichen Gesamtplanung, der sich auf die zukünftige Gestaltung und den Ablauf der Fertigung und deren Verfahren bezieht. Sie umfasst die Fertigungsprogrammplanung, die Planung der Bereitstellung der Produktionsfaktoren und die Ablaufplanung(Vgl.[Wir14]).

Die Produktionsplanung ist das Kernstück eines jeden Herstellungsprozess. Sein Zweck ist es Zeit und Kosten zu minimieren, effizient zu organisieren und den Einsatz von Ressourcen und Effizienz am Arbeitsplatz zu maximieren.

Die Produktionsplanung umfasst eine Vielzahl von Produktionselemente, die von den alltäglichen Aktivitäten der Mitarbeiter auf die Fähigkeit abhängig sind z. B. genaue Lieferzeiten für den Kunden zu realisieren.

3.1.4 Beschaffungsplanung

Die Beschaffungsplanung umfasst die Deckung eines gegebenen Nettobedarfs an Repetierkosten. Dieser Bedarf wird im Anschluss an Bereitstellungsplanungsinformationen von beispielweise Lieferkonditionen, Qualität und Preis der zu beschaffenden Werk-, Hilfs- und Betriebsstoffe ermittelt. Hierzu wird auf die Beschaffungsmarktforschung zurückgegriffen um Informationen über den Zustandsraum, Aktionsraum und die Ergebnisfunktion des beschaffungsrelevanten Entscheidungsfeld zu gewinnen.
Durch die Beschaffungsplanung sollen die Beschaffungskosten optimiert, die Versorgungsrisiken vermindert, Verbesserung der Steuerung, Kontrolle der Beschaffungsdurchführung und Einhaltung des Qualitätsstandards wie in der Abb. 4 ersichtlich gehalten werden(Vgl.[Rol01], S. 50).

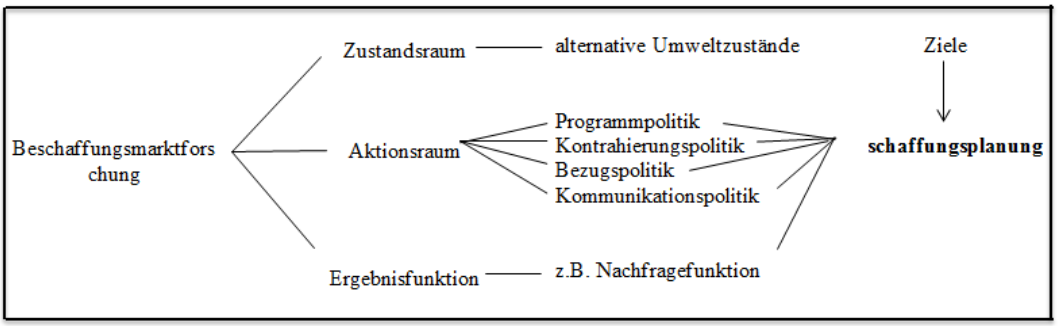


Abbildung 4: Einflussfaktoren der Beschaffungsplanung([Rol01], S. 52)

3.1.5 Finanzierungsplanung

Ohne die Finanzierung kann das Unternehmen nicht geführt werden. Deshalb ist es eine Hauptfunktion des Unternehmens. Mit der Finanzplanung wird der Kapitalbedarf ermittelt und seine Deckung durch die Auswahl geeigneter Finanzierungsformen bestimmt. Gleichzeitig muss jeder-

zeitige Zahlungsfähigkeit des Unternehmens gesichert werden(Vgl.[Fis96], S. 137 f).

Die Finanzplanung hat die Aufgaben, die optimale finanzwirtschaftliche Struktur der Planung zu sichern, Finanzprognosen aufzustellen (Schätzung der künftigen Ein- und Auszahlungen) Alternativen für die Mittelbeschaffung und Mittelanlage für die zu erwartenden Fehlbeträge bzw. Finanzüberschüsse, ständige Plankontrolle und Planrevision. Die wesentlichen Grundsätze der Finanzplanung sind folgende(Vgl.[Fis96], S. 137 f):

- Regelmäßigkeit: z. B. in finanzwirtschaftliche problematischen Situationen muss die Finanzplanung regelmäßig und dauernd vorgenommen werden und nicht nur teilweise.
- Die Vollständigkeit: hier müssen alle Zahlungsströme die innerhalb des Planungszeitraums liegen berücksichtigt werden. Nur so ist es möglich die Entwicklung der Liquidität abzuschätzen.
- Zeitpunktgenauigkeit: Der Zeitpunkt der einzelnen Geldströme ist so genau wie möglich anzugeben, um genaue Aussagen über die Entwicklung der Liquidität vornehmen zu können.
- Betragsgenauigkeit: Es sollen realistische Beträge für die Ausgaben und Einnahmen angesetzt werden.
- Bruttoausweis: Es dürfen keine Saldierungen von Zahlungsströmen vorgenommen werden, weil dadurch die Transparenz der Finanzplanung eingeschränkt wird.
- Elastizität: Bezieht sich auf eine erwartete bestimmte veränderbare Situationen der Unternehmens- und Umweltsituationen, d.h. die Plansätze sind unsicher und deshalb den Situationen angepasst werden können.
- Kontrollierbarkeit: Vergleich von Soll- und Ist-Werte um Abweichungen analysieren.
- Wirtschaftlichkeit: Der Aufwand muss in einem wirtschaftlichen vernünftigen Verhältnis zum Ergebnis und insbesondere der Genauigkeit stehen.

Abb. 5 zeigt, wie ein Finanzplan laut Fischer aufgebaut ist. Dieser Plan hängt von Absatzplan und dadurch auch von Produktions- und Materialplan ab.

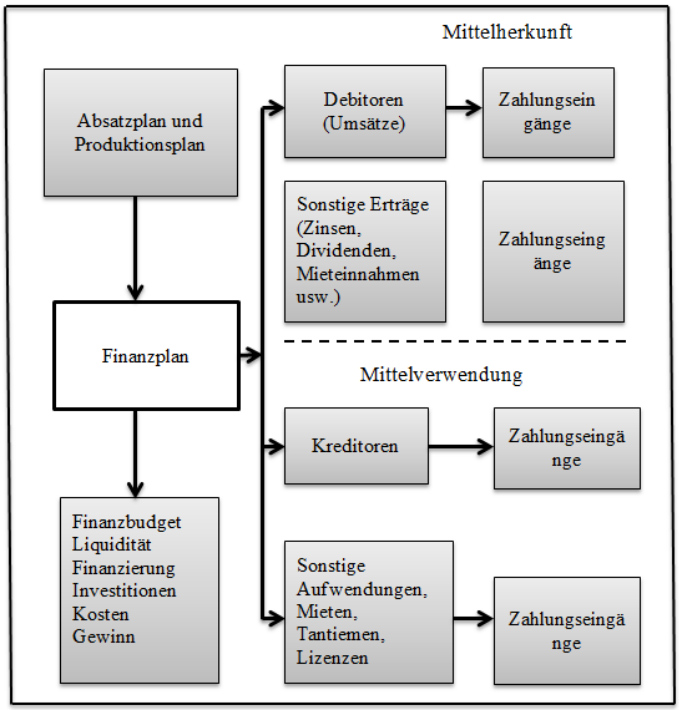


Abbildung 5: Aufbau eines Finanzplans([Fis96], S. 139)

3.2 Integrierte Unternehmensplanung

Integrierte Unternehmensplanung bedeutet die Vernetzung aller Teilpläne der operativen Planung und erfasst das Gesamtunternehmen. Es werde sämtliche Bereiche und Variablen des Unternehmens insgesamt unter gegenseitiger Abstimmung geplant(Vgl.[Fis96], S. 8).
Integrierte Unternehmensplanung bezieht sich auf die Technologien, Anwendungen und Prozesse die den Anschluss der Planungsfunktion im Unternehmen zur organisatorischen Ausrichtung und finanziellen Leistungen verbessern.
Integrierte Unternehmensplanung stellt genau ein ganzheitliches Modell des Unternehmens da, um die strategische Planung und operative Planung mit der Finanzplanung zu verknüpfen.

Durch den Einsatz eines einzigen Modells im gesamten Unternehmen und die Nutzung von Informationsressourcen der Organisation , Führungskräften, Geschäftsbereichsleitern und Planungsmanager verwendet integrierte Unternehmensplanung Pläne und Aktivitäten auf der Grundlage der tatsächlichen wirtschaftlichen Auswirkungen.

Die Planung stellte eine Auseinandersetzung mit der Zukunft da, wobei die Zukunft durch Ziele fixiert ist. Durch die integrierte Planung reagiert das Unternehmen nicht mehr passiv mit sondern agiert oft am Marktgeschehen.

In Unternehmen gibt es Pläne, die in einem hierarchischen Verhältnis zueinander stehen. Diese werden in drei Hierarchistufen unterteilt:

3.2.1 Strategische Planung:

Die Ausmaße der strategischen Planung sind die Ertragspotentiale Markt, Produkt und Mitteleinsatz (Ressourcen) des Unternehmens. Diese sind die Aktivitätsfelder des Unternehmens, in denen sich das Unternehmen Vorteile in Form wie von Marktanteilen und Qualitätsvorteilen aufbaut. Die strategische Planung muss markt- und zielorientiert sein, um knappe Mitteln zielorientiert und konzentriert in den Markt einzusetzen. Der Planungszeitraum erstreckt sich über fünf bis zehn Jahr hinaus. Durch rechtzeitige Kurskorrekturen kann eine strategische Lücke in der Zukunft vermieden oder zumindest abgemildert werden.

Die strategische Planung ist in Primärstrategien und Sekundärstrategien unterteilt. Bei der Primärstrategie werden im Unternehmen Absatzstrategien und Entwicklungsstrategien eingesetzt. Die Sekundärstrategie unterteilt sich in Produktions-, Beschaffungs-, Personal-, Investitions-, Finanz- und Standortstrategien(Vgl.[Fis96], S. 38 ff).

Abb. 6 stellt auf, mit welchen Fragen sich die strategische Planung befasst und was dadurch erreicht werden soll.

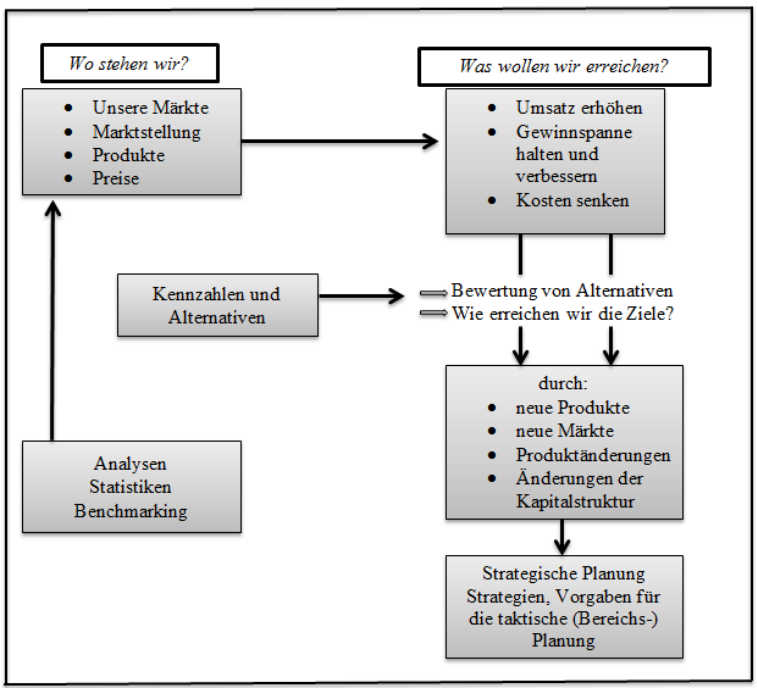


Abbildung 6: Strategische Planung ([Fis96], S. 40)

3.2.2 Taktische oder Bereichsplanung:

Die taktische Planung oder Bereichsplanung konkretisiert die Rahmenvorgaben der strategischen Planung und umfasst alle im Unternehmen vorkommenden operativen Bereiche. Sie operiert innerhalb dieser Bereiche von den Rahmenvorgaben der strategischen Planung, die dann in ope-

rative Planung in Einzelzahlen aufgelöst werden. Taktische oder Bereichsplanung ist meist kurz oder eine mittelfristige Planung von 2 bis 3 Jahren(Vgl.[Fis96], S. 41).

Abb. 7 zeigt die taktische Bereichsplanung eines Industriebetriebes von der strategischen Planung.

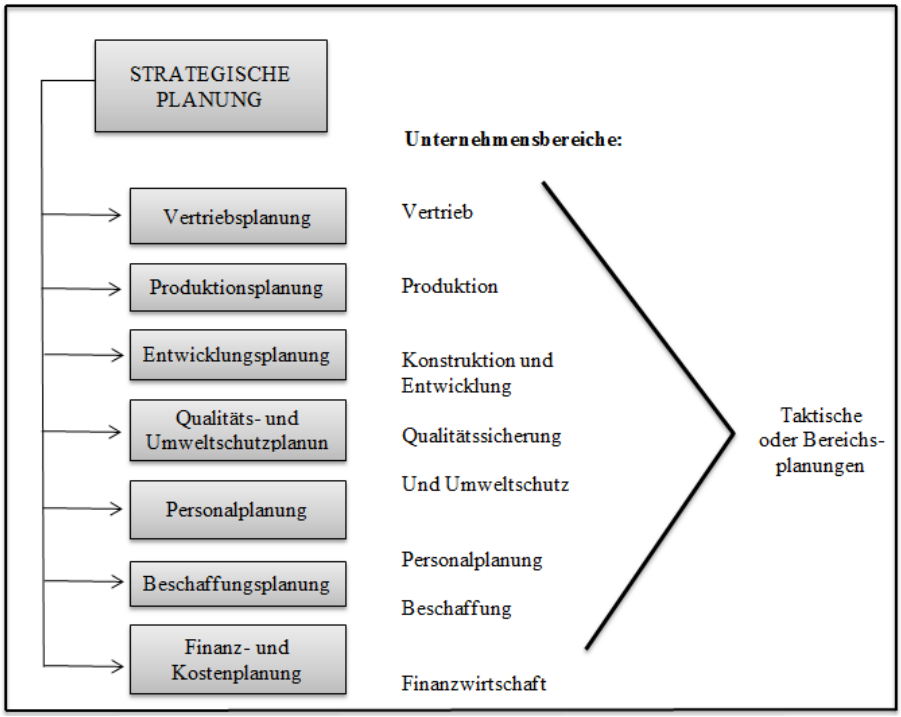


Abbildung 7: Taktische oder Bereichsplanung([Fis96], S. 41)

3.2.3 Operative Planung:

Die operative Planung stellt eine Die operative Planung stellt eine kurzfristig ausgelegte Planung innerhalb eines Jahres der Prozesse im Rahmen gegebener Kapazitäten da. Sie ist eine Feinplanung mit dem Ziel der Minimierung der negativen Auswirkungen taktischer Fehlplanung. Die Zahlen der operativen Planung gehen in die Fertigsteuerung bei produzierenden Unternehmen und in den Einkauf bei Handelsbetrieben ohne zusätzlich noch eine Korrektur vorzunehmen. Fischer spricht von einer sogenannten „Frozen Zone“(Vgl.[Fis96], S. 41).
Die Abb. 8 zeigt den Ablauf der strategischen über die taktische zur operativen Planung.

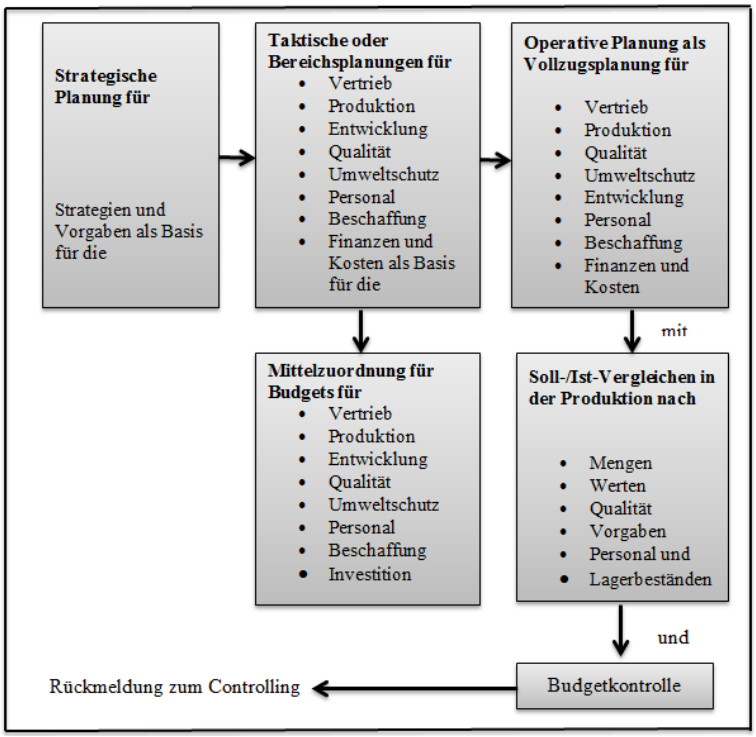


Abbildung 8: Operative Planung (Durchführungsplanung)[Fis96], S. 42

3.3 Werkzeuge in integrierte Unternehmenssteuerung

Die Vollständigkeit der integrierten Unternehmensplanung kann nur erfolgen, wenn Werkzeuge für die Planung eingesetzt werden.

Die BI (Business Intelligence) sind wichtige Bestandteile der Werkzeuge des Unternehmens im Management. Diese ermöglichen Projekte qualitativen Nutzens für Kontrolle, Planung und Steuerung der Unternehmensleistung und entwickeln Produkte kontinuierlich weiter. Das Scheitern hängt in dem Fall nicht vom Programm ab, sondern vom Unternehmen, welches geeignete Programme für seine Zwecke finden muss. Deshalb ist es wichtig zur Umsetzung das passende Anwendungspaket bzw. Programm zu suchen und diese anzuwenden.

Durch die richtigen Anwendungspakete werden Planungsprozesse beschleunigt und verbunden, Kosten werden gesenkt, Komfortabilität und Benutzerfreundlichkeit sowie die Konzentration auf Nutzanwendungen steigen. Die Einsparung von Zeit und Deutung einer Prognose ergeben sich und erleichtern die Abstimmungen bzw. Bewertungen. Zudem ist es die sicherste Deckung der Unternehmensdaten. Die Planungsabläufe sind bereichsübergreifend verfügbar und somit entsteht eine Kontrolle der Unternehmensbereiche. Dahingegen entsteht ein zu hoher Kostenaufwand zum Beispiel für die Schulungen von Arbeitskräften. Das nimmt zusätzlich viel Zeit in Anspruch und führt dazu, dass nur wenige Unternehmen die Anwendungspakete nutzen(Vgl.[Klu06], S. 11 f).

Die Abbildung der Business Application Research Center (BARC) stellt ein Vergleich zwischen den Top 10 BI der Werkzeuganbieter in Deutschland. SAP ist vor Oracle in Deutschland der anerkannteste Marktführer im Geschäft mit Business Intelligence (BI) Software. Das zeigt die erhöhte Entwicklung von SAP in den Jahren 2007 bis 2011.

Rang	Unternehmen	Softwareumsatz 2007 (Mio. Euro)	Softwareumsatz 2008 (Mio. Euro)	Marktanteil 2008	Wachstumsrate 2007-2008	Rang	Unternehmen	Softwareumsatz 20011 (Mio. Euro)	Marktanteil 2011	Veränderung zu 2010
1	SAP	101	110	14,6 %	8,9 %	1	SAP	171	16 %	8 %
2	Oracle	96	104	13,8 %	7,8 %	2	Oracle	140	13 %	6 %
3	SAS	85	89	11,8 %	4,7 %	3	IBM	135	13 %	14 %
4	IBM	73	72	9,5%	-1,4 %	4	SAS	115	11 %	8 %
5	Microsoft	44	53	7,1%	20,0 %	5	Microsoft	83	7,8 %	14 %
6	Teradata	30	33	4,4%	9,2 %	6	Informatica	43	4 %	5 %
7	Micro Strategy	23	24	3,2%	3,0 %	7	Qlik Tech	36	3,3 %	40 %
8	Informatica	13	16	2,1%	23,1 %	8	Micro Strategy	33	3,1 %	25 %
9	Qlik Tech	9,5	14	1,9%	47,4 %	9	Teradata	29	2,7 %	-8 %
10	SPSS	13	14	1,8%	9,5 %	10	Software AG/IDS Scheer	17	1,6 %	6 %

Abbildung 9: Top 10 Umsatz BI in Deutschland ([GB14])

Die SAP ist ein wichtiger Werkzeuganbieter der integrierten Unternehmensplanung. Mit BW-integrierten Planungslösung (BW-IP) und Business Planning and Consolidation (BPC) hat SAP zwei Lösungen für unterschiedliche Planungsaufgaben.
Durch den Zukauf von OutlookSoft und Business Objects im Jahr 2007 entstanden zwei verschiedene SAP-Planungswerkzeuge. Beide Werkzeuge zeichnen sich durch individuelle Funktionen und Einsatzmöglichkeiten aus, deren Vorteile die Fach- und IT-Abteilungen gegeneinander abwägen müssen(Vgl.[Com14]).

Allgemeiner Überblick	BW-IP	BPC
Genereller Ansatz beziehungsweise Einsatz	Für zentrale unternehmensweite Planung. IT-und Backend-getriebene Umsetzung (hohes Datenvolumen, Stabilität, Sicherheit, Konsistenz).	Für dezentrale Fachbereichs- und Abteilungsplanung. Fachbereichs- und Frontend-getriebene Umsetzung(hohe Flexibilität, Benutzerfreundlichkeit).
Planungsprozess	Vordefiniert, zentral gesteuert, hierarchisch.	Flexibel, abteilungsspezifisch.
Administration	Zentral durch IT.	Dezentral durch Enduser.
Integration	Backend-Ansatz (aber Zero-Footprint-Applikation): Volle Data Warenhouse- und SAP-Integration (z.B. ERP, Portal, BCS...).	Frontend-Ansatz. Integration in MS Office. Einheitliche Planung, Konsolidierung und Reporting innerhalb von BPC.
Reporting und Analyse	Vordefinierte Reports durch IT oder Poweruser, Nutzung durch Enduser.	Eigenständiges Reporting innerhalb von BPC(Ad-hoc-fähig).
Konsolidierung	Nicht vorhanden. BCS oder EC-CS muss verwendet werden.	Implementiert, inklusive IFRS Starter Kit.

Tabelle 1: SAPs Planungslösungen im Vergleich([Com14])

BW-IP und BPC Gemeinsamkeiten sind die Verkürzung der Planungszyklen von Unternehmen, Reduzierung der manuellen Tätigkeiten und die Steigerung der Planungsqualität. Ein wichtiger Unterschied ist , dass sich BW-IP eher für eine zentrale unternehmensweite Planung eignet, weil harmonisierte Daten aus dem Business Warehouse genutzt werden können, während BPC mehr für eine dezentrale Fachbereichs- und Abteilungsplanung in Frage kommt. Für Unternehmen die über die Geschäftsbereiche hinweg einheitlich planen wollen, ist die integrierte Planungslösung BW-IP vom Vorteil. Die Unternehmen müssen dann aber auf die Flexibilität und Benutzerfreundlichkeit verzichten, die BPC den Fachanwendern bietet. Bei BW-IP sind sämtliche Planungsprozesse vordefiniert, zentral gesteuert und hierarchisch geordnet, was die Verwendung einheitlicher Planungsobjekte und Planungslayouts einschließt. Dadurch kann eine noch höhere Planungsgenauigkeit erzielt werden als mit BPC. Eine weitere Besonderheit von BW-IP ist, dass die IT-Abteilung während des gesamten Produktlebenszyklus die Hoheit über die Administration behält, wie z. B. der Installation, Wartung und der technischen Unterstützung der Umsetzung von Planungsthemen(Vgl.[Com14]).

3.4 Vor-Nachteile einer integrierten Unternehmensplanung

Die klassische Unternehmensplanung wird in integrierter Unternehmensplanung geführt. Durch eine integrierte Unternehmensplanung soll ein Unternehmen erfolgreich langfristige Planungen festlegen. Diese vermeiden zukünftige Krisen und schätzen vor allem finanzielle Risiken besser ein. Es erfolgt eine Operation der Planung für Rentabilität. Die Aktivitäten werden intern besser koordiniert und es führt zur Beschleunigung von Wachstum und Ertrag.

Ein großer Nachteil der heutigen Unternehmensplanung ist der große Zusammenhang und der Dynamik zwischen Unternehmen und der Umwelt. Die Umwelt verändert sich häufig. Einige Ursachen dafür sind die gesellschaftlichen, politischen, wirtschaftlichen und technischen Bereiche, die sich auf Märkte, Produkte und die Unternehmen selbst einwirken. So ist es, dass vom Jahr zu Jahr die Energie- und Rohstoffe knapper werden, eine zunehmende Weltverschuldung vorhanden ist und die zahlreichen politischen Krisen sowie regionale Kriege Auswirkungen darauf haben.

Aufgrund der oben genannten Vor und Nachteile einer strategisch planenden Firma und einer rein auf die operativen Geschäftstätigkeiten ausgerichteten Wettbewerber, stellt sich heraus, dass das Prinzip einer strategisch planenden Firma sinnvoller und erfolgreicher ist. Dies wird durch das folgende Zitat unterstützt:

„Es ist heute nachgewiesen, dass strategische planende Firmen erfolgreicher sind als ihre rein auf die operativen Geschäftstätigkeiten ausgerichteten Wettbewerber“[Pue84], S. 19-30

4 Zusammenfassung

Das nachgegangene Verfahren der integrierten Unternehmensplanung liegt vor, wenn Teilplanungen aufeinander bezogen erfolgen. Es handelt sich in dem Fall um ein Planungsverfahren. Das Ziel des Verfahrens ist ein abgestimmter Gesamtplan.

Die integrierte Unternehmensplanung wurde im zweiten Teil spezifisch mit ihren Teilplanungen dargestellt. Die Integration von Teilplanungen soll zu einem abgestimmten Planungssystem führen, in dem die bestehenden Interdependenzen Berücksichtigung finden. Als geordnetes System von aufeinander bezogenen Teilplanungen entsteht integrierte Unternehmensplanung durch methodisches und systematisches Zusammenfügen und Verknüpfungen von Teilplanungen. Dabei ist es erstmal erforderlich von dem komplexen Untersuchungsgegenstand der isolierten Teilplanung mit Absatz-, Werbe-, Beschaffungs-, Fertigung-, Produktions-, und Finanzplanung sich schrittweise der integrierten Unternehmensplanung zu nähern. Die isolierte Planung führt zwar auf abgestimmten Teilplanungen, gilt aber nicht als integrierte Unternehmensplanung.

In der integrierten Unternehmensplanung steht im Unternehmen das Verhältnis dreier Hierarchiestufen untereinander, zwischen der strategischen, taktischen und operativen Planung. Die strategische Planung befasst sich mit den Aktivitätsfeldern des Unternehmens vorwiegend nach den Kriterien des günstigsten Gewinnprofils bzw. der maximalen Gewinnpunktsumme. Die taktische Planung konkretisiert die Rahmenvorgaben der strategischen Planung. In der operativen Planung mag es möglich sein, mit dem Kriterium des maximalen typischen Gewinns je Jahr zu arbeiten aber es ist sinnvoll, bei mittelfristigen Optimierungen in einem Geschäftsbereich eine mehrperiodige Optimierungen durchzuführen.

Literatur

[Com14] Computerwoche: *SAP-Planungswerkzeuge*. http://www.computerwoche.de/a/sap-planungswerkzeuge-welches-ist-das-richtige,2547991. Version: 2014, Abruf: 23.07.2014

[Edi14] Editorial, Gründerszene: *Gründer Szens*. http://www.gruenderszene.de/lexikon/begriffe/absatzplanung. Version: 2014, Abruf: 09.05.2014

[Ehr13] Ehrmann, Harald: *Unternehmensplanung*. Bad Reichenhall, 2013. – ISBN 978–3–470–46836–5

[Fis96] Fischer, Hellmuth: *Unternehmensplanung*. Vahlen, München, 1996. – ISBN 3–8006–2019–7

[GB14] Geschäftsführer BARC, Dr. Carsten B.: *Business Intelligence Werkzeuge*. http://mybac.googlecode.com/svn/trunk/Material/Beispiele/themen/Vortraege/Vortrag_Bange.pdf. Version: 2014, Abruf: 18.07.2014

[Klu06] Kluge, T.M.: *Voraussetzungen für die Einführung einer integrierten Unternehmensplanung für Unternehmen der Serienfertigung mit SAP/R3*. Diplom.de, 2006

[Koc82] Koch, Helmut: *Integrierte Unternehmensplanung*. Gabler Wiesbaden, 1982. – ISBN 3–409–34671–6

[MS14] Müller-Stewens, Prof. Dr. G.: *Gabler Wirtschaftslexikon*. http://wirtschaftslexikon.gabler.de/Definition/unternehmensplanung.html. Version: 2014, Abruf: 06.05.2014

[Pue84] Puempin, Cunio: Unternehmenskultur, Unternehmensstrategie und Unternehmenserfolg. In: *gdi impuls* 2 (1984)

[Rol01] Rollberg, Roland: *Integrierte Unternehmensplanung*. Wiesbaden, 2001. – ISBN 3–8244–0584–9

[Sim14] Simon, Fabian: *Rechnungswessen*. http://www.rechnungswesen-verstehen.de/bwl-vwl/marketing/werbeplanung.php. Version: 2014, Abruf: 12.06.2014

[Wir14] Wirtschaftslexikon24: *Fertigungsplanung*. http://www.wirtschaftslexikon24.com/d/fertigungsplanung/fertigungsplanung.htm. Version: 2014, Abruf: 16.06.2014

Vor- und Nachteile von in Memory Computing

Seminararbeit
im Rahmen der Projektgruppe „inMemory Planung mit SAP HANA“

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	M.Eng.Tech. Viktor Dmitriyev
Vorgelegt von:	Rima Adhikari (K.C.) Ammerländer Heerstraße 94 26129 Oldenburg 015123789980 rima.adhikari.kc@uni-oldenburg.de
Abgabetermin:	02. August 2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

1 Einleitung **4**
1.1 Historie von Databanksystem : 4
1.2 In Memory computing 4

2 Motivation **4**

3 Hauptteile **6**
3.1 Vorteile des In Memory Computing 6
3.2 Nachteile des In Memory Computing 8

4 Einsatz von In Memory Computing:SAP HANA **9**
4.1 Fallstudie:Vaillant Group optimiert den ökonomischen Wirkungsgrad 9
4.1.1 Einleitung 9
4.1.2 Herausforderung 9
4.1.3 Ziel der Fallstudie 9
4.1.4 Ansatz 9
4.1.5 Prozess 9
4.1.6 Ergebnis 9

5 Zusammenfassung **11**

6 Fazit **12**

Literaturverzeichnis **13**

Abbildungsverzeichnis

1 Herausforderung für die Unternehmen 5
2 Reduktion von RAM Preis 6

1 Einleitung

1.1 Historie von Databanksystem :

Die Datenbank ist heute ein sehr wichtiger Teil für die Unternehmen. Die einfache Definition von Datenbank ist die Ansammlung von Daten. Datenbanksysteme und Datenverwaltung hat lange Geschichte.

- Ab 18. Jhd.: Lochkarten
- 1956: Erfindung der Festplatte
- 1968 – 1975: Hierarchisches Datenmodell
- 1975 – 1980: Netzwerkdatenmodell
- Ab 1980: Relationales Datenmodell
- Objektorientierte und objektrelationale Datenbanken: Entwicklung in den 90er Jahren, Einsatz in Produktivumgebungen ist im Kommen[JL05].

Jetzt in diesen Tagen hat in Memory Computing eine zunehmende Bedeutung im Databankmanagementsystem.

1.2 In Memory computing

In-Memory speichert alle Daten bzw. Informationen im Arbeitsspeicher des jeweiligen Servers. Das heißt, die Daten werden im Hauptspeicher gespeichert und verwaltet. „Durch die Entwicklung der Speichertechnologie ist es möglich, Datenvolumen im Hauptspeicher eines Computers zu halten, welche früher auf langsamere Speichermedien wie Datenbanken auf Festplatten ausgelagert werden mussten. Diese Verarbeitung im Hauptspeicher In-Memory-Computing genannt“. Bekannte Beispiele sind Google, Facebook oder auch Amazon[Gmb14].

2 Motivation

Technologie wächst sehr schnell. Die wachende Technologie erzeugen überwiegende Daten. Z.B. Während der Produktherstellung werden große Datenmenge durch Fließband oder Fertigung-Robert erzeugt sowie Sozial Media haben auch einen Beitrag dafür. Je mehr Daten , desto mehr kompliziert sie zu verwalten. Deswegen haben Neue Technologien einen wachsenden Einfluss auf den Unternehmenserfolg. Die großen Datenmengen bringt folgende Herausforderung für Unternehmen.

- **Die schiere Menge:**Das Datenvolumen steigt sehr schnell an. Heute werden Datenmengen im Terabyte-Bereich analysiert, in die Zukunft werden sie im Petabyte und Exabyte – Bereich analysiert.

- **Der Zeitdruck:** Die massive Datenmenge sollten idealerweise in Echtzeit analysieren. Damit die Unternehmen zeitnah auf Marktänderungen reagieren können.
- **Die mangelnde Struktur:**Daten kommen aus verschieden Quellen. Strukturierten und unstrukturierten Datenquellen steigt die Komplexität zur Datenanalysen. Daten aus Weblogs und Sozial-Media-Plattformen, RFID etc.
- **Die wachsende Anwenderzahl:**Die Anzahl der potenziellen internen und externen Benutzer steigt immer mehr. Je mehr Anwender sind, desto mehr wird Daten erzeugt und mehr Daten heißt mehr Herausforderung[Lin12].

.

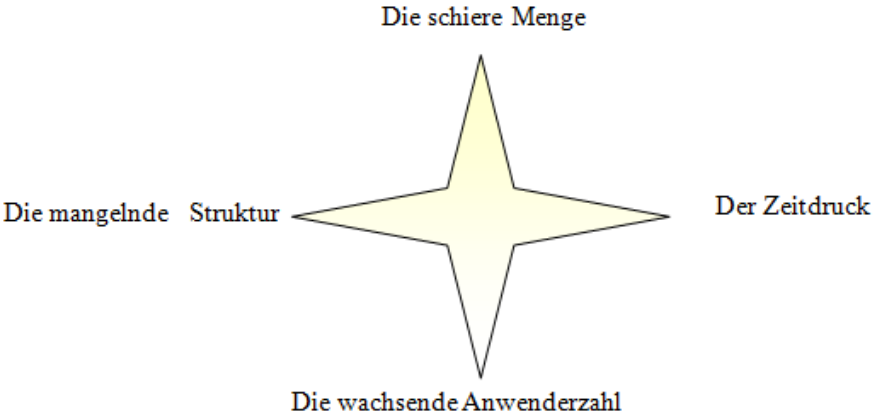


Abbildung 1: Herausforderung für die Unternehmen

Abheben der Last der großen Daten spielt in Memory Computing eine wichtige Rolle.

3 Hauptteile

3.1 Vorteile des In Memory Computing

- **Reduzierung von Kosten und Verbesserung der IT-Effizienz:** Durch die Weniger Hardware, weniger Datenbanken, Kürzere Zeit bis zur Auslieferung und vereinfachte IT-Landschaften bietet in Memory Computing eine große Kostenvorteile für IT. Früher haben die Organisationen viele Anwendungen. Und durch Einsatz von In Memory Computing mehr Anwendung in einem Server ist heute möglich. Dann wird eine Menge Geld gespart. Z.B. nicht nur Strom, Stellfläche, Kühlung sondern auch die Ersatzkosten alle drei bis vier Jahre für zehn oder zwanzig Server wird weniger. Außerdem brauchen die Unternehmen keine Menschen Speicher zu verwalten. Die Hardware Kosten senken auch jedes Jahr. In die Zukunft können die allen Unternehmen in Memory Computing leisten[Ell13]. Z.B. Ram Preis sinkt 30% alle 12 Monte. Der Preis von 1 TB RAM ist $20k-$ 40k.

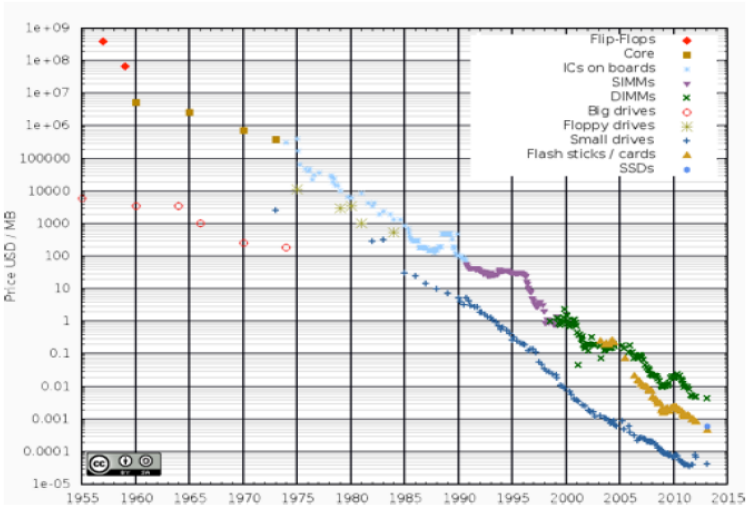


Abbildung 2: Reduktion von RAM Preis

[1]

- **Reduzierung redundante Datenhaltung:** Die Daten aus verschieden Quellen werden auf einer Datenbank gelagert. Eine separate Datenbank für Transaktion oder Analyse ist unnötig. Außerdem benötigt In Memory Computing schlanker Hardware und weniger Systemkapazität. Das reduziert redundante Datenhaltung[AG11].

- **Potenzial in vier Wertdimensionen:**In Memory Computing bietet enormes Potenzial nicht nur in der Senkung der TCO, sondern in allen vier Wertdimensionen: Effizient, Prozess-Innovation, Vereinfachung und Flexibilität.

  – **Effizient:** In Memory Computing können die Daten sehr schnell zugreifen, weil die Daten im Hauptspeicher abgelagert werden und sind nach dem Spaltenprinzip abge-

[1] Ivanov Nikita, Four Myths of in Memory Computing, 23.09.2013

legt. Die Zugriffszeiten sind im von 100 Nanosekunden möglich, während eines Zugriffs auf Festplatten etwa fünf Millisekunden benötigt.

- **Prozess Innovation:**Der Gewinn der Performance bietet das Potenzial für innovative Anwendungen, von Wettbewerbern zu differenzieren.
- **Vereinfachung:** Der Verringerung Schichten kann die Komplexität von Datenmodellen signifikant reduziert werden. Die Reduktion der Komplexität reduziert der mögliche Fehler auch.
- **Flexibilität:** Die nahezu in Echtzeit Kalkulation der Analytik Ergebnisse aus den Rohdaten entspricht die Flexibilität in Bezug auf zwei Dimensionen:
  - **Modifikation von Analysis:**Eine Analyse kann ohne großen Aufwand ändern, nur ein Wechsel zu einer Abfrage wird benötigt.
  - **Integration von zusätzlichen Datenquellen:** Neue Datenquellen können leicht in als zusätzliche Informationsquelle angeschlossen werden, weil jeder Berechnung wird aus Rohdaten gestartet. [MG13]

- Diese Methode kann vor allem von BI -Systemen genutzt werden, da Analyse-Abfragen stärker in spalten- als Zeilenorientiert arbeiten [MF12a].

- **Der Geschäftswert der In-Memory:**In Memory Computing hat die Fähigkeit, in Echtzeit Entscheidungen zu geben. Unternehmen können Wettbewerbsvorteile durch Einsatz von In-Memory gewinnen.
  - **Auf operative Ebene:** Durch die Beschleunigung der Datenerfassung und-Vereinfachen der Prozesse können Unternehmen Lagerbestände reduzieren, Geschäftsrisiken minimieren außerdem geringere Betriebskosten, beschleunigen die Geschwindigkeit auf den Markt, die Förderung von Produktivität und Kundenbedürfnisse besser erreichen.
  - **Auf der Management-Ebene:** durch die Beschleunigung der Entscheidungsfindung und Planung Führungskräfte können Marktchancen schneller nutzen, Wettbewerbsbedrohungen früher identifizieren. Z. B.
    - **Finance and performance management:** Business-Analysten oft verbringen 90 Prozent ihrer Zeit mit Abfragen und deutlich weniger Zeit mit Ist-Analyse. Schnellere Abfragen, mehr Zeit für die Analyse sollte letztlich zu einer insgesamt verbesserten Geschäftsentwicklung führen. Das ist möglich von IMC.
    - **Innovation in CRM:** In-Memory-basierte BI-Tools werden Daten in Echtzeit nicht nur aus Transaktionssystemen schaffen, sondern auch ermöglicht es die Unternehmen, unstrukturierte Daten aus dem Sozial-Media-Bereich zu ernten und zu verwalten. Das hilft die Kunden und Unternehmen nähen zu kommen[Aud12].

3.2 Nachteile des In Memory Computing

- **kosten:**Die Entscheidung des Umstiegs auf in Memory Datenbank verursacht am Anfang erhebliche Kosten für zusätzliche Hardware, Softwarelizenzen sowie einen erhöhten Aufwand für Wartung und Backup/Recovery-Mechanismen. Neue Technologie heißt neue Technikal know how(neues Wissen), deswegen ist die Technologie bisher nicht für alle Unternehmen geeignet. Außerdem spielt eine Rolle dabei, ob das nötige Know-how zur Anpassung vorhandener Anwendungen im eigenen Unternehmen vorhanden ist. Falls die Mitarbeiter mit der Technologie nicht zurechtkommen, sollten sie geschult werden. Trainings, Schulung usw. verursacht auch Kosten.

- **Datenverlust:**Das Risiko von Datenverlust ist höherer in In Memory Computing als Festplatten Technologie, weil in in Memory für die Speichern von Daten Storm benötigt wird. Falls Storm ausfällt, werden Daten Verlust. Ein hoher Aufwand wird benötigt, um dieses Problem durch Back-up-Szenarien zu verhindern. Aber spielt wieder der höhere Kostenfaktor eine wichtige Rolle[MF12b].

- **Abhängig von Anbieter:** Es gibt sehr vielen Anbieter von In Memory Technologie. Die Unternehmen sind abhängig von Anbietern. Z.B. Lizenz außerdem gibt es unterschiedliche Software und die werden von unterschiedlicher Hardware unterstützt. Die Benutzer können nicht selbst entscheiden. Bei der Lizenzverlängerung kostet auch Geld. Falls gleiche Software bzw. Hardware in die Zukunft nicht gibt, sollte Unternehmen auf eine andere Software bzw. Hardware umsteigen und dann wieder neuen Aufwand. Z.B. Microsoft ist nicht mehr verantwortlich für Windows XP.

- **Teurer als Festplatte**Die Unternehmen, die jeden Tag mit sehr großen Datenmengen beschäftigt sind, spielt Kosten nicht so große Rolle. Z.B. Kapitalmarkt, Telekommunikation. Sonst ist Hauptspeicher Technologie im Gegensatz zu Festplatten bis heute immer noch recht teuer, obwohl der Preis von Hardware jedes Jahr sinkt.

- **Daten Sicherheit:**Die Datensicherheit ist auch ein Problem in Memory sein könnte. Auf Festplatten kann der IT Manager Daten verschlüsseln bzw. kontrollieren.

4 Einsatz von In Memory Computing:SAP HANA

SAP HANA ist eine Datenbanktechnologie von SAP, die 2010 vorgestellt wurde.

4.1 Fallstudie:Vaillant Group optimiert den ökonomischen Wirkungsgrad

(Umstieg auf Plattform zur Unternehmenssteuerung auf Basis von SAP HANA mit Hilfe von HP)

4.1.1 Einleitung

In den Bereichen Heizlüftungs- und Klimatechnik tätiges Unternehmen „Vaillant Group“ entwickelt und produziert mit mehr als 12.000 Mitarbeitern an 13 Standorten in sechs europäischen Ländern und China maßgeschneiderte Produkte, Systeme und Dienstleistungen für Wohnkomfort. Das befindet sich seit Gründung Jahr 1874 (140-jährigen Tradition) und die Vaillant Group ist das zweitgrößte europäische Unternehmen in dieser Branche mit einem Jahresumsatz von rund 2,38 Mrd. Euro in 2013 und exportiert ihre Produkte in 60 weitere Länder[Gmb].

4.1.2 Herausforderung

Bereits vor mehr als einem Jahrzehnt wurde die IT-Landschaft der Vaillant Group konsequent standardisiert. Im Markt zu konkurrieren und auch zu existieren, er war notwendig die wachsenden IT-Ansprüche zu erfüllen.

4.1.3 Ziel der Fallstudie

Schaffung einer IT Infrastruktur für die Echtzeitbetrachtung der Kosten und Erträge auf Produckte- und Kundenebene mit SAP Business Intelligence auf SAP HANA.

4.1.4 Ansatz

HP, die Implementierung und der Betrieb der IT- Infrastruktur für In Memory Computing.

4.1.5 Prozess

Bei Einsatz von SAP HANA als Applicance-Software läuft die Datenbank im Hauptspeicher statt im Storage-System auf Platte.

4.1.6 Ergebnis

Ergebnisse für die IT: Durch Schlanke, kosteneffiziente IT-Lösung mit vereinfachter Dateninfrastruktur, Performance-Steigerung um Faktor 10, Datenverdichtung um Faktor 8 biete in Memory Vorteile für IT von Vaillant Group.

Ergebnisse für das Business: Verbesserte Unternehmenssteuerung durch stichhaltige Reports und Analysen in Echtzeit, Blick auf die konzernweiten Kosten und Umsätze nach Produkt

und Kunden, Besseres Forderungsmanagement durch fortlaufende Analyse der Cashflows nach Kunde, Unterstützung der Monats- und Jahresabschlussprozesse, Fundierte Voraussetzungen für schnelle und gezielte Entscheidungen sind die Vorteile, die Vaillant Group aus dem Einsatz von in Memory gezogen hat. Außerdem ist die Datenbankgröße von rund 4.2 TB auf 350 GB gesunken. Der Datenladevorgang läuft um bis zu Faktor 170 schneller. Die Gesamtlaufzeit der Datenladung hat sich von achtzehn auf sieben Stunden verkürzt[Com14].

Dieses Ergebnis beweist, dass In Memory Computing positive Rolle für Unternehmenserfolg spielt.

5 Zusammenfassung

- Schneller und effektiver Abruf und Aufbereitung der großen Datenvolumen ist durch IMC möglich, damit die Unternehmen die wachsenden Herausforderungen des Marktes effektiv lösen könnten. SAP gibt an, durch Einsatz von In-Memory-Technologie 10.000 Abfragen gegen 1,3 Terabytes Daten in einigen Sekunden ausführen zu können[McI11].

- Die Entscheidung des Umstiegs auf in Memory Datenbank verursacht nicht erhebliche Kosten sondern auch spielt eine Rolle dabei, ob das nötige Know-how zur Anpassung vorhandener Anwendungen im eigenen Unternehmen vorhanden ist.

- Nicht nur im Unternehmensbereich(Industrie) kommt In Memory Computing auch in vielen anderen Sportarten zum Einsatz. Fußball: SAP HANA im DFB-Einsatz bei der Fußball-WM 2014 in Brasilien „Mit SAP HANA kann der DFB heute jedes Spiel in Echtzeit analysieren. Laufwege und Pässe werden für jeden Spieler präzise erfasst. So kann das Trainerteam die Leistung objektiver bewerten und Trainings individuell anpassen“. Rennfahrer(Formel 1): Mit SAP HANA 20.000 Szenarien durchrechnen[FO14].

- In-Memory in der Krebstherapie: Die Analyse der großen Patienten bezogene Datenmengen, im Fall von Krebs Patienten werden mit SAP Technologie auf Sekunden bringen. früher das 3-4 Tage gekostet aber mit Hana 2-3 Sekunden[TV].

- Neue Technologie bietet sowohl neue Gelegenheit als auch neue Herausforderung im Unternehmen an. Neue Gelegenheiten in dem Sinn, dass die Unternehmen Marktänderung frühzeitig erkennen und können sofort reagieren. Herausforderung in dem Sinn, dass die Unternehmen in ein neues System investieren und niemand weiß, ob diese Investition sich lohnt oder ob das Ziel erreicht wird.

6 Fazit

> „Tape is Dead, Disk is Tape. Flash is Disk. RAM Locality is King.“JIM GRAY, Dec. 2006

Die Aussage sagt, Tape und Disk sind schon veraltet. Heutzutage ist die Speichertechnologie Tape und Disk durch In Memory (RAM) ersetzt und Speichern im RAM ist König geworden.

Vergleich mit Disk oder Tape ist in Memory noch teuer aber schneller und effektiver Abruf und Aufbereitung der großen Datenvolumen ist hier möglich. Deswegen ist Datenspeichern im RAM im Trend. Wegen des Datenverlustproblems im RAM wird ein Hybridsystem benötigt. Wenn RAM als Primärspeicher und Flash als Sekundärspeicher(Backup) genutzt würden, würde Datenverlustproblem von In Memory(Daten im RAM) beseitigt und RAM würde König. Jedes Jahr sinkt Flashkosten auch. Der Preisunterschied zwischen Disk und Ram wird nicht groß und alle Unternehmen, ob die groß oder klein sind, können in Memory Computing leisten und davon profitieren.

Literatur

[AG11] AG, SAP: *SAP in Memory Computing Technology.* http://fm.sap.com/data/UPLOAD/files/SAP_In-Memory_Computing_Technology_.pdf. Version: 2011, Abruf: 05.07.2014

[Aud12] Auditore, Peter: *In Memory Technologie: Innovation in Business Intelligence?* http://sandhill.com/article/in-memory-technology-innovation-in-business-intelligence/. Version: Mai 2012, Abruf: 26.06.2014

[Com14] Company, Hewlett-Packard D.: *Vaillant Group optimiert den ökonomischen Wirkungsgrad: Umstieg auf Plattform zur Unternehmenssteuerung auf Basis von SAP HANA mit Hilfe von HP.* http://www8.hp.com/h20195/v2/GetPDF.aspx%2F4AA5-1442DEW.pdf. Version: April 2014, Abruf: 25.06.2014

[Ell13] Elloitt, Timo: *Why In-Memory Computing Is Cheaper and Changes Everything",Business-analytics.* http://timoelliott.com/blog/2013/04/why-in-memory-computing-is-cheaper-and-changes-everything.html. Version: April 2013, Abruf: 10.07.2014

[FO14] Friederike Orths, Andreas S.: DFB: Big Data zur Fußball-WM. In: *SAP News center* (2014), März. http://de.news-sap.com/2014/03/11/cebit-2014-merkel-bierhoff-sap/, Abruf: 20.06.2014

[Gmb] http://www.vaillant-group.com/

[Gmb14] Gmbh, Empolis Information M.: *In Memory Computing.* http://www.empolis.com/smart-information-management/technologie/in-memory-computing.html. Version: 2014, Abruf: 12.06.2014

[JL05] Jochen Löhl, Ingo S. Mario Lörcher L. Mario Lörcher: *Historische Entwicklung von Datenbanken.* http://www.chrisix.net/fhhn/ss05/1_hist_entwicklung_handout.pdf. Version: September 2005, Abruf: 20.06.2014

[Lin12] Linden, Klaus H.: *Data Governance ist das A und O, Computer Woche40.* http://www.computerwoche.de/a/data-governance-ist-das-a-und-o,2509488. Version: Mai 2012, Abruf: 16.06.2014

[McI11] McIlvaine, Heather: SAP HANA Surges Ahead. In: *SAP News center* (2011), März. http://www.news-sap.com/hana-inmemory-ibm-x3850-analytics/, Abruf: 06.07.2014

[MF12a] Martin Funk, Michel P. Boris Marinkov M. Boris Marinkov: Trends in der IT „In-Memory Technologie. (2012). http://trends-in-der-it.de/?Fachartikel/In-Memory-_Technologie, Abruf: 01.07.2014

[MF12b] Martin Funk, Michel P. Boris Marinkov M. Boris Marinkov: Trends in der IT „In-Memory Technologie. (2012). http://trends-in-der-it.de/?Fachartikel/In-Memory-_Technologie, Abruf: 01.07.2014

[MG13] Marcel Grandpierre, Ralf E. Georg Buss B. Georg Buss: *In Memory Computing Technologie "the holy grail of analytics?* http://www2.deloitte.com/content/dam/Deloitte/de/Documents/technology-media-telecommunications/TMT_Studie_In_Memory_Computing.pdf. Version: Juli 2013, Abruf: 16.06.2014

[TV] TV, SAP: *In-Memory in der Krebstherapie.* http://www.sap-tv.com/video/#/7611/in-memory-in-der-krebstherapie, Abruf: 22.06.2014

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Analytical capabilities of SAP HANA: Integration with R & Excel

Seminararbeit
im Rahmen der Projektgruppe 2014

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	M.Eng. & Tech. Viktor Dmitriyev
Vorgelegt von:	Eduard Rajski eduard.rajski@uni-oldenburg.de
Abgabetermin:	02. August 2014

Inhaltsverzeichnis

Glossar **3**

Symbolverzeichnis **3**

Abbildungsverzeichnis **4**

Tabellenverzeichnis **4**

1 Einleitung **5**
1.1 Was ist R? 5
1.2 Was ist Excel? 6

2 Integration der Tools in SAP HANA **6**
2.1 R 7
2.1.1 Beispiele für Funktionen und Prozeduren 8
2.2 Excel 12

3 Analytische Möglichkeiten in SAP HANA **14**
3.1 Vor- und Nachteile einer Integration 14
3.2 SAP HANA mit R 15
3.3 SAP HANA mit Excel 16

4 Fazit und Ausblick **17**

Literaturverzeichnis **19**

Glossar

Symbolverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Unternehmen haben heutzutage mit einer stetig steigenden Datenmenge zu kämpfen [RKA14]. Wenn diese nun auch noch nach bestimmten Kriterien analysiert werden müssen, sind bei herkömmlichen Datenbanken oft längere Analyse- beziehungsweise Verarbeitungszeiten zu beobachten [RKA14]. Dies ist einerseits auf die Lese- und Schreibraten von Festplatten, die im Vergleich zu Arbeitsspeichern oder ähnlichen Speichern Daten langsamer verarbeiten, zurückzuführen und andererseits ist die Art wie die Daten gespeichert werden nicht zu unterschätzen, da eine Speicherung von Daten als Tabellen bei größeren Datenmengen zu vielen ressourcen- sowie zeitkonsumierneden Zusammenführungen führen kann [RKA14]. In-Memory Computing kann hier Abhilfe schaffen, da bei dieser Technologie sowohl das Speichermedium als auch die Art und Weise der Speicherung optimiert wurden [RKA14]. Näheres zu In-Memory Computing findet sich in der Seminararbeit von Rima Adhikari Kc der Projektgruppe OLiMP. SAP HANA ist ein Beispiel für eine solche Implementierung der In-Memory Technologie [RKA14]. Doch bietet SAP HANA für sich alleine nur die nötigsten Analysemöglichkeiten [RKA14]. Dafür bietet SAP HANA mehrere Möglickeiten externe Tools einzubinden, wie ein RLANG Parameter für R-Prozeduren und eine MDX-Schnittstelle [RKA14].

Diese Arbeit beschäftigt sich mit der Integration zwei solcher Tools mit SAP HANA und den daraus resultierenden Analysemöglichkeiten. Die zwei Tools sind die Softwareumgebung und Programmiersprache R sowie das BI-Tool (Business Intelligence Tool) Microsoft Excel. Auf der einen Seite wird die Bewerkstelligung der Integration betrachtet und auf der anderen Seite die resultierneden Analysemöglichkeiten, die sich durch die Integration von SAP HANA mit einem Tool ergibt. Dabei soll jeweils ein Beispiel pro Tool für mögliche zusätzliche Analysemöglichkeiten. Abgeschlossen wird dies durch ein entsprechendes Fazit und ein Ausblick auf mögliche zukünftige Entwicklungen.

1.1 Was ist R?

R [Dol04] ist eine Softwareumgebung für statistische Berechnungen und Grafiken, die eine eigene Programmiersprache besitzt [SAP14b]. Neben der Nutzung von R für fortgeschrittene Analysen werden mithilfe der Programmiersprache R speziell angepasste Software für Data Mining und Statistik entwickelt [Dol04]. Sowohl die Entwicklung in R als auch das Ausführen des Codes können in allen größeren Betriebssystemen (Windows/UNIX/Linux/Mac OS) durchgeführt werden [Dol04]. Die Programmiersprache R ist *case-sensitive* und achtet bei Ausführung auf Groß- und Kleinschreibung [Dol04]. Benutzereingaben werden mithilfe einer Kommandozeilenkonsole anhand einer festgelegten Syntax verarbeitet

und als Vektoren oder Matrizen ausgegeben [Dol04]. Für statistische Auswertungen werden *data frames* genutzt [Dol04]. Diese matrizenförmigen Datensätze können im Gegensatz zur üblichen Syntax auch Daten verschiedener Typen enthalten, solange die jeweilige Spalte Einträge desselben Typs hat [Dol04]. Als letzte Ausgabeform besitzt R Listen, die innerhalb dieser Listenstrukturen verschiedene Typen enthalten kann [Dol04].

1.2 Was ist Excel?

Die Tabellenkalkulation Excel [Wal13] [Sch07] kommt vom Softwarehersteller Microsoft und ist Teil des Microsoft Office Paketes, dass entweder Windows oder Mac OS als Betriebssystem voraussetzt. Excel erlaubt durch verschiedene Hilfsmittel in Form von statistischen Funktionen und Assistenten auch Personen, die nicht mit umfangreichen mathematischen Vorkentnissen ausgestattet sind, die transparente Erzeugung von komplizierten Statistiken [Sch07]. Die verarbeitbaren Datentypen sind nur im Bezug auf die für den Kontext genutzten Hilfsmittel begrenzt [Sch07]. Ein Beispiel ist die Formel für die Mittelwertberechnung, die aufgrund der Funktion entsprechend Zahlen voraussetzt. Die Eingaben können dabei mithilfe eines Assistenten getätigt werden, der den Anwender durch diese Eingabe führt [Sch07]. Bei fortgeschrittenem Kenntnisstand können entsprechende Formeln der Excel-Syntax entsprechend eingegeben werden [Sch07]. Je nach Ziel können Statistiken sowohl in Tabellen als auch in verschiedenen Diagrammen dargestellt werden [Wal13] [Sch07].

2 Integration der Tools in SAP HANA

SAP HANA ist eine sogenannte *In-Memory Datenbank* [RKA14]. Dies bedeutet, dass die Datenbank RAM-Speicher anstatt Festplatten zur Speicherung von Daten nutzt [RKA14] [SAP14b], wodurch eine Effizienzsteigerung bei Verarbeitung von Daten in SAP HANA herbeigeführt werden kann. Die Kernkomponente von SAP HANA ist dabei die *In-Memory Processing Engine* [RKA14], die die Verarbeitung innerhalb der Datenbank durchführt. SAP HANA kann neben ihrer eigenen auch fremde *Calcualtion Engines* ausführen oder durch entsprechende Schnittstellen mit fremden Applikationen kommunizieren [RKA14]. Zu diesen Schnittstellen und Protokollen gehören JDBC (Java Database Connectivity), ODBC (Open Database Connectivity), ODBO, auch OLE DB for OLAP (Object Linking and Embedding, Database for Online Analytical Processing) genannt, sowie SQLDBC (SQL Database Connectivity) [RKA14]. Das Ansprechen dieser Schnittstellen wird in der Regel mit SQL (Structured Query Language), MDX (Multidimensional Expressions) und

BICS (Business Intelligence Consumer Services) Sprachen bewerkstelligt [RKA14]. Dies erlaubt die Integration verschiedener Analysetools, unabhängig davon, ob sie nun in direkter Relation zu SAP-Produkten stehen. Im Zuge dieses Kapitels soll die Integration der Tools R und Excel betrachtet werden.

2.1 R

Die Integration von R in SAP HANA findet nicht direkt in HANA statt, sondern erfordert eine externe Errichtung einer R-Umgebung sowie eines dazugehörigen Server [RKA14] [SAP14b]. Diese werden dann mit einer auf SAP HANA laufenden R-Clientsoftware angesprochen, welches mithilfe eines *R-Operators* Befehle aus SAP HANA heraus interpretieren kann, um sie passend an die R-Umgebung weiterzureichen [SAP14b]. SAP HANA nutzt dafür die hauseigene SQL-Methode RLANG, die den R-Code an den R-Clienten weitergibt [SAP14b]. Dieser Client gibt mithilfe eines R-Servers diesen Code an die R-Umgebung weiter, in der dieser Code ausgeführt sowie die benötigten Daten aus dem Clienten abgefragt und geholt werden [SAP14b]. Der R-Server nutzt dafür eine TCP/IP-Verbindung [RKA14]. Das Ergebnis wird anschließend aus der R-Umgebung zurück an den Clienten zurückgegeben [SAP14b]. In diesem wird es mithilfe des Operators in ein für SAP HANA verständliches Format umgewandelt und ausgegeben [SAP14b]. Um eine weitere Effizienzsteigerung in diesem Prozess zu erwirken, wird das Aufbrechen der Anfragen in kleinere Einzelabfragen empfohlen. Diese können nahezu parallel ausgeführt, was bei einer großen Prodezur nicht im selben Maße möglich ist.

Grenzen tun sich in diesem Zusammenhang darin auf, dass als Parameter in dieser Kommunikation nur Tabellentypen erlaubt sind [SAP14b] [RKA14]. Andere Typen aus R werden erst in ein SAP HANA konformes Format umgewandelt, damit sie von HANA ausgegeben werden können [SAP14b] [RKA14]. Weiterhin sollte ausschliesslich Kleinschreibung in der RLANG-Methode genutzt werden, um eine fehlerfreie Verarbeitung zu gewährleisten [SAP14b] [RKA14]. Ausserdem muss die Methode für einen fehlerfreien Ablauf mindestens ein Ergebnis zurückgeben [SAP14b] [RKA14].

Die unterstützten Datenstrukturen und -Formate aus R [Dol04] sind neben Raw-Daten auch numerische Typen, wie Integer und Double, Strings, wie Buchstaben und Faktoren, sowie Zeitangaben, wie Date und Datestamp [SAP14b] [RKA14]. Diese korrespondieren entsprechend mit den SAP HANA SQL Typen, wie INTEGER, DOUBLE, FLOAT, VARCHAR, DATE und TIMESTAMP [SAP14b] [RKA14].

Weiterhin kann durch SAP HANA auf die prädikative Analysebibliothek, der zuvor integrierten R-Umgebung, zugegriffen werden, wodurch sich die Analysemöglichkeiten von

SAP HANA erweitern. Auch hier wird entsprechend SQLScript genutzt, um auf diese Bibliothek zuzugreifen.

2.1.1 Beispiele für Funktionen und Prozeduren

Im Folgenden sollen verschiedene Beispiele für Funktionen und Prozeduren, die in einer solchen Integration genutzt werden würden, erörtert werden.

Als erstes sollte die Vorbereitung beziehungsweise Erstellung der R-Daten für die weitere Verarbeitung in SAP HANA angegangen werden. Aus dem Codebeispiel Listing 1 ist zu entnehmen, dass zu Begin in jedem Fall eine neue Tabelle sowie eine Reihe an Spalten erzeugt werden. Im nächsten Schritt wird eine RLANG-Prozedur erstellt, die einerseits eine R-Bibliothek zur Verarbeitung definierenl und andererseits die zu importierenden Daten in SAP HANA laden soll. Die zuvor erstellten Tabellen *spam* dienen im folgenden Schritt als Vorlage für neue Tabellen, die die zu ladenden Daten enthalten sollen. Im letzten Schritt werden die Daten in die SAP HANA Datenbank geladen und entsprechend den Spezifikation in dieser gespeichert.

```
1  DROP TABLE "spam";
2  CREATE COLUMN TABLE "spam"(
3  "make" DOUBLE, "address" DOUBLE, "all" DOUBLE, "num3d" DOUBLE, "our" DOUBLE,
4  "over" DOUBLE, "remove" DOUBLE, "internet" DOUBLE, "order" DOUBLE, "mail"
5  DOUBLE, "receive" DOUBLE, "will" DOUBLE, "people" DOUBLE, "report" DOUBLE,
6  "addresses" DOUBLE, "free" DOUBLE, "business" DOUBLE, "email" DOUBLE, "you"
7  DOUBLE, "credit" DOUBLE, "your" DOUBLE, "font" DOUBLE, "num000" DOUBLE, "
       money"
8  DOUBLE, "hp" DOUBLE, "hpl" DOUBLE, "george" DOUBLE, "num650" DOUBLE, "lab"
9  DOUBLE,
10 "labs" DOUBLE, "telnet" DOUBLE, "num857" DOUBLE, "data" DOUBLE, "num415"
       DOUBLE,
11 "num85" DOUBLE, "technology" DOUBLE, "num1999" DOUBLE, "parts" DOUBLE,
12 "pm" DOUBLE, "direct" DOUBLE, "cs" DOUBLE,"meeting" DOUBLE, "original"
       DOUBLE,
13 "project" DOUBLE, "re" DOUBLE, "edu" DOUBLE, "table" DOUBLE, "conference"
14 DOUBLE, "charSemicolon" DOUBLE, "charRoundbracket" DOUBLE, "
       charSquarebracket"
15 DOUBLE, "charExclamation" DOUBLE, "charDollar" DOUBLE, "charHash" DOUBLE,
16 "capitalAve" DOUBLE, "capitalLong" DOUBLE, "capitalTotal" DOUBLE,
17 "type" VARCHAR(5000), "group" INTEGER);
18 DROP PROCEDURE LOAD_SPAMDATA;
19 CREATE PROCEDURE LOAD_SPAMDATA(OUT spam "spam")
20 LANGUAGE RLANG AS
21 BEGIN
```

```
22  ##--if the kernlab package is missing see Requirements
23  library(kernlab)
24  data(spam)
25  ind <- sample(1:dim(spam)[1],2500)
26  group <- as.integer(c(1:dim(spam)[1]) %in% ind)
27  spam <- cbind(spam, group)
28 END;
29 DROP TABLE "spamTraining";
30 DROP TABLE "spamEval";
31 CREATE COLUMN TABLE "spamTraining" like "spam";
32 CREATE COLUMN TABLE "spamEval" like "spam";
33 DROP PROCEDURE DIVIDE_SPAMDATA;
34 CREATE PROCEDURE DIVIDE_SPAMDATA()
35 AS BEGIN
36 CALL LOAD_SPAMDATA(spam);
37 Insert into "spamTraining" select * from :spam where "group"=1;
38 Insert into "spamEval" select * from :spam where "group"=0;
39 END;
40 CALL DIVIDE_SPAMDATA();
41 Alter Table "spamTraining" DROP ("group");
42 Alter Table "spamEval" DROP ("group");
```

Listing 1: Vorbereitung von R-Daten [SAP14b]

Sind nun passende Daten im System vorhanden, so werden verschiedene Prozeduren für die Analyse dieser Daten benötigt. Neben den bereits in SAP HANA vorhandenen Funktionen können mithilfe der Integration auch Prozeduren aus R genutzt werden. Im Falle des Codebeispiels Listing 2 findet ein direkter Aufruf einer R-Prozedur statt. Im Beispiel wird eine Tabelle *spamClassified* nach dem Schema der *spamEval* erstellt und eine Spalte für die Klassifikationsergebnisse hinzugefügt. Das erstellte Schema wird in der folgenden Prozedurerstellung als Outputschema definiert. Weiterhin wird in der Prozedur durch R auf die entsprechenden zu verwendenden Daten in SAP HANA verwiesen, die im Zuge der Ausführung als data frames mit übergeben werden. *result* verweist auch auf die entsprechende Variable im Bezugssystem. Dementsprechend müssen die Ergebnisse der *USE_SVM* Prozedur, die im result-Feld gespeichert werden auch als data frames übergeben werden.

```
1 DROP TABLE "spamClassified";
2 CREATE COLUMN TABLE "spamClassified" LIKE "spamEval" WITH NO DATA;
3 ALTER TABLE "spamClassified" ADD ("classified" VARCHAR(5000));
4 DROP PROCEDURE USE_SVM;
5 CREATE PROCEDURE USE_SVM(IN train "spamTraining", IN eval "spamEval",
6 OUT result "spamClassified")
```

```
7  LANGUAGE RLANG AS
8  BEGIN
9   library(kernlab)
10  model <- ksvm(type~. , data=train, kernel=rbfdot(sigma=0.1))
11  classified <- predict(model, eval [,-(which(names(eval) %in% "type"))])
12  result <- as.data.frame(cbind(eval, classified))
13 END;
14 CALL USE_SVM("spamTraining", "spamEval", "spamClassified") WITH OVERVIEW;
15 SELECT * FROM "spamClassified";
```

Listing 2: Aufrufen einer R-Prozedur [SAP14b]

Eine R-Prozedur kann jedoch auch als Teil einer übergeordneten SQLScript-Prozedur gestartet werden. Im Codebeispiel Listing 3 wird dies noch einmal deutlicher dargestellt. Der Beginn ist dabei analog zum Codebeispiel 2, in der die Prozedur *USE_SVM* definiert wird. In diesem Falle werden die zu verarbeitenden Daten so umarangiert, dass sie eine verbessertes Analysetraining bieten. Anschließend wird innerhalb der SQLScript-Methode (siehe Attribut *LANGUAGE SQLSCRIPT*) die vorher definierte R-Methode *USE_SVM* gestartet. Somit können mehrere Aufgaben zu einer zusammengefasst werden, wodurch die gegebenen Ressourcen besser genutzt werden können.

```
1  DROP TABLE "spamClassified";
2  CREATE COLUMN TABLE "spamClassified" LIKE "spamEval" WITH NO DATA;
3  ALTER TABLE "spamClassified" ADD ("classified" VARCHAR(5000));
4  DROP PROCEDURE USE_SVM;
5  CREATE PROCEDURE USE_SVM( IN train "spamTraining", IN eval "spamEval",
6  OUT result "spamClassified")
7  LANGUAGE RLANG AS
8  BEGIN
9   library(kernlab)
10  model <- ksvm(type~. , data=train, kernel=rbfdot(sigma=0.1))
11  classified <- predict(model, eval [,-(which(names(eval) %in% "type"))])
12  result <- as.data.frame(cbind(eval, classified))
13 END;
14 DROP PROCEDURE R_PARTOFMORE;
15 CREATE PROCEDURE R_PARTOFMORE(OUT result "spamClassified")
16 LANGUAGE SQLSCRIPT AS
17 BEGIN
18  subset1 = select * from "spamEval" where "capitalLong" > 14;
19  subset2 = select * from "spamEval" where "capitalLong" <= 14;
20  train = select * from "spamTraining";
21  newtrain = CE_UNION_ALL(:subset1, :train);
22  CALL USE_SVM(:newtrain, :subset2, result);
23 END;
```

```
24 CALL R_PARTOFMORE("spamClassified") WITH OVERVIEW;
25 SELECT * FROM "spamClassified";
```

Listing 3: Aufrufen einer R-Prozedur innerhalb einer SQLScript-Prozedur [SAP14b]

In den vorangegangenen Codebeispielen tauchte in den R-Prozeduren immer wieder der Ausdruck *model* auf. Mithilfe solcher Modelle können Analyseprozesse nach Belieben spezifiziert und eine bestimmte Abhandlung der überreichten Daten gewährleistet werden. Das Codebeispiel Listing 4 zeigt dabei, wie ein solches Model beim Prozeduraufruf angewendet werden kann. Dabei muss das Model einerseits im Präambel der SQL-Prozedur (Form: IN modelname MODELTYP) und andererseits im eigentlichen R-Code definiert werden. Um ein Model permanent zu speichern und dieses auch in den Spezifikationen zu trainieren, werden Trainingsprozeduren genutzt. Im Codebeispiel Listing 5 wird dies beispielhaft anhand von *SPAM_MODEL* verdeutlicht. Dabei wird nach Erstellung des Models dieses mit passenden Daten angereichert, um eine spezifzierte Abarbeitung in späteren Vorgängen zu gewährleisten.

```
1  DROP TABLE "spamClassified";
2  CREATE COLUMN TABLE "spamClassified" LIKE "spamEval" WITH NO DATA;
3  ALTER TABLE "spamClassified" ADD ("classified" VARCHAR(5000));
4  DROP PROCEDURE USE_SVM;
5  CREATE PROCEDURE USE_SVM(IN eval "spamEval", IN modeltbl SPAM_MODEL_T, OUT
6   result "spamClassified")
7  LANGUAGE RLANG AS
8  BEGIN
9    library(kernlab)
10   svmModel <- unserialize(modeltbl$MODEL[[1]])
11   classified <- predict(svmModel, eval [,-(which(names(eval) %in% "type"))])
12   result <- as.data.frame(cbind(eval, classified))
13 END;
14 CALL USE_SVM("spamEval", SPAM_MODEL, "spamClassified") WITH OVERVIEW;
15 SELECT * FROM "spamClassified";
```

Listing 4: Benutzen eines Models aus R in SAP HANA [SAP14b]

```
1  DROP TYPE SPAM_MODEL_T;
2  CREATE TYPE SPAM_MODEL_T AS TABLE (
3    ID INTEGER,
4    DESCRIPTION VARCHAR(255),
5    MODEL BLOB
6  );
7  DROP TABLE SPAM_MODEL;
8  CREATE COLUMN TABLE SPAM_MODEL (
```

```
 ID INTEGER,
 DESCRIPTION VARCHAR(255),
 MODEL BLOB
);
DROP PROCEDURE SAPM_TRAIN_PROC;
CREATE PROCEDURE SPAM_TRAIN_PROC (IN traininput "spamTraining", OUT
    modelresult
SPAM_MODEL_T)
LANGUAGE RLANG AS
BEGIN
 generateRobjColumn <- function(...){
 result <- as.data.frame(cbind(
 lapply(
 list(...),
 function(x) if (is.null(x)) NULL else serialize(x, NULL)
 )
 ))
 names(result) <- NULL
 names(result[[1]]) <- NULL
 result
 }
 library(kernlab)
 svmModel <- ksvm(type~. , data=traininput, kernel=rbfdot(sigma=0.1))
 modelresult <- data.frame(
 ID=c(1),
 DESCRIPTION=c("SVM_Model"),
 MODEL=generateRobjColumn(svmModel)
 )
END;
CALL SPAM_TRAIN_PROC("spamTraining", SPAM_MODEL) WITH OVERVIEW;
select * from SPAM_MODEL;
```

Listing 5: Speichern und Trainieren eines Models aus R in SAP HANA [SAP14b]

2.2 Excel

Die Integration von Microsoft Excel in SAP HANA wird mithilfe der MDX-Schnittstelle beider Programme bewerkstelligt [Wal13]. MDX bedeutet *Multidimensional Expressions Language* und ist im Kern eine Art multidimensionales SQL, dass primär für OLAP Datenbanken genutzt wird [Wal13] [Moe11]. Die Verbindung wird dabei in der Regel über OLEDB oder OBDO realsiert [Moe11]. Für das Durcharbeiten der Daten in SAP HANA nutzt Excel sogenannte *Pivottables* [Wal13] [Moe11]. Diese erlauben es größere Datenmen-

gen, beispielsweise aus Datenbanken, auf verschiedenste Weise darzustellen, zu organisieren und zu analysieren ohne das diese bei Veränderungen der zu betrachtenen Spalten wieder komplett geladen werden müssen [Wal13] [Moe11].

Die Verbindung von SAP HANA mit Excel und das Füllen mit Daten geschieht dabei in drei Schritten [Wal13] [Moe11]. Zu Beginn muss in Excel eine Datenverbindung erstellt beziehungsweise spezifiziert werden [Wal13] [Moe11]. Dies geschieht in Excel über einen *Data Connection Wizard*, der den Nutzer durch die Verbindungserstellung führt [Moe11]. Für die ordnungsgemäße Verbindung von SAP HANA und Excel sollte der *SAP HANA MDX Provider*, der der primäre SAP HANA OLE DB Provider für die Verbndung von HANA mit Business Intelligence Tools ist, genutzt werden [Moe11]. Es ist hierbei jedoch anzumerken, dass andere SAP Produkte, wie SAP BW, ihre eigenen Provider haben und eine entsprechende Verbindung nicht zwangsläufig über den *SAP HANA MDX Provider* läuft [Moe11]. Abschließend wird in der Regel nach Login-Informationen gefragt, um die Verbindung zur passenden HANA Instanz mit den passenden Rechten abzuschließen [Wal13] [Moe11]. Der nächste Schritt ist zu entscheiden, welche Daten nun in Excel geladen werden sollen [Moe11]. Dabei werden im Assistenten die entsprechenden Datenwürfel und die gewünschte Ansicht ausgewählt [Moe11]. Im letzten Schritt wird die Pivottabelle mit den gewünschten Daten befüllt [Moe11] [Wal13]. Beim ersten Laden der Datenwürfel wählt man die zu betrachtenden Einheiten und Dimensionen, in denen auch die Hierarchien der Daten enthalten sind [Moe11]. Die enthaltenen Spalten können nach Belieben per Drag and Drop in die entsprechenden Zellen transportiert werden [Wal13] [Moe11], womit eine komplett benutzerdefinierte Ansicht auf die zu analysierenden Daten erreicht werden kann. Daten können hierbei nach Wunsch auch aggregiert angezeigt, nach bestimmten Kriterien gefiltert und die Eigenschaften der Einzelteile betrachtet werden [Moe11].

Die einzelnen Modelle und Ansichten der Daten in SAP HANA bekommen durch die MDX-Schnittstelle eine passende Repräsentation, mit der sie in externen Programmen, wie Excel, abgerufen und dargestellt werden. Dabei werden Analyseansichten sowie Kalkulationsansichten als Datenwürfel angezeigt und Attributansichten werden wiederum als Dimensionen dargestellt [Moe11]. Dabei bietet SAP HANA eine native Unterstützung von Hierarchien in Attributsansichten, was für eine verständlichere Ansicht der Daten sorgen kann [Moe11]. Hierbei ist es für die MDX-Schnitstelle irrelevant, ob es in einer Ebenenhierarchie oder einer Eltern-Kind-Hierarchie angeordnet ist [Moe11].

3 Analytische Möglichkeiten in SAP HANA

In diesem Kapitel sollen die analytischen Möglichkeiten, die sich durch die Integration von Analysetools, wie R und Excel, in SAP HANA ergeben, näher beleuchtet werden. Dazu gehören allgemeine Vorteile sowie Nachteile einer solchen Integration im Bezug auf die resultierenden analytischen Möglichkeiten. Anschließend werden die Analysemöglichkeiten des jeweiligen Tools erläutert und anhand eines Beispiels verdeutlicht. Dabei soll auch darauf eingegangen werden, bei welchen Situation beziehungsweise Problemstellungen sich welches Tool am ehesten zur Integration eignet.

3.1 Vor- und Nachteile einer Integration

Die Integration von verschiedenen Tools mit SAP HANA kann sowohl einen positiven als auch einen negativen Einfluss aufweisen.

Ein Vorteil einer solchen Integration findet sich im Anstieg der Verarbeitungsgeschwindigkeit, die ein Tool durch die Integration mit HANA, erreichen kann [RKA14] [SAP14a]. Dieser Anstieg ist im Vergleich zur Verbindung der Tools mit anderen Datenbanksystemen zu sehen. Vor allem sehr große Datenbestände können mithilfe der üblichen Tools ohne lange Analysezeiten verarbeitet werden [RKA14] [SAP14a]. Die Verbindung zwischen bekannten Tools und HANA hat außerdem den Vorteil, dass man die Vorteile von HANA auch in einer bereits bekannten Umgebung genießen kann [RKA14] [SAP14a] [Wal13]. Bereits mit dem Tool vertraute Mitarbeiter können weiterhin problemlos analysieren ohne auf komplett unbekanntes Territorium zu stoßen. Weiterhin eröffnet die Integratrion eines externen Tools, wie R oder Excel, zusätzliche Analysemöglichkeiten, Funktionen und Darstellungsmöglichkeiten [RKA14] [SAP14a] [Cun14] [Sim11]. So kann beispielsweise mithilfe von R ein WordCloud erstellt werden, wo eine größere Verkaufsmenge auch die Größe des Produkttitels anpasst [Asw12]. Auch können bestehende Analyse- und Darstellungsmöglichkeiten durch die Integration eines Tools benutzerdefinierter gestaltet werden [Cun14] [RKA14] [Sim11].

Es gibt auch Nachteile bei der Integration. So könnte der Arbeitsspeicher, den SAP HANA nutzt, mit passenden Mitteln ausgelesen werden [RKA14]. Eine Integration mit einem externen Tool, könnte dieses Sicherheitsproblem nochmals dadurch verschlimmern, dass Sicherheitslücken in den externen Tools dafür verwendet werden können kritische Daten aus HANA zu stehlen. Ein weiteres Problem, dass sich nach der Integration bemerkbar machen könnte, ist das „Flaschenhals“-Syndrom [RKA14]. Dies bedeutet, dass es im Betrieb des Systems aufgrund der Integration zu Engpässen bei der Rechenleistung beziehungsweise Verarbeitung kommt. Dabei ist jedoch zu beachten, dass andere SAP-Produkte

sowie R von Vornherein auf die Integration optimiert werden [RKA14] [SAP14a]. Auch die Schnittstellen, wie beispielsweise SAP HANAs MDX-Schnittstelle, werden so gut wie möglich zur Integration optimiert, wodurch Engpässe auch bei größeren Datenbeständen relativ unwahrscheinlich sind. Schlussendlich kann es situationsabhängig zu eventuellen zusätzlichen Kosten kommen, was ein weiterer Nachteil sein kann. Dies können unter anderem zusätzliche Kosten für Hardware oder Softwarelizenzen sein. Auch könnten für bestimmte Veränderungen in der Benutzung Schulungen nötig sein, die wiederum Kosten mit sich bringen [RKA14].

3.2 SAP HANA mit R

R wird sowohl alleinstehend als auch im Zusammenhang mit einem dritten Tool, wie SAP Predictive Analysis, verwendet [Asw12] [RKA14]. Dabei erweitert R die Funktionalität, die bereits durch SQL verfügbar ist [Asw12] [RKA14]. Weiterhin können bei R Anfragen optimierter formuliert werden, wodurch unnötige Mehrfachanfragen vermieden werden können und somit kann auch bei größeren Datenbeständen die schnelle Verarbeitung des In-Memory Computing gewährleistet werden [Asw12] [RKA14]. Die Vielzahl der Pakete beziehungsweise Bibliotheken, die in R zur Verfügung stehen, erlauben stärker an den Benutzer angepasste Analysen [Asw12] [RKA14]. Die Daten können dabei sowohl kalkulativ, beispielsweise eine Durchschnittsberechnung, als auch visuell, beispielsweise durch ein Kreisdiagramm, analysiert werden [Asw12] [RKA14]. Durch eine Zusammenarbeit von HANA und R können vor allem auch in sich komplexere Datenbestände einfacher veranschaulicht werden. Dieser Aspekt kann vor allem in der Forschung und Medizin genutzt werden, um komplexe Zusammenhänge zwischen Epedemien sowie deren Ursachen zu erkennen und zu analysieren. In Branchen, in denen eine Echtzeitauswertung wichtig ist, können HANAs schnelle Verarbeitung mit den spezifierten Bibliotheken aus R vereint werden und so an die Branche angepasste Methoden direkt genutzt werden [Asw12] [RKA14]. Dies führt dazu, dass auch minimalste Parametereingaben schon für die gewünschten Aufgaben ausreichen [Asw12] [RKA14]. Somit können branschenspezifische, aussagekräftige Analyseergebnisse effezient geholt und ausgewertet werden [Asw12] [RKA14]. Ein Beispiel für solche Branchen ist die Energiebranche [Asw12] [RKA14]. So können anhand der gespeicherten Windparkdaten Analysen zu zukünftigen meteoroligischen Veränderungen durchgeführt werden. Dies kann mithilfe von Durchschnitt und Standardabweichungsfunktionen geschehen [Asw12] [RKA14]. Weiterhin könnte man die Daten auf einem Graphen plotten, periodische Sequenzen erkennen lassen und anhand dieser den weiteren Verlauf bis zu einem gewissen Grad vorhersagen [Asw12] [RKA14]. Aus dieser Vorhersage kann

dann ein entsprechendes Angebot auf den Energiemarkt gebracht werden und Abweichungen rechtzeitig durch Ab- oder Zuschalten von Anlagen entgegengewirkt werden. Hierfür relevante R-Bibliotheken wären *forecast*, *RadioSonde* und *bReeze*, die allesamt Funktionen zur Winddatenaufbereitung, Darstellung sowie für prediktive Analysen bieten [RKA14]. Weitere visuelle Analysemöglichkeiten, die Funktionalitäten von HANA durch R erweitern, sind Clustering, WordClouds und Netzwerkgraphen [Asw12] [RKA14]. Die üblichen Möglichkeiten, wie Linien-, Kreis-, Streu-, Balken-, Säulen- und Kurvendiagramme sind auch vorhanden und können unter anderem in ihrer Skalierung weiter angepasst werden [Asw12] [RKA14]. Insgesamt ist R in Kombination mit SAP HANA vor allem dann vom großen Nutzen, wenn branchenspezifische Analysen regelmäßig gemacht werden müssen, Merfachanfragen vermieden werden sollen und die benötigten Parameter möglichst gering gehalten werden sollen.

3.3 SAP HANA mit Excel

Wie bereits in der Einführung angesprochen wurde, ist Excel ein verbreitetes BI-Tool (Business Inteligence) [Wal13] [Moe11] [Sim11]. Eine direkte Verbindung dieses BI-Tools mit HANA erlaubt eine schnelle, effiziente Analyse von Daten in einer familiären Umgebung [Wal13] [Moe11] [Sim11]. Dies wird durch eine assistentengestützte Menüführung unterstützt, sodass auch ohne große Vorkenntnisse die Daten, die aus HANA ausgelesen werden, erkundet und analysiert werden können [Wal13] [Moe11] [Sim11]. Excel bietet hierfür zahlreiche Berichterstattungsfunktionen, die Unternehmen bei der Analyse größerer Datenmengen unterstützen können [Moe11] [Sim11]. Neben bereits bekannten kalkulativen Analysen, wie beispielsweise die selektive Summierung von Daten mit bestimmten Eigenschaften, können auch verschiedene visuelle Analysen, wie Liniendiagramme, aus den Daten erstellt und auch im Nachhinein direkt an den gewünschten Datenrahmen angepasst werden [Moe11] [Sim11] [Cun14]. Die Ergebnisse können hierbei direkt wieder auf SAP HANA übertragen werden, da beide die MDX-Schnittstelle problemlos interpretieren können [Moe11] [Sim11]. Einschränkungen ergeben sich primär in den Funktionen, die sich außerhalb der Schnittmenge von HANA und Excels unterstützten Funktionen befinden [Moe11]. Ein Praxisbeispiel für die Nutzung von Excel mit HANA ist die Analyse von Verkaufszahlen bestimmter Produktgruppen in einem bestimmten Zeitraum. Bei Visualisierung dieser Daten können entsprechend besondere Abweichungen schneller erkannt und der Rest ohne Neuberechnung ausgeblendet werden. Weiterhin kann der hervorgehobene Graph weiter analysiert werden, beispielsweise durch das Hineinzoomen in ein Teilintervall des Gesamtgraphen [Moe11] [Sim11]. So können Trends beim Verkauf herauskristallisiert

werden und die Unternehmensstrategie im Allgemeinen und die Vertriebsstrategien im Speziellen angepasst werden, sodass das Unternehmen sich stärker auf profitable Geschäftsbereiche konzentrieren kann. Excel unterstützt visuelle Analysen in Form von Linien-, Kreis-, Streu-, Balken-, Säulen- und Kurvendiagramme [Moe11] [Sim11] [Cun14]. Je nach Diagramm können durch Anklicken bestimmter Punkte beziehungsweise Abschnitte weitere Informationen eingeblendet werden, um die Korrelation zwischen der Ursache und den Folgen besser zu erkennen [Moe11] [Sim11]. Alle wichtigen Analysen können anschließend in einen entsprechenden Bericht zusammengefasst werden, was an die verschiedenen Ebenen des Unternehmens gereicht werden kann, um so die Entwicklung und Anpassung von Strategien zu erleichtern [Moe11] [Sim11]. Insgesamt eignet sich Excel in Zusammenhang mit SAP HANA vor allem für Aufgaben aus dem Bereich des Business Intelligence. Sowohl Trendanalysen als auch die Indentifizierung nicht lohnender Sparten, beispielsweise anhand von Verkaufsanalysen oder zusätzlich gekauften Artikeln, können mit SAP HANA schneller ausgeführt werden, während Excel verschiedene Funktionen zur eigentlichen Durchführung der Analyse bietet.

4 Fazit und Ausblick

Alles in allem kann eine Integration für ein Unternehmen in vielerlei Hinsicht vorteilhaft sein. Neben schnellerer Analysen in bereits bekannten Umgebungen, erweitert die Integration von externen Tools SAP HANA um weitere Optionen zu kalkulativen und visuellen Analyse. Diese Optionen können unter anderem zusätzliche Darstellungsmöglichkeiten, wie weitere Visualisierungen oder verbesserte Filterung der bestehenden Daten, umfassen. Durch die große Anzahl an Paketen, die in R verfügbar sind, sowie deren stetiger Anstieg, können Unternehmen ihre Analysen noch benutzerdefinierter gestalten und die Möglichkeiten, die SQL bietet, spezifiziert erweitern, was Unternehmen in der Steigerung ihrer Produktivität, Effizienz und Effektivität unterstützen kann. Die Analyseergebnisse können durch Tools, wie Excel und R, insoweit aufbereitet werden, dass diese Ergebnisse in allen zugehörigen Bereichen ohne große Schwierigkeiten verstanden werden. Es ist jedoch zu beachten, dass die Sicherheitsprobleme von SAP HANA durch eine Integration von externen Tools verstärkt werden, da SAP HANA durch Sicherheitslücken der externen Tools angreifbarer gemacht wird. Dies ist einer der Gründe, warum Unternehmen bei der Entscheidung zur Einführung oder Integration eines solchen Systems zögern könnten.

Mögliche Entwicklungen für die Zukunft einer Integration externer Tools in SAP HANA können sich in vielerlei Hinsicht auftun. So kann die Integration externer Tools zur größeren Akzeptanz führen, da bereits bekannte Umgebungen dafür sorgen, dass die Hemm-

schwelle für Unternehmen sowie deren Mitarbeiter erniedrigt wird. Dies folgt primär aus dem Aspekt, dass sich diese nur geringfügig umgewöhnen müssten und somit bestimme negative Aspekte entfallen. Ein weitere Entwicklung ist die Verbesserung der Integrationsunterstützung in SAP HANA. Damit ist gemeint, dass bestehende Integrationsmöglichkeiten, wie R und Excel, in der Menge an unterstützten Funktionen zunehmen und weitere Integrationsmöglichkeiten in Form von BI-Tools, die nicht von SAP produziert werden, entwickelt werden.

Literatur

[Asw12] Aswani J. und Doerpmund J. Advanced Analytics with R and HANA. http://de.slideshare.net/JitenderAswani/na-6693-r-and-sap-hana-dkom-jitenderaswanijensdoeprmund Zuletzt aufgerufen: 18.05.2014, 2012.

[Cun14] Cundus. SAP Business Objects Reporting mit Crystal Reports. http://www.cundus.com/de-de/technologie/sap-bi-beratung/sap-businessobjects/sap-businessobjects-reporting-analyse/ Zuletzt aufgerufen: 18.05.2014, 2014.

[Dol04] Dubravko Dolic. *Statistik mit R - Einführung für Wirtschafts- und Sozialwissenschaftler*. Oldenbourg Verlag, München, 2004.

[Moe11] Amyn Rajan; Darryl Eckstein; George Chow; King Long Tse; Roman Moehl. Connecting to SAP HANA with Microsoft Excel 2007 PivotTables and ODBO. http://www.sdn.sap.com/irj/scn/go/portal/prtroot/docs/library/uuid/e03fef5e-d82f-2f10-8898-859c4ed57e62?quicklink=index&overridelayout=true Zuletzt aufgerufen: 18.05.2014, November 2011.

[RKA14] Tomasz Rudny, Monika Kaczmarek, and Witold Abramowicz. *Analytical Possibilities of SAP HANA: On the Example of Energy Consumption Forecasting*. Advances in Intelligent Systems and Computing. Springer International Publishing, 2014.

[SAP14a] SAP AG. SAP HANA Master Guide. http://help.sap.com/hana/SAP_HANA_Master_Guide_en.pdf Zuletzt aufgerufen: 18.05.2014, 2014.

[SAP14b] SAP AG. SAP HANA R Integration Guide. https://help.sap.com/hana/SAP_HANA_R_Integration_Guide_en.pdf Zuletzt aufgerufen: 18.05.2014, 2014.

[Sch07] Udo Schweitzer. *Statistik mit Microsoft Excel* -. W3l GmbH, Witten, 1. Aufl. edition, 2007.

[Sim11] Simba Technologies Inc. Using Excel with SAP HANA. https://www.youtube.com/watch?v=tu4ULFHbw7s Zuletzt aufgerufen: 18.05.2014, 2011.

[Wal13] Mark Walker. SAP HANA Starter: SAP HANA Integration with Microsoft Excel. http://www.packtpub.com/article/

`sap-hana-integration-with-microsoft-excel` Zuletzt aufgerufen: 18.05.2014, 01 2013.

Abschließende Erklärung

Ich versichere hiermit, dass ich meine Seminararbeit „Analytical capabilities of SAP HANA: Integration with R & Excel“ selbständig und ohne fremde Hilfe angefertigt habe, und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Oldenburg, den 30. März 2015 Eduard Rajski

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Statistische Verfahren zur Fortschreibung historischer Daten

Seminararbeit
im Rahmen der Projektgruppe „inMemory Planung mit SAP HANA“

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dipl.-Math Jens Siewert
Vorgelegt von:	Daniel Stratmann daniel.stratmann@uni-oldenburg.de
Abgabetermin:	02. August 2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

Tabellenverzeichnis **3**

1 Einleitung **4**

2 Der KDD-Prozess **5**

3 Data Mining **7**
3.1 Techniken und Methoden des Data Mining 7

4 Stat. Verfahren zur Fortschr. hist. Daten **9**
4.1 Zeitreihenanalyse 9
4.2 Elman-Netze 11
4.3 M5'-Modellbaumverfahren 14

5 Fazit **18**

Literaturverzeichnis **20**

Abbildungsverzeichnis

1 Der KDD-Prozess [FPSS96a] 6
2 Modell eines künstlichen Neurons [Lei13] 11
3 Beispielstruktur eines Elman-Netzes (Vgl. [Pet05]) 13

Tabellenverzeichnis

1 Komponentenmodell der Zeitreihenanalyse (Vgl. [Lip93]) 10

1 Einleitung

Eine immer größer werdende Herausforderung für die Informationsgesellschaft im Allgemeinen und Unternehmen im Speziellen ist die Generierung neuen Wissens [RK96]. Täglich werden Millionen Gigabytes Daten im World Wide Web beziehungsweise von Data Storage Devices erhoben und verarbeitet [FPSS96a]. In diesen Datenbeständen steckt ein großes Potential, denn sie enthalten implizites Wissen, welches den verschiedenen Akteuren als Orientierungshilfe in Entscheidungs- und Planungsprozessen dienen oder Auswirkungen getroffener Entscheidungen vorhersagen kann [LLS10]. Eine Anwendung dieses Wissens kann zu mehr Transparenz in Planungsprozessen führen, indem Entscheidungen auf Grundlage objektiver Tatsachen und nicht auf subjektiven Einschätzungen einzelner Akteure getroffen werden. Im Rahmen der Projektgruppe „In-Memory Planung mit SAP Hana“ soll ermittelt werden, wie durch den Einsatz von In-Memory-Technologien sowie Planungs- und Prognosewerkzeugen eine bessere Simulation zukünftiger Auswirkungen auf heute zu treffende Entscheidungen unterstützt werden kann. Ein wichtiger Teilaspekt dieser Frage- bzw. Aufgabenstellung ist dabei die Anwendung von Prognoseverfahren auf sehr große Mengen von historischen Daten, wodurch mögliche zukünftige Entwicklungen eines Sachverhaltes ermittelt werden können. Damit soll der Entscheidungsträger mit den nötigen objektiven Informationen für eine optimal zu treffende Entscheidung aus einer Menge von Entscheidungsalternativen ausgestattet werden [WY05]. Demnach liefern Prognoseverfahren die integralen Bestandteile einer transparenten und auf objektiven Tatsachen beruhenden Entscheidungsfindung. Diese Seminararbeit beschäftigt sich detailliert mit zwei Prognoseverfahren des maschinellen Lernens, namentlich Elman-Netze und M5'-Modellbaumverfahren. Beide Verfahren ermöglichen Prognosen auf Basis historischer Daten und sind dem Data Mining zuzuordnen. Unter Data Mining versteht man die Anwendung verschiedenster, computergestützter Techniken um verborgene Muster oder Zusammenhänge zwischen Datenbeständen zu finden, mit deren Hilfe anschließend neues Wissen oder Regeln abgeleitet werden können [HK00]. Data Mining wiederum ist ein Teilschritt des Prozesses der Wissensgewinnung in Datenbanken (KDD-Prozess) [1]. Dieser Prozess bietet Anwendern eine systematische Vorgehensweise zur Wissensgewinnung, indem alle relevanten Schritte – von der Selektion der Daten bis zur Interpretation und Evaluation der gefundenen Muster und Regeln – abgebildet werden.

Um die Zusammenhänge zwischen KDD-Prozess, Data Mining und den genannten Prognosetechniken zu verdeutlichen, werden im folgenden Kapitel zunächst der KDD-Prozess und seine Teilschritte beschrieben. Kapitel 3 geht anschließend detaillierter auf den Teilschritt Data Mining und die darin enthaltenen Methoden zur Wissensgewinnung in Daten ein. In Kapitel 4 werden Grundlagen zur Analyse historischer Daten aufgearbeitet; darauf aufbauend erfolgt die detaillierte Beschreibung des Elman-Netzes und des M5'-Modellbaumverfahrens. Im Fazit erfolgt eine kurze Einordnung der vorgestellten Verfahren hinsichtlich betrieblicher Planungsprozesse und es wird ein Ansatz zur Verbesserung der Prädiktionsgenauigkeit der Verfahren benannt.

[1] engl. Knowledge Discovery in Data.

2 Der KDD-Prozess

In der Literatur wird der Begriff Knowledge Discovery in Databases auf verschiedene Arten erklärt. Oftmals wird auch die Bezeichnung des Data Mining stellvertretend für den gesamten Prozess verwendet. Um eine einheitliche Bezeichnung des Begriffes zu gewährleisten, folgt diese Arbeit der Definition von [FPSS96b]:

> „KDD is the nontrivial process of identifying valid, novel, potentielly useful, and ultimately understandable pattern in data."

Es handelt sich hierbei um den vollständigen, mehrstufigen, nichttrivialen Prozess der Gewinnung von Modellen und Strukturen aus einer Mengen von Daten, deren Auswertung bisher unbekanntes, verständliches und nützliches Wissen liefert [Lei13].

Abbildung 1 skizziert die einzelnen Schritte des KDD-Prozesses. Neben dem zentralen Schritt des Data Mining werden zuvor die Phasen Datanselektion, -vorbereitung, -transformation und nachfolgend die Interpretation und Bewertung der gefundenen Muster durchlaufen. Aus dem Schaubild wird ersichtlich, dass der Prozess iterativ aufgebaut ist. Das heißt, dass jede Phase beliebig oft durchlaufen werden kann. Ebenso kann von einer Phase zur nächsten als auch zur vorherigen Phase zurückgesprungen werden: Das - subjektiv empfundene - beste Ergebnis eines Teilschrittes dient als Basis für die Durchführung der nächsten Phase. Insbesondere kann am Ende des KDD-Prozesses in beliebige vorherige Phasen zurückgesprungen werden. Hiermit kann eine Ergebnisverbesserung erzielt werden, falls die gefundenen Muster die Aufgaben- beziehungsweise Zielsetzung nicht zufriedenstellend lösen. Auch ein erneutes Durchlaufen des KDD-Prozesses mit bereits gefundenen Muster/Wissen ist denkbar, um damit genauere oder zusätzliche Ergebnisse zu erhalten [FPSS96a]. Im Folgenden werden die zentralen Funktionen und Aufgaben der einzelnen Teilschritte des KDD-Prozesses erläutert.

Unternehmensdomäne kennenlernen: [2] Im Initialschritt werden Hintergrund und Zielsetzung der Anwendung des KDD-Prozesses aus der Kundensicht definiert. Beispielsweise wird hier herausgearbeitet, welches Wissen der KDD-Prozess generieren soll. Des Weiteren erhalten die Experten alle nötigen Meta-Informationen, die zur Durchführung des KDD-Prozesses erforderlich sind. [3]

Datenselektion: Anhand der im vorigen Schritt definierten Zielsetzung werden diejenigen (Roh-) Daten aus den (heterogenen) Datenquellen selektiert, die für die Generierung des gewünschten Wissens als relevant angesehen werden. Zusätzlich wird eine geeignete Datenbasis[4] erstellt, welche als Basis für die folgenden Schritte dient.

[2] Nicht in der Abbildung enthalten, wird jedoch in [FPSS96a] als Initialschritt benannt.
[3] Zum Beispiel welche Daten in welchen Datenbanken abgelegt sind.
[4] Zum Beispiel in Form eines Data Warehouses [FPSS96a].

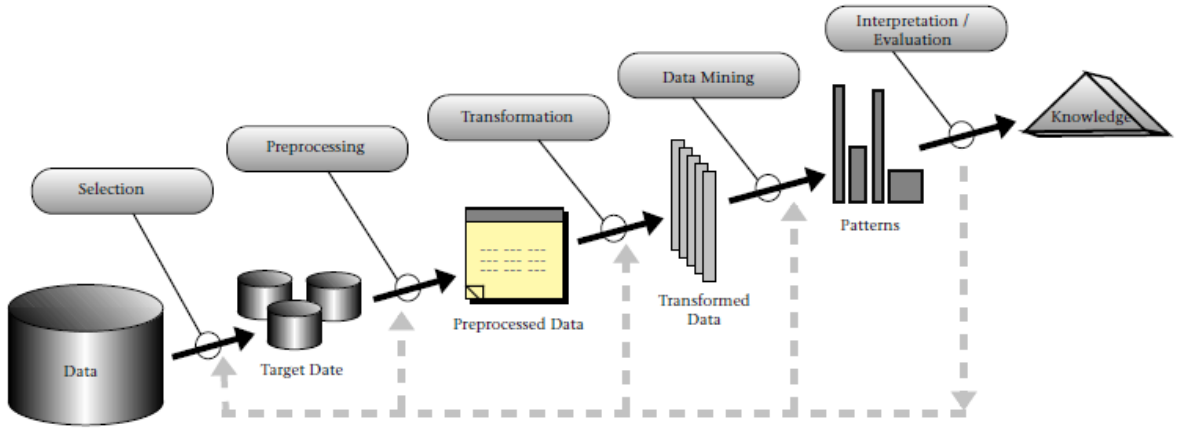


Abbildung 1: Der KDD-Prozess [FPSS96a]

Datenvorverarbeitung: In dieser Phase werden die Rohdaten auf Inkonsistenzen, Vollständigkeit, Redundanzen sowie Datenkonflikte geprüft und bereinigt. Diesem Schritt muss eine hohe Priorität beigemessen werden, da fehlerhafte Datentupel das spätere Ergebnis des Data Mining stark verfälschen, die Interpretation der Muster erschweren oder im schlimmsten Fall sogar falsche Ergebnisse hervorbringen können [ZZY03]. Gegenfalls werden die Datenbestände um weitere Attribute angereichert, wenn diese für die Lösung der Aufgabenstellung erforderlich sind oder es werden Attribute entfernt, sofern sie für die Zielsetzung kein Belang haben.

Datentransformation: Die von Anomalien bereinigten Datentupel werden in diesem Schritt in die Datenbasis aus Schritt 2 eingefügt und integriert, so dass sie nach Abschluss dieser Phase in einem passenden Format für eine konkrete Data Mining Technik vorliegen.

Data Mining: Das Data Mining ist der zentrale Schritt des KDD-Prozesses. Hier werden die eigentlichen Muster aus den integrierten Datenbeständen durch die Anwendung verschiedener Algorithmen erzeugt.

Interpretation: In dieser Phase werden die gefundenen Muster und Strukturen anwendungsorientiert interpretiert, um somit neues Wissen zu generieren. Die Ergebnisse müssen den Entscheidungsträgern oftmals visuell dargestellt werden, da der Output der Data Mining Methoden keinen Mehrwert für sie darstellen. Zusätzlich gilt es, das gewonnene Wissen anhand früherer Erkenntnisse auf Konsistenz zu überprüfen und es entsprechend des Anwendungsgebietes zu nutzen oder zu dokumentieren [Lei13].

Die Vorstellung des KDD-Prozesses ist hiermit abgeschlossen. Das folgende Kapitel befasst sich detaillierter mit der Phase des Data Mining. In dieser Phase erfolgt die konkrete Anwendung von

Algorithmen auf die Daten, zu denen auch die bereits benannten Prognoseverfahren gehören.

3 Data Mining

Data Mining ist nach [FPSS96b]

> „...a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumaration of patterns (or models) over the data.“

Data Mining ist der maßgebliche Schritt des KDD-Prozesses. Hier wird ein sorgfältig und zielorientiert ausgewählter Algorithmus auf die zuvor selektierten und präparierten Daten angewendet, um daraus neue Muster, Strukturen oder Kenngrößen und somit Wissen zu gewinnen [Lei13]. Data Mining Algorithmen sind Methoden des maschinellen Lernens. Dazu werden Computer so programmiert, „dass ein bestimmtes Leistungskriterium anhand von Beispieldaten oder Erfahrungswerten aus der Vergangenheit optimiert wird“ [Alp08]. Im Gegensatz zur Statistik werden beim Data Mining gegebene Hypothesen jedoch nicht bestätigt, vielmehr werden durch die Anwendung der in Kapitel 3.1 vorgestellten Techniken und Methoden neue Hypothesen automatisch aufgestellt [AN00] [Cha13]. Diese Hypothesen müssen anschließend von Experten interpretiert und evaluiert werden. Man unterscheidet verschiedene Arten von Methoden und damit auch verschiedene Arten von Informationen, die mit Data Mining Techniken gewonnen werden. Diese Methoden lassen sich grob in folgende Kategorien unterscheiden: Klassifizierungen, Sequenzen, Assoziationen, Cluster und Prognosen [LLS10]. Der folgende Abschnitt befasst sich näher mit den benannten Methoden und den daraus resultierenden Informationen.

3.1 Techniken und Methoden des Data Mining

Jede Data Mining Methode wird anhand eines kurzen Beispieles konkretisiert und es werden entsprechende Techniken für jede Methode benannt.

Klassifikation Hauptziel der Klassifikation ist die Einordnung von (neuen) Objekten in zuvor definierte Klassen. Dazu werden auf Basis historischer Daten[5] Muster und Regeln abgeleitet, welche die jeweiligen Klassen möglichst genau beschreiben. Anschließend können Objekte, deren Klassenzugehörigkeit unbekannt ist, mit den gefundenen Mustern und Regeln einer Klasse zugeordnet werden. Beispiel: Einem Unternehmen wandern Kunden ab. Mit Hilfe von Klassifizierungstechniken können die Eigenschaften der abgewanderten Kunden (Trainingsdaten) analysiert werden. Mit diesem Wissen können weitere Kunden identifiziert werden, welche wahrscheinlich abwandern wollen. Mit diesem Wissen können Entscheidungsträger Gegenmaßnahmen einleiten, um diese Kunden zu halten. Klassifizierungstechniken ermöglichen demnach ebenfalls eine Art Prädiktion, indem sie das Verhalten von Objekten auf Basis bestimmter Eigenschaften vorhersagen. Jedoch ermöglichen Klassifizerungstechniken nicht die Vorhersage von

[5] Im Data Mining Kontext oft auch Trainingsdaten genannt [Alp08].

unbekannten Variablen oder Trendentwicklungen[6]. Konkrete Klassifizierungstechniken sind: der naive Bayes Klassifikator, Entscheidungsbäume oder neuronale Netzwerke [HK00] [LLS10].

Cluster Beim Clustering erfolgt ebenfalls eine Zuordnung von Objekten mit ähnlichen Eigenschaften in Klassen. Jedoch sind die Klassen sowie (teilweise)[7] deren Anzahl zunächst nicht bekannt. Dies wird von den Techniken des Clustering selbst vorgenommen. Das Ziel des Clustering ist also nicht nur die Zuordnung von Objekten zu Klassen, sondern auch die Definition dieser. In diesem Kontext spricht man auch von „uninformierten Techniken“, da sie nicht auf Klassen-Vorwissen angewiesen sind. Beispiel: Ein Unternehmen könnte seine Kundendatenbank segmentieren um mit diesen Wissen auf das jeweilige Kundensegment angepasste Marketingmaßnahmen durchzuführen. Aus dem Beispiel geht hervor, dass Clusteringtechniken Datensätze beschreiben, ohne dabei Vorhersagen zu treffen. Demnach sind Clusteringtechniken den deskriptiven Data Mining Methoden zuzuordnen. Konkrete Clusteringtechniken sind der k-means-Algorithmus, Multiview-Clustering oder Self-Organizing Maps [WF99] [ES00].

Sequenzen Sequenzen analysieren das Verhalten von Objekten über die Zeit.[8] Dabei sollen diejenigen Sequenzen gefunden werden, welche am häufigsten in den Trainingsdaten vorkommen. Beispiel: Das Klickverhalten von Kunden auf einem Webshop könnte analysiert werden. Mit diesen Daten kann festgestellt werden, welche Seitenbesuche am häufigsten zum Kauf eines bestimmten Produktes führen. Außerdem lässt sich vorhersagen, welche Seite mit der höchsten Wahrscheinlichkeit als Nächstes besucht werden. Sequenztechniken ermöglichen demnach ebenfalls Prädiktion - in Abhängigkeit der zeitlichen Reihenfolge von Ereignissen - und ist damit eng verwandt mit der in Kapitel 4 vorgestellten Zeitreihenanalyse. Konkrete Sequenztechniken sind der GSP- und der PrefixSpan-Algorithmus [Duc08] [Coo14].

Assoziationen Assoziationstechniken suchen innerhalb jedes Trainingsdatensatzes nach Ereignissen, die andere Ereignisse innerhalb des Datensatzes implizieren. Wird eine solche Suche über eine große Menge von Datensätzen durchgeführt, lassen sich so genannte „starke Regeln“ identifizieren, die aussagen, dass mit einer gewissen Wahrscheinlichkeit auf ein Ereignis A auch Ereignis B eintritt. Ein weit verbreiteter Anwendungsbereich solcher Techniken ist die Warenkorbanalyse: Dabei wird untersucht, welche Produkte gleichzeitig mit anderen Produkten gekauft werden. So könnte beispielsweise festgestellt werden, dass in 80% aller Bierkäufe ebenfalls Kartoffelchips eingekauft werden. Diese Informationen werden anschließend oftmals als Grundlage für die räumliche Anordnung von Produkten im Verkaufsbereich oder für das Cross-Marketing im Internet genutzt. Auch dieses Verfahren unterstützt in gewisser Weise die Prädiktion, da zukünftige Ereignisse auf Basis von Trainingsdaten vorhergesagt werden. Typische Assoziationstechniken sind der AIS-Algorithmus sowie Apriori-Algorithmus mit seinen Varianten [LLS10] [Lü10].

[6] Das ist den Prognosetechniken des Data Mining, z.B. Künstliche Neuronale Netze oder Modellbaumverfahren, vorbehalten [Pet05].
[7] In Abhängigkeit des verwendeten Algorithmus.
[8] Der Tagesablauf eine Menschen stellt zum Beispiel eine Sequenz von Tätigkeiten dar.

Prognose Ähnlich der Klassifikation sollen bei den Prognosetechniken des Data Mining anhand einer Eingabe eine entsprechende Ausgabe erzeugt werden. Bei den Prognosetechniken besteht die Ausgabe Y jedoch aus nummerischen Werten. Das bedeutet, es muss eine Funktion erlernt werden, die alle Trainingsdaten widerspiegelt und eine möglichst genaue Berechnung des Ausgabewertes für unbekannte Daten liefert. Hilfsmittel hierfür sind die Zeitreihen- sowie die Regressionsanalyse. Beispiel: Auf Grundlage historischer Verkaufsdaten eines Produktes soll ein Trend der Entwicklung des Absatzes des Produktes für das zukünftige Quartal berechnet werden. Konkrete Data-Mining-Methoden, die solche Berechnungen ermöglichen sind beispielsweise Elman-Netze und das M5'-Modellbaumverfahren. Beide Methoden werden im folgenden Abschnitt im Detail behandelt.

4 Statistische Verfahren zur Fortschreibung historischer Daten

Dieses Kapitel beschäftigt sich mit Prognosetechniken des Data Mining. Zunächst werden die notwendigen Grundlagen der Zeitreihenanalyse[9] vorgestellt. Anschließend werden zwei Möglichkeiten zur Umsetzung von Prognosen mit Verfahren des Data Mining vorgestellt. Diese sind das Elman-Netz, ein künstliches neuronales Netz, welches Zeitreihenanalysen ermöglicht sowie der M5'-Algorithmus, ein Modellbaumverfahren, welches mit Hilfe von Regressionstechniken Prognosen erstellt.

4.1 Zeitreihenanalyse

Im Gegensatz zu den im Kapitel 3.1 vorgestellten Methoden der Klassifizierung umfassen Prognose- bzw. Zeitreihenanalyseverfahren Techniken, mit denen es möglich ist, Modelle zur Prognose eines nummerischen Wertes zu entwickeln. Dies können künftige Entwicklungen, etwa Zinskurse, Umsatzzahlen, Temperaturwerte oder Durchlaufzeiten in Prozessen sein. Eine Zeitreihe besteht aus Beobachtungswerten für eine Variable, die meist zu äquidistanten Zeitpunkten erhoben wurde. Als Modell einer Zeitreihe kann eine Regel angesehen werden, die den stochastischen Prozess beschreibt, der die Beobachtungswerte erzeugt haben könnte. Ein stochastischer Prozess lässt sich als Folge $(X_t)_{t \in T}$ von Zufallsvariablen X_t definieren, wobei t Element der maximal abzählbaren Indexmenge T ist. Voraussetzung für die Nutzung eines derartigen Modells zur Prognose ist, dass sich der zugrunde liegende stochastische Prozess nicht verändert. Diese Eigenschaft wird als Stationarität bezeichnet. Das Vorgehen mathematisch-statistischer Zeitreihenanalyse umfasst drei Schritte [Pet05], [Lip93]:

1. Analyse: Charakterisierung und Aufspaltung der herausragenden Eigenschaften und Merkmale einer Zeitreihe (zum Beispiel Saisonschwankungen oder der Trendanteil).
2. Modellierung: Modellierung von Zeitreihen, mit deren Hilfe ein Modell zur Prognose künftiger Werte erstellt wird.

[9] Der Begriff Zeitreihenanalyse wird in der Literatur oftmals synonym für Prognosetechniken verwendet [Pet05].

3. Prognose: Prognostizieren von zukünftigen Werten, mit dem Ziel, der realen Entwicklung zu entsprechen.

Die Verfahren zur Zeitreihenanalyse lassen sich in lineare und nichtlineare und diese wiederum je in univariate und multivariate einteilen. Univariate Verfahren modellieren auf Basis einer Zeitreihe mit einer abhängigen Variablen, Lags[10] und Prognosehorizont[11] zur Prognose der abhängigen Variable, die in diesem Fall mit der unabhängigen identisch ist. Multivariate Verfahren modellieren auf Basis keiner und/oder einer und/oder mehrerer Zeitfenster von mehreren Zeitreihen jeweils mehrerer unabhängiger Variablen, Lags und Prognosehorizonte zur Prognose der abhängigen Variablen, die mit einer unabhängigen Variable zwar identisch sein kann, aber nicht zwingend sein muss. In diesem Zusammenhang stellt die Modellierung des Zusammenhangs von unabhängigen und abhängigen Variablen bezüglich einer oder mehreren Zeitreihen ein spezielles Zeitreihenanalyseproblem dar: Je nach Anwendung wird die Entwicklung einander beeinflussender Größen im Zeitverlauf in einem oder mehreren Modellen berücksichtigt [Pet05].

Neben der Vorhersage der zukünftigen Entwicklung von Werten zielt das Arbeitsfeld der Zeitreihenanalyse auch auf das Erkennen von Ursachen der bisherigen Entwicklung sowie dessen Beschreibung ab. Hierzu wird ein Modell mit Regeln konzipiert, welches den bisherigen Verlauf einer Zeitreihe möglichst genau beschreibt und zeigt, wie die beobachteten Werte produziert sein könnten. In diesem Kontext spielt das so genannte „Komponentenmodell“ der Zeitreihenanalyse eine wichtige Rolle: Es interpretiert eine Zeitreihe y_t als Überlagerung einfacher Funktionen der Zeit, die formal auf ihre Periodizität definiert sind und „Komponenten“ genannt werden [Lip93].

Komponente	**ökonomische Beschreibung**	**formale Beschreibung**
Trend	langfristige Niveauänderung, Wachstum, Entwicklung, Grundtendenz	monoton steigende bzw. fallende Funktion oder Polynom geringen Grades
Konjunktur	Schwankungen im Auslastungsgrad des Produktionspotentials	wurde früher (heute nicht mehr üblich) als zyklisch aufgefasst
Saison	jahreszeitliche (z. B. Klima) und institutionelle Einflüsse (z. B. Steuer-, Ferientermine)	weitgehend regelmäßige Schwingung mit einer Periode (Wellenlänge) von (genau) einem Jahr
Zufallskomponente	keine bekannten oder einmalige, nicht vorhersehbare Einflüsse	irreguläre Einflüsse

Tabelle 1: Komponentenmodell der Zeitreihenanalyse (Vgl. [Lip93])

Tabelle 1 zeigt das Komponentenmodell nach K. Pearson. Ziel der Modellbildung von Zeitreihen ist in diesem Zusammenhang die Zerlegung der Zeitreihe in die beschriebenen Komponenten. Dieses Wissen lässt anschließend wieder Rückschlüsse auf zukünftige Entwicklungen der Zeitreihe zu. Die Klassifizierung einzelner Beobachtungswerte hinsichtlich des Komponentenmodells entspricht einer Klassifizierung beim Data Mining. Für den speziellen Fall der Zeitreihenanalyse existieren bestimmte Verfahren, die den Faktor Zeit in ihre Berechnungen mit einfließen

[10] Lags bezeichnen den Zeitabstand der betrachtete Werte zueinander und sind ein Hilfsmaß für die Bestimmung des Zusammenhanges zwischen Beobachtungsdaten mit einem zeitlichen Bezug zueinander [Lue99].

[11] Der Prognosehorizont definiert den Zeitraum für den eine Prognose abgegeben wird [Wir14].

lassen. Ein konkretes Verfahren zur Umsetzung von Zeitreihenanalysen im Data Mining ist das Elman-Netz, welches im folgenden Abschnitt näher betrachtet wird.

4.2 Elman-Netze

Künstliche neuronale Netze (KNN) sind „ein sortiertes Tripel (N, V, w) mit zwei Mengen N, V sowie einer Funktion w, wobei N die Mengen der Neuronen bezeichnet und V eine Menge $\{(i, j)|i, j \in \mathbb{N}\}$ ist, deren Elemente Verbindungen von Neuron i zu Neuron j heißen. Die Funktion $w : V \rightarrow \mathbb{R}$ definiert die Gewichte, wobei $w((i, j))$, das Gewicht der Verbindung von Neuron i zu Neuron j, kurz mit $w_{i,j}$ bezeichnet wird.“[Kri05]. Dabei sind die Neuronen in Schichten angeordnet. Alle Neuronen einer Schicht sind mit allen Neuronen der nachfolgenden Schicht durch die erwähnten Kanten miteinander verbunden. Über diese Verbindungen werden Daten zwischen den einzelnen Neuronen ausgetauscht. Abbildung 2 zeigt den Datenverarbeitungsprozess eines Neurons j.

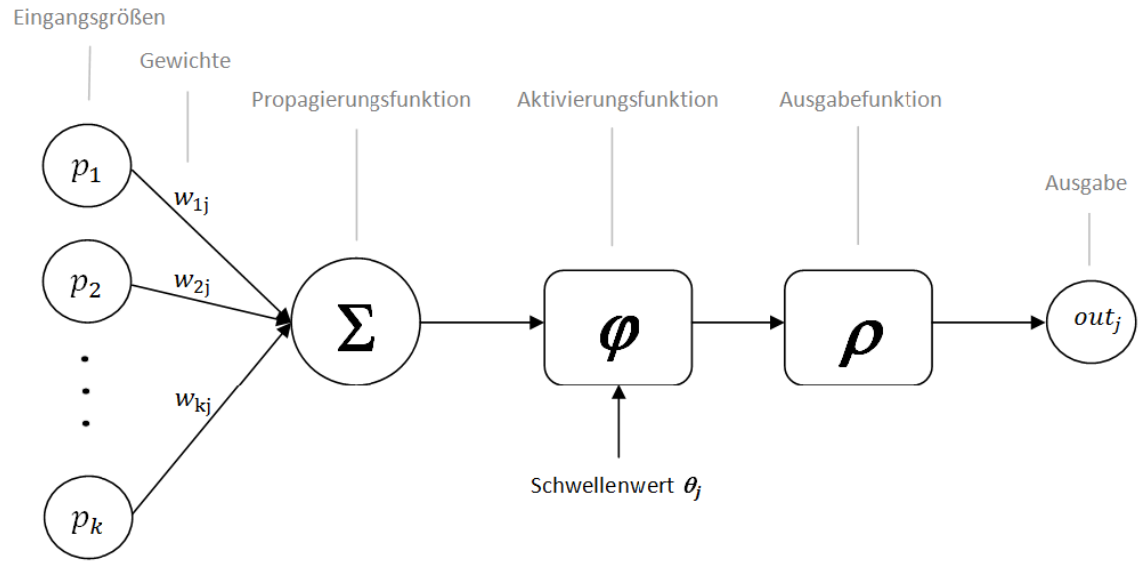


Abbildung 2: Modell eines künstlichen Neurons [Lei13]

Die Propagierungsfunktion nimmt für ein Neuron j Ausgaben $p_{i_1}, ...p_{i_n}, i \in \{1, ..., n\}$ anderer Neuronen $p_1, ...p_n$ entgegen und verarbeitet diese unter Berücksichtigung der Verbindungsgewichte $w_{i,j}$ zu einem skalaren Netzinput[12]. Dazu werden die Eingaben beispielsweise entsprechend ihrer Gewichtungen aufsummiert. Die Aktivierungsfunktion des Neurons j definiert, wie sich der aktuelle Zustand des Neurons zum Zeitpunkt t durch die Netzeingabe in einem anderen Aktivierungszustand zum Zeitpunkt $t + 1$ ändert. Ein Neuron wird erst aktiv, d.h. es produziert Ausgaben, wenn sein Schwellenwert Θ_j überschritten wird [13]. Der Schwellenwert definiert, wie empfindlich ein Neuron auf die Eingaben anderer Neuronen reagiert. Je höher der

[12] Der Netzinput stellt die skalaren Ausgaben der vorgelagerten Neuronen $p_1, ...p_n$ dar [Kri05]
[13] Initial wird dieser Schwellenwert mit einem Standardwert definiert und im Laufe der Trainingsphase durch den Trainingsalgorithmus angeglichen [Kri05].

Schwellenwert, desto unempfindlicher ist das Neuron und vice versa. Die Ausgabefunktion bildet anschließend den Aktivierungszustand auf einen zu definierenden Wertebereich ab. Ein einfaches Beispiel hierfür ist die binäre Ausgabefunktion, bei der lediglich die Ausgaben 1 und 0 erfolgen: Bei Unterschreiten des Schwellenwertes wird der Wert 0 ausgegeben, beim Überschreiten des Schwellenwertes wird der Wert 1 ausgegeben. In Abbildung 2 stellt out_j das skalare Ausgabesignal des Neurons j dar [Lei13].

KNN verfolgen den Ansatz des selbstständigen Lernens. Dazu muss das Netz vor dem tatsächlichen Einsatz trainiert werden, indem ihm Instanzen von Trainingsdaten so lange präsentiert werden, bis es die gewünschten Ausgaben produziert [Pet05]. Dabei „lernt“ das Netz durch die Anpassung der Gewichte in der Summationsfunktion sowie der Anpassung der Schwellenwerte der einzelnen Neuronen [AN00], bis es die gewünschten Ausgaben produziert. Das Wissen eines KNN steckt also nicht in den Neuronen selbst, sondern in den Gewichtungen der Kanten zwischen den Neuronen sowie den jeweiligen Schwellenwerten [Pet05]. Ein mögliches Verfahren, um diesen Lernprozess zu realisieren ist der Backpropagation-Algorithmus, welcher im nächsten Abschnitt in Verbindung mit dem Elman-Netz skizziert wird.

Elman-Netze sind partiell rekurrente[14] neuronale Netze. Ein partiell rekurrentes Netz besteht aus jeweils einer Eingabe- und Ausgabeschicht, sowie n versteckten Verarbeitungsschichten.[15] Im Gegensatz zu Feedforward-Netzen[16] ist der Informationsfluss in rekurrenten Netzen nicht strikt von der Eingabe- zur Ausgabeschicht definiert, sondern es gibt zusätzlich rückwärts gerichtete Kanten, die einen Informationsfluss von einer Schicht zu einer ihr vorgelagerten Schicht ermöglichen [Pet05]. Zusätzlich enthalten Elman-Netze weitere verdeckte Zellen, welche als Kontextzellen bezeichnet werden. Durch diese Zellen erhält das Netz ein „Gedächtnis“, welches es ihm ermöglicht, zeitliche Abhängigkeiten von Eingaben implizit zu verarbeiten [Elm90].

Abbildung 3 zeigt ein Elman-Netz. In einem Elman-Netz stimmt die Menge der Neuronen der verdeckten Schicht und der Kontext-Zellen immer überein [Elm90]. Es existieren trainierbare Verbindungen zwischen allen Neuronen der Eingabeschicht und allen Neuronen der verdeckten Schicht. Jedes Neuron der verdeckten Schicht hat eine rekurrente Verbindung zu einer Kontextzelle. Diese rekurrenten Verbindungen sind nicht trainierbar. Alle Kontextneuronen widerrum haben zu jeder Neurone der verdeckten Schicht eine trainierbare Verbindung. Der Informationsfluss ist dabei wie folgt [Pet05]: Gewichte und Schwellenwerte von der Eingangs- Ausgangsverdeckten- und Kontextneuronen werden mit Standardwerten initialisiert. Anschließend erfolgt das Anlegen eines Musters einer Musterfolge an den Eingabeneuronen, welche ihre Ausgaben wiederum an die verdeckte Schicht weiterleiten. Die Neuronen der verdeckten Schicht leiten ihre Ausgaben an die Ausgabeschicht sowie die Kontextzellen weiter. Die Kontextzellen besitzen als Aktivierungsfunktion die Identitätsfunktion. Das bedeutet, dass sich die Eingabewerte der

[14] Die rückwärtsgerichteten Kanten führen zum Begriff „rekurrent“. Da nicht alle Neuronen rückwärtsgerichtete Kanten besitzen, wird der Begriff „partiell“ verwendet [Pet05].
[15] Diese Schichten sind nach außen nicht sichtbar, agieren also als Black-Box.
[16] In Feedforward-Netzen ist der Informationsfluss strikt von der Eingabe- zur Ausgabeschicht definiert [Pet05].

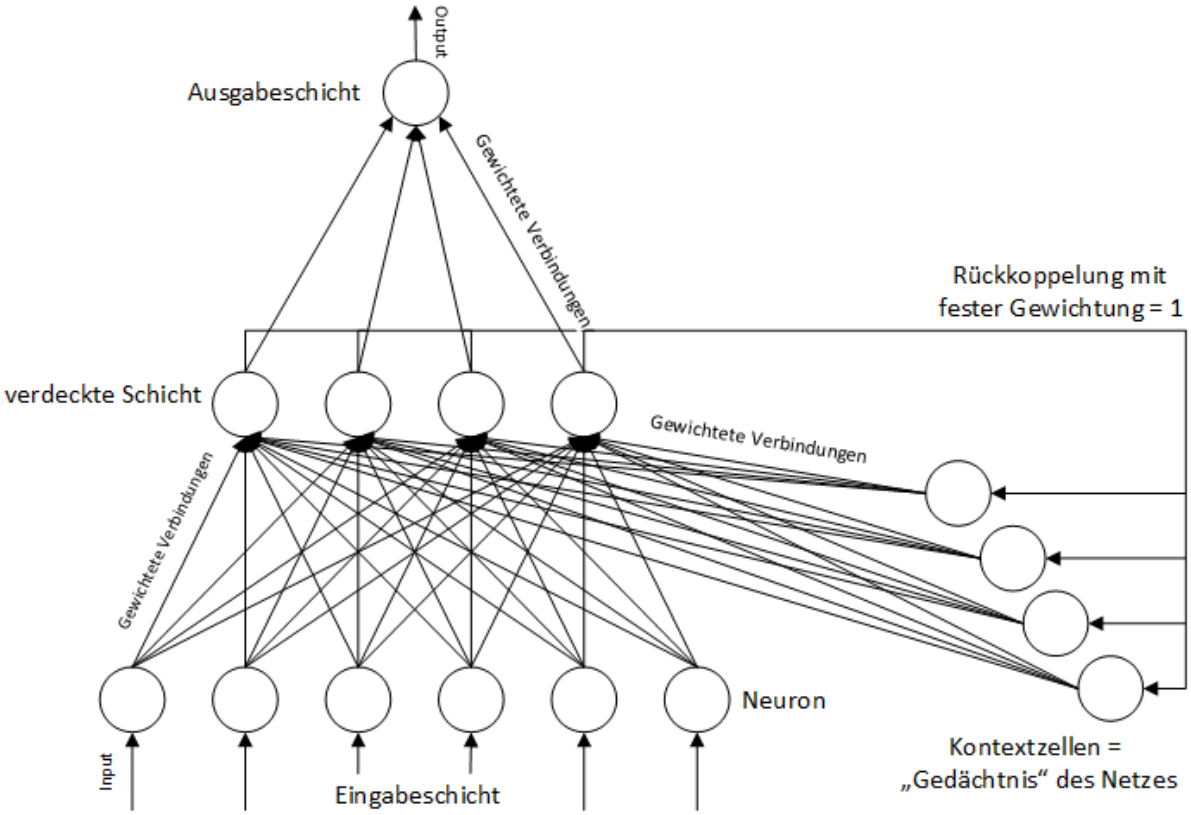


Abbildung 3: Beispielstruktur eines Elman-Netzes (Vgl. [Pet05])

Kontextzellen als Kopie der Ausgaben der verdeckten Zellen ergeben. Beim nächsten Durchlauf enthalten die Kontextzellen die Aktivierungen der verdeckten Zellen des vorigen Eingabemusters $(t-1)$, welche beim nächsten Durchlauf neben dem zweiten Muster als Eingabe in die verdeckte Schicht eingehen. So wird auf diese Weise ein impliziter Bezug zu früheren Mustern hergestellt [Pet05].

Die verdeckten Zellen der Elman-Netze entwickeln während des Trainings eine interne Repräsentation der Eingabemuster mit dem dazugehörigen Ausgabemuster. Durch die Kontextzellen werden die Aktivierungen der verdeckten Zellen durch vorangegangene Muster gespeichert. Die Aufgabe der verdeckten Zellen besteht darin, bestimmten Mustern, die an die Eingabezellen angelegt werden, im Zusammenhang mit dem Zustand der Kontextzellen bestimmten Ausgabemustern zuzuordnen. Damit werden die zeitlichen Eigenschaften der Eingabefolge in den verdeckten Zellen kodiert [Pet05].

Lernverfahren rekursiver Netze Abgesehen von den rekurrenten Verbindungen entsprechen partiell rekurrente Netze Feedforward-Netzen. Da die rückwärts-gerichteten Verbindungen in partiell rekurrenten Netzen nicht trainiert werden[17] finden Trainingsalgorithmen, die bei Feedforward Netzen angewendet werden auch bei den partiell rekurrenten Netzen mit leichten Mo-

[17] Die Kantengewichtung ist immer 1.

difikationen Anwendung [Pet05]. Ein solcher Algorithmus ist Backpropagation. Das Verfahren des Backpropagation-Algorithmus, angewendet auf partiell rekurrente Netze, wird im Folgenden vorgestellt [Pet05], [Lei13]:

1. Initialisierung der Kontextzellen.
2. Anlegen des Eingabemusters und Vorwärtspropagierung der Ausgabesignale der Neuronen bis zur Ausgabeschicht (ohne Beachtung der rekurrenten Verbindungen).
3. Vergleich der tatsächlichen Ausgabe mit der erwünschten Ausgabe. Ist der Fehler hinsichtlich eine zuvor definierten Schwelle klein genug, so wird das Training abgebrochen und es wird eine Testphase gestartet. Andernfalls wird eine Berechnung des Fehlersignals für jede Ausgabezelle durchgeführt.
4. Rückwärtspropagierung der Fehlersignale von den Ausgabezellen zu den Eingabezellen (ohne Beachtung der rekurrenten Verbindungen).
5. Berechnung der Gewichtsänderungen mit Hilfe des Fehlersignals und Adaption der Gewichte anhand einer vorgegebenen Lernregel.
6. Berechnung des Folgezustandes der Kontextzellen gemäß ihrer Eingangsverbindungen.

Vor- und Nachteile Künstlicher Neuronaler Netze Künstliche Neuronale Netze benötigen kein Wissen über ein Lösungsverfahren. Ebenso gelingt mit diesen Netzen die Lösung von linearen als auch nichtlinearen Problemenstellungen ohne Einschränkungen hinsichtlich des Wertebereiches der Eingangsgrößen. Zusätzlich sind KNN robust gegenüber verrauschten Daten. Sie bieten also eine hohe Fehlertoleranz bei guter Prädiktionsgenauigkeit [Lei13]. Mit dem Elman-Netz gelingt die Betrachtung des zeitlichen Zusammenhangs von Daten und deren Einfluss auf zukünftige Entwicklungen. Für sich häufig ändernde Zeitreihen (z.B. Aktienkurse) ist es oftmals notwendig, dass Modell ebenfalls neu zu erstellen. Hier bietet das Elman-Netz den Vorteil, das relativ schnelle Trainingsalgorithmen (Backpropagation) angewendet werden können [Pet05]. Die Berechnungen von KNN ähneln häufig jedoch einem Black-Box-Verhalten, was eine Interpretation bzw. ein Verständnis des eingeschlagenen Lösungsweg des KNN für den Nutzer erschwert. Ein weiterer Nachteil ist, dass die Architektur eines KNN von der Problemstellung abhängt. Der Nutzer muss die Anzahl der benötigten Schichten sowie die Anzahl der Neuronen vorgeben, die das Problem wahrscheinlich lösen [Alp08]. Hiervon sind auch die Elman-Netze betroffen: Sobald das zu lösende Problem nichtlinearen Charakter hat, reicht eine Schicht verdeckte Neuronen nicht mehr aus. Abhilfe schaffen die hierarchischen Elman-Netze, in denen mehrere verdeckte Schichten existieren. Jede verdeckte Schicht sowie die Ausgabeschicht enthält eigene Kontextzellen mit den bereits beschriebenen rekurrenten Verbindungen zu den Neuronen der jeweiligen Schicht [Pet05].

4.3 M5'-Modellbaumverfahren

Ein Entscheidungsbaum ist ein geordneter, gerichteter Baum. Seine hierarchisch angeordneten Knoten sind durch Kanten miteinander verbunden. Sie dienen dazu, anhand des Vergleiches von Konstanten, Attributen oder durch Anwendung einer Funktion einem Objekt eine Klasse

zuzuordnen [Pet05]. Ein Objekt o_j wird durch seine Eigenschaften und/oder Attribute $a_i (i = 1, ...n)$ gekennzeichnet. Eine Klasse wird als eine Entscheidung definiert, die zu treffen ist, wenn ein Objekt bestimmte Attributsausprägungen besitzt [Pet05]. Dazu wird der Entscheidungsbaum Top-Down mit dem zu klassifierenden Objekt durchlaufen bis ein Blatt erreicht ist, das seine Klasse definiert. Dieses Vorgehen entspricht einem Entscheidungsbaum für die Klassifikation von Objekten. Entscheidungsbäume können jedoch auch für Regressionen genutzt werden. Dazu werden zwei Varianten unterschieden [Lei13]:

- **Regressionsbäume:** Die Blätter speichern den Mittelwert aller Trainingsinstanzen, die dieses Blatt erreicht haben.
- **Modellbäume:** Die Blätter (und Knoten) enthalten ein lineares Regressionsmodell, welches für jede Trainingsinstanz anhand ihrer Attributsausprägungen ausgewertet wird.

Beide Baumarten bieten wiederum Lösungsmöglichkeiten für die in in Kapitel 4 beschriebenen uni- und multivariaten Verfahren [Lei13]. Modellbäume haben jedoch einige Vorteile gegenüber Regressionsbäume: Je nach Aufgabenstellung erzeugen Regressionsbaumverfahren deutlich größere, komplexere Bäume als Modellbaumverfahren [Pet05] und erfordern dadurch einen höheren Speicher- und Rechenaufwand. Regressionsbäume diskretisieren alle stetigen Variablen, indem diese mit Schranken in Intervalle eingeteilt und darauf Klassifikationsalgorithmen angewendet werden [Lei13]. Dadurch können für unbekannte Daten nur Werte vorhergesagt werden, die während des Trainings erlernt wurden, was in Hinblick auf Regressionsanalysen ein deutlicher Nachteil ist [Pet05]. Modellbaumverfahren umgehen diesen Nachteil, indem lineare Modelle für die Berechnung unbekannter Daten genutzt werden, was im Vergleich auch zu einer höheren Prädiktionsgenauigkeit führt [Lei13]. Aufgrund der genannten Vorteile von Modellbaumverfahren befasst sich der folgende Abschnitt mit einem konkreten Algorithmus für Modellbäume. Das ist der M5'-Algorithmus welcher von Y. Wang und Ian H. Witten im Oktober 1996 vorgestellt wurde [WW97], und als Modifizierung des M5-Algorithmus von [Qui92] gilt.

Funktionsweise des M5' Für die Erstellung eines Entscheidungsbaumes wird eine Trainingsmenge T so lange nach den Ausprägungen der Attribute in disjunkte Teilmengen T_i aufgespalten, bis jede Menge nur noch Objekte aus einer Klasse enthält [Pet05]. Dieser Knoten wird als Blatt bezeichnet, da keine weitere Verzweigung stattfindet. Der so entstandene Entscheidungsbaum soll die Trainingsmenge T mit einer möglichst kleinen Fehlerquote klassifizieren [Pet05]. Dies funktioniert nach der sogenannten „Teile-und-Herrsche-Strategie“ [HK00]: Das komplexe Problem wird rekursiv in kleine Teilprobleme zerlegt, die lösbar sind. Aus diesen Ergebnissen wird dann eine Lösung für das Gesamtproblem entwickelt. Diese – u.a. für Entscheidungsbäume typische Strategie – wird auch für Modellbäume genutzt. Modellbäume werden in zwei Schritten erstellt: Zunächst erfolgt das Aufbauen des Baumes (Baumінduktion) und anschließend erfolgt das Pruning (Zurückschneiden) des Baumes [Pet05]. Beide Phasen werden nachfolgend genauer erläutert.

Aufbau des Entscheidungsbaumes Zunächst muss ein binärer Entscheidungsbaum anhand einer Trainingsmenge T erstellt werden. Dazu wird die Trainingsmenge beginnend am Wurzelknoten des Baumes an jeden Knoten rekursiv in jeweils zwei Tochterknoten aufgeteilt, bis keine weitere Verzweigung mehr nötig bzw. möglich ist[18][Lei13]. Zur Bestimmung des Aufteilungskriteriums am jeweiligen Knoten muss dasjenige Attribut gefunden werden, welches über den höchsten Informationsgehalt[19] verfügt [Pet05]. Dieses Attribut wird in Modellbäumen mit Hilfe der Reduktion der Standardabweichung (SDR)[20] an jedem Knoten bestimmt [Pet05], [WF99]. Hierzu wird am betrachteten Knoten für alle Attribute der Instanzen aus T, die diesen Knoten erreichen, die Standardabweichung berechnet. Für ein Attribut j der Trainingsmenge T wird die Standardabweichung wie folgt berechnet (Vgl. [Aue05], [Lei13]):

$$sd(T_j) = \sqrt{var(T_j)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_{.j})^2}$$

Dabei entspricht n die Anzahl der Instanzen in T, x_{ij} entspricht dem konkreten Wert der Instanz X_i zu Attribut j und $\overline{x}_{.j}$ ist der Mittelwert des Attributes j über alle Instanzen $i = 1, ..., n$ der Trainingsdaten. Mit Hilfe der so berechneten Standardabweichungen wird anschließend für jedes Attribut die erwartete Fehlerreduktion anhand folgender Formel errechnet [WW97]:

$$SDR = sd(T) - \sum_{i} \frac{|T_i|}{|T|} \times sd(T_i)$$

Dabei bezeichnet $sd(T)$ die Standardabweichung aller Instanzen des betrachteten Knotens, i die Anzahl der zu bildenden Tochterknoten[21], $|T_i|$ die Anzahl an Trainingsinstanzen die bei einem Split mit dem betreffenden Attribut im Tochterknoten i entstehen, $|T|$ die Anzahl aller Trainingsinstanzen im betrachteten Knoten und $sd(T_i)$ die Standardabweichung der Trainingsinstanzen im Tochterknoten i [Pet05]. Das Attribut, welches die erwartete Fehlerreduktion maximiert, wird für die Aufteilung an dem betreffenden Knoten herangezogen [WF99]. Das Verfahren endet, d.h. ein Blatt wird gebildet, sobald die Standardabweichung der Instanzen, die einen Knoten erreichen, nur einen Bruchteil (zum Beispiel 5%) der ursprünglichen Instanzmenge entspricht [WF99]. Das Verfahren endet ebenfalls, wenn eine vom Nutzer definierte Schranke bezüglich der minimalen Anzahl von Instanzen in einem Knoten unterschritten wird [Lei13]. Beide Abbruchkriterien werden vom Nutzer definiert und gehören zum Vorbeschneiden (Pre-Pruning[22]) eines Baumes [Lei13].

Nach der Baumінduktion ist jeder Knoten und jedes Blatt mit einem linearen Modell zu versehen [WF99]. In den betrachteten Knoten gehen allerdings nur diejenigen Attribute in das Modell ein, welche in dem aktuell betrachteten und den unmittelbar darunter befindlichen Knoten als

[18] In diesem Fall wird der Knoten zum Blatt.
[19] Das Attribut, welches die höchstmögliche genaue Klassifikation der Trainingsdaten in einem Teilbaum realisiert.
[20] SDR (**S**tandard **D**erivation **R**eduction).
[21] In Modellbäumen werden immer zwei Tochterknoten gebildet, daher ist $i = 2$.
[22] Ein Pre-Pruning bezeichnet das Beschneiden eines Baumes noch bevor dieser vollständig aufgebaut ist [Lei13].

Verzweigungsmerkmal dienten [Lei13]. Die Modelle haben dann folgende Form [WF99], [WW97]:

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im}$$

$x_{i1}, x_{i2}, ..., x_{im}$ sind dabei die Werte der Attribute $j (j = 1, ..., m)$ für die betrachtete Instanz X_i und $w_0, w_1, ..., w_m$ die Gewichtungen der Attribute, welche mit Hilfe der Standardregression berechnet werden [WF99]. $\hat{y}_i$ gibt die Prädiktion des betrachteten Modells für die Instanz X_i an [Lei13].

Beschneidung des Baumes Um den Baum zu optimieren, wird dieser beschnitten (Post-Pruning) [Lei13]. Dadurch wird eine Überanpassung (Overfitting)[23] des Baumes an die Daten vermieden; zusätzlich wird die Generalisierungfähigkeit erhöht [Pet05]. Dies ist sinnvoll, da generalisierte und damit kleinere Bäume höhere Prädiktionsgenauigkeiten ermöglichen [Lei13]. Das Post-Pruning erfolgt in zwei Schritten: Im ersten Schritt werden vom Wurzelknoten ausgehend alle unwichtigen Variablen aus den linearen Modellen entfernt [Pet05]. Im zweiten Schritt wird der Baum von den Blättern ausgehend um seine Teilbäume gekürzt. Dabei wird wie folgt vorgegangen: Wenn der Prädiktionsfehler des betrachteten Knotens größer als der Prädiktionsfehler des übergeordneten Knotens ist, so wird der betrachtete Knoten entfernt und der übergeordnete Knoten wird zum Blatt [WF99]. In beiden Schritten wird hierfür der erwartete Fehler für ungesehene Daten berechnet, indem die Differenz zwischen dem Zielwert und der durch das Modell berechneten Prädiktion bestimmt und mit dem nachfolgenden Korrekturfaktor multipliziert wird [Lei13]:

$$corr = \frac{n + v}{n - v}$$

n stellt dabei die Anzahl der Trainingsinstanzen dar, die den betrachteten Knoten erreichen und v ist die Anzahl der Attribute, die in dem betrachteten Regressionsmodell dieses Knotens als Parameter genutzt werden [Lei13].

Anwendung von Testdaten Nach dem Aufbau und dem Beschneiden des Baumes können diesem anschließend Testdaten präsentiert werden. Die Testinstanz durchläuft dabei den Baum anhand ihrer Attributsausprägungen von der Wurzel ausgehend, bis ein Blatt erreicht wird. Das dort vorhandene lineare Regressionsmodell berechnet anschließend die Prädiktion für unbekannte Daten [Lei13]. Die Genauigkeit dieser Prädition lässt sich allerdings durch einen Glättungsprozess (Smoothing), welcher nachfolgend erläutert wird, verbessern [WW97].

Smoothing Beim Smoothing wird die Präditkion eines Blattes entlang der durchlaufenen Knoten zurück zur Wurzel mit Hilfe der jeweiligen Regressionsmodelle in den Knoten geglättet [WF99]. Die Berechnungen der linearen Modelle in den Knoten nehmen dadurch gewichteten Einfluss auf die Prädiktion des Blattes. Dies führt zu einer deutlich höheren Prädiktionsgenau-

[23] Overfitting bezeichnet die Tendenz, Trainingsdaten auswendig zu lernen, anstatt einen funktionalen Zusammenhang der Daten zu erlernen [Lei13].

igkeit [Pet05], [Qui92]. Eine geeignete Formel hierfür lautet [WW97]:

$$p' = \frac{np + kq}{n + k}$$

Dabei ist p' die an den nächsthöheren Knoten weitergereichte, geglättete Vorhersage. q ist die Vorhersage des betrachteten Knotens, p ist die Vorhersage, die der betrachtete Knoten vom untergeordneten Knoten erhalten hat. n ist die Anzahl der Trainingsinstanzen, die den betrachteten Knoten erreichen und k stellt eine Glättungskonstante dar.[24].

Vor- und Nachteile von Modellbäumen Zu den bereits genannten Vorteilen von Modellbäumen gegenüber Regressionsbäumen kommt hinzu, dass sie mit nummerischen als auch diskreten Attributen umgehen können [Lei13]. Im Vergleich zu Standardverfahren der linearen Regression bieten Modellbäume gerade für Daten, die sich nicht durch ein entsprechendes Modell darstellen lassen, genauere Vorhersagen [HK00]. Das Smoothing erhöht zusätzlich die Prädiktionsgenauigkeit des Modellbaumes [WW97]. Weiter ist der M5'-Algorithmus schnell im Training und robust im Umgang mit fehlenden Daten [Lei13] und Overfitting wird weitestgehend vermieden [Pet05]. Auch der Speicher- und Trainingsaufwand des Baumes ist verhältnismäßig gering: Es müssen nicht zwingend die Trainingsdaten selbst, sondern lediglich die Struktur des Baumes sowie die Parameter der Knoten bzw. Verzweigungen der Modelle und Blätter gespeichert werden [Lei13]. Da der Baum nach dessen Induktion beschnitten wird, ist dieser schneller und weniger komplex, so dass eine einfachere Interpretation ermöglicht wird [Pet05]. Insbesondere gründet dies auf der Tatsache, das Modellbäume mit Hilfe einfacher Wenn-Dann-Regeln schriftlich dargestellt werden können [Lei13].

5 Fazit

In [AN00] wurden Ansätze zur Unterstützung betrieblicher Planungsprozesse mit Hilfe von KNN und Entscheidungsbäumen gezeigt. Konkret geht es hier um die Unterstützung des Managements bei der Werbeträgerplanung im Internet, der Kundensegmentierung und der Einschätzung der Kreditwürdigkeit von Kunden. [AN00] führt hier Lösungswege mit Beispieldaten auf, allerdings wird die Prädiktionsgenauigkeit nicht immer eindeutig bewertet. In [Ste97] wurden verschiedene Entscheidungsbaumverfahren zur täglichen Wechselkursprognose untersucht. Hier lieferten die Entscheidungsbaumverfahren gute Ergebnisse, jedoch schnitten alle Baumverfahren bei der längerfristigen Prognose[25] durchweg schlecht ab, wobei das M5-Modellbaumverfahren das beste Ergebnis erzielte [Pet05]. In diesem Kontext könnte eine Kombination beider Verfahren eine Erhöhung der Prädiktionsgenauigkeit mit sich bringen, indem die jeweiligen Vorteile der Verfahren miteinander kombiniert werden: Baumverfahren benötigen weniger Trainingsinstanzen zur Induktion als KNN, dafür klassifizieren KNN genauer als Baumverfahren. [Pet05] schlägt hierzu die Erstellung und Überführung eines Entscheidungsbaumes in ein KNN vor, so dass das KNN

[24] In [Qui92] ist $k = 15$

[25] Es sollten Prognosen über den Zeitraum eines Monats erstellt werden [Pet05].

zunächst die Klassifikationsentscheidungen des Entscheidungsbaumes nachahmt. Anschließend wird das Netz mittels Backpropagation weiter auf die gewünschten Ausgaben trainiert. Eine Untersuchung ausgewählter kombinierter Verfahren in [SPW01] hat ergeben, dass diese – im Vergleich zum ursprünglichen Verfahren – leicht verbesserte Prädiktionsgenauigkeiten liefern.

Literatur

[Alp08] Alpaydin, Ethem: *Maschinelles Lernen*. München: Oldenbourg Wissenschaftsverlag, 2008

[AN00] Alpar, Paul ; Niedereichholz(Hrsg.), Joachim: *Data Mining im praktischen Einsatz*. Vieweg, 2000. – ISBN 3–528–05748–3

[Aue05] Auer, Joachim von: *Ökonometrie - Eine Einführung*. Springer, 2005. – ISBN 3–540–24978–8

[Cha13] Chamoni, Prof. Dr. P.: *Data Mining*. `http://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/daten-wissen/Business-Intelligence/Analytische-Informationssysteme--Methoden-der-/Data-Mining/index.html`. Version: September 2013, Abruf: 27.05.2014

[Coo14] Cooperation, Microsoft: *Microsoft Sequence Clustering-Algorithmus*. `http://technet.microsoft.com/de-de/library/ms175462(v=sql.105).aspx`. Version: 2014, Abruf: 29.05.2014

[Duc08] Duc, Kien N.: *Data Mining von Sequenzdaten*, Leibniz Universität Hannover, Diplomarbeit, 2008

[Elm90] Elman, Jeffrey L.: Finding Structure in Time. In: *Cognitive Science* (1990), Nr. 14, S. 179–211

[ES00] Ester, Martin ; Sander, Jörg: *Knowledge Discovery in Databases - Techniken und Anwendungen*. Springer, 2000. – ISBN 3–540–67328–8

[FPSS96a] Fayyad, U. ; Piatetsky-Shapiro, G. ; Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: *AI Magazine* 17 (1996), Nr. 3, S. 37–50

[FPSS96b] Fayyad, U. ; Piatetsky-Shapiro, G. ; Smyth, P.: The KDD Process for Extraction Useful Knowledge from Volumes of Data. In: *Communications of the ACM* 39 (1996), Nr. 11, S. 27–34

[HK00] Han, Jiawei ; Kamber, Micheline: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000. – ISBN 1–55860–489–8

[Kri05] Kriesel, D.: *Ein kleiner Überblick über Neuronale Netze*. `http://www.dkriesel.com/_media/science/neuronalenetze-de-zeta2-2col-dkrieselcom.pdf`. Version: 2005, Abruf: 24.07.2014

[Lei13] Leicht, Caroline: *Analyse und Optimierung von Algorithmen des Maschinellen Lernens in der Virtuellen Messtechnik*, Universität Leipzig, Diss., 2013

[Lip93] Lippe, Prof. Dr. P. d.: *Deskriptive Statistik*. Gustav Fischer Verlag, Jena, 1993. – ISBN 3–437–40268–4

[LLS10] Laudon, K. C. ; Laudon, J. P. ; Schoder, D.: *Wirtschaftsinformatik - Eine Einfuehrung*. Bd. 2. Pearson Studium, 2010

[Lue99] Luebbert, D.: *Statistik*. `http://www.luebbert.net/uni/statist/zr/zr2.php`. Version: 1999, Abruf: 29.07.2014

[Lü10] Lüdecke, Dipl. Tech. Red. E.: *Ermittlung von Assoziationsregeln aus großen Datenmengen*. Hochschule Merseburg, April 2010

[Pet05] Petersohn, H.: *Data Mining:*. Bod Third Party Titles, 2005 `http://books.google.de/books?id=U8B3tx4f--kC`. – ISBN 9783486577150

[Qui92] Quinlan, J. R.: Learning With Continuous Classes, World Scientific, 1992, S. 343–348

[RK96] Rehäuser, Jakob ; Krcmar, Helmut: *Wissensmanagement in Unternehmen*. Lehrstuhl für Wirtschaftsinformatik, Univ. Hohenheim, 1996

[SPW01] Seewald, Alexander K. ; Petrak, Johann ; Widmer, Gerhard: Hybrid Decision Tree Learners with Alternative Leaf Classifiers: An Empirical Study. In: *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference*, AAAI Press, 2001. – ISBN 1–57735–133–9, 407–411

[Ste97] Steurer, Elmar: *Ökonometrische Methoden und maschinelle Lernverfahren zur Wechselkursprognose : Theoretische Analyse und empirischer Vergleich mit 124 Tabellen*. Heidelberg : Physica-Verl., 1997

[WF99] Witten, Ian H. ; Frank, Eibe: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999. – ISBN 1–55860–552–5

[Wir14] Wirtschaftslexikon24: *Prognosehorizont*. `http://www.wirtschaftslexikon24.com/d/prognosehorizont/prognosehorizont.htm`. Version: 2014, Abruf: 29.07.2014

[WW97] Wang, Yong ; Witten, Ian H.: Induction of Model Trees for Predicting Continuous Classes. In: *Proceedings of the European Conference on Machine Learning Poster Papers. Prag, Tschechische Republik* (1997), S. 128 – 137

[WY05] Wang, Wei ; Yang, Jiong: *Advances in Database Systems*. Bd. 28: *Mining Sequential Patterns from Large Data Sets*. Kluwer, 2005. – 1–161 S. – ISBN 978–0–387–24246–0

[ZZY03] Zhang, Shichao ; Zhang, Chengqi ; Yang, Qiang: Data preparation for data mining. In: *Applied Artificial Intelligence: An International Journal* 17 (2003), Nr. 5-6, S. 375–381

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Bewertung des Einsatzes von prediktiven Methoden und Werkzeugen im SAP-Umfeld (SAP Predictive Analysis)

Seminararbeit
im Rahmen der Projektgruppe VLBA inMemory Planung mit SAP HANA

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	M.Sc. Deyan Antonov Stoyanov
Vorgelegt von:	Jonas Schlemminger jonas.schlemminger@uni-oldenburg.de
Abgabetermin:	02.08.2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

Tabellenverzeichnis **3**

1 Einleitung **4**

2 Prädiktive Analysen **4**
2.1 Ziele für das Prädiktive Modell 6
2.2 Funktionen von Analysewerkzeugen 7
2.3 Werkzeuge für Prädiktive Analysen 10

3 SAP Predictive Analysis **11**
3.1 R als Programmiersprache 12
3.2 Die HANA Predictive Analysis-Bibliothek 13
3.3 Predictive Analysis Architektur 15

4 Fazit **16**

Literaturverzeichnis **18**

Abbildungsverzeichnis

1 SAP Predictive Analysis Architektur [SAP14c] 16

Tabellenverzeichnis

1 Algorithmen Kategorie [SAP14c] 13

1 Einleitung

In der heutigen Welt fallen immer mehr Daten an. Diese enthalten wertvolle Informationen, jedoch wie und womit Nutze ich diese sinnvoll? Zudem stellen sich viele Unternehmen die Frage, wie sieht das Geschäft in der Zukunft auf? Welche Trends wird es geben? Mittels Prädiktive Analysen können diese Fragen beantwortet werden. So lassen sich aus bereits vorhandenen Daten, Zukunftsprognosen erstellen. Dies ist für Unternehmen erforderlich um auf dem Markt erfolgreich agieren zu können. Innerhalb dieser Seminararbeit werden Prädiktive Analysen und das Produkt SAP Predictive Analysis im Zusammenhang mit SAP HANA näher betrachtet und Bewertet. So werden zunächst auf die Grundlagen von Prädiktive Analysen eingegangen. Hierbei werden auf Ziele 2.1 der Prädiktiven Analysen aufgezeigt. Darauffolgend werden die Funktionen von Analysewerkzeugen 2.2 detailliert beschrieben und ein Überblick über die Werkzeuge 2.3 gegeben. Danach folgt das 3 Kapitel mit dem Schwerpunkt von SAP Predictive Analysis.

2 Prädiktive Analysen

Vorhersagemodelle sind wichtig, denn sie ermöglichen es Unternehmen, die Ergebnisse der Alternativstrategien vor der Implementierung zu prognostizieren und zu bestimmen, wie eine effektiv Verteilung von knappe Ressourcen, wie Marketing-Budget oder Arbeitsstunden möglich ist. Typische Anwendungen für Vorhersagemodelle sind:

- Responsemodelle - Reaktion vorherzusagen, welche Kunden reagieren am ehesten auf eine Marketing-Maßnahme (z.B E-Mail, Newsletter, Promotion)
- Cross-Selling und Up-Selling-Modelle, welche Produktvorschläge sind sinnvoll für einen zusätzlichen oder erhöhten Verkauf an einen bestimmten Kunden.
- Abwanderungsmodelle/Reaktivierungsmodelle, welche Kunden wandern am ehesten ab oder kaufen in der näheren Zukunft nicht mehr.
- Betrugsmodelle vorhersagen, welche Transaktionen und Interaktionen sind wahrscheinlich betrügerisch oder erfordern eine weitere Überprüfung

[SAP14c]

Die gemeinsamen Geschäftsprobleme die durch Vorhersagemodelle angesprochen werden,

sind nicht mit prädiktiven Algorithmen zu verwechseln. Jede der oben genannten Probleme können mit einer Reihe verschiedener Algorithmen gelöst werden. Das Verständnis der Eigenschaften eines Business-Problem und verbinden der Daten mit dem am besten geeigneten Vorhersagealgorithmus ist der Teil der statistischen Modellierung, die oft mehr Kunst als Wissenschaft ist. Zum Beispiel, ein Unternehmen möchte eine einfaches Ergebnis (z. B. ein Kunde will ein Angebot annehmen) vorhersagen, kann der Modellierer einen Entscheidungsbaum oder ein Regressionsmodells zu verwenden. Jeder dieser Vorhersageverfahren hat Vorteile und Nachteile hinsichtlich der Einfachheit der Durchführung, Präzision, Genauigkeit und Entwicklungsaufwand. [SAP14c]

Während der Wert der Vorhersagemodelle variiert von Firma zu Firma, ist es leicht, den Wert der Vorhersageergebnisse besser zu quantifizieren. Aus einer Marketing-Perspektive, kann die Zuteilung knapper Marketing-Ressourcen, um des Kunden Antwortraten zu erhöhen und die Kosten in der gleichen Zeit, oft mit einem Return on Investment in der Größenordnung von Millionen Dollar pro Jahr betragen. Vorhersagemodelle erlauben auch Unternehmen, mehrere Vorschläge in einer Simulationsumgebung, Vorhersagen von Ergebnis und Einnahmen zu testen, anstatt sich auf das Bauchgefühl des Management bei der Entscheidung zwischen Alternativen zu verlassen. Finanzdienstleistungen oder Versicherungen, können besser vorhersagen, welche Kunden wahrscheinlich einen Anspruch auf ein Darlehen haben. So ist das Risiko genauer ein schätzbar und eine höhere Wahrscheinlichkeit einen guten Kunden zu gewinnen. Ebenso mit wiederholbaren, messbaren Geschäftsregeln für die Erstellung dieser Modelle ermöglicht es Unternehmen, schneller auf den Markt reagieren und ermöglicht sich ändernden Geschäftsumgebungen zu identifizieren. Typischerweise entwickeln Firmen Vorhersagemodelle für einen bestimmten Bereich oder Abteilung, daraus lassen sich schnell viele Möglichkeiten identifizieren, für ähnliche Anwendungen oder andere Funktionsbereiche die Vorhersagemodelle anzuwenden. Die benötigten Daten werden aus einem Data Warehouse extrahiert und häufig für die Übertragung auf das Vorhersagetool umgewandelt. Diese Datenübertragung erfolgt über Textdateien, Exporte oder Datenbankverbindungen. In den besten Fällen ist das Vorhersagemodellierungswerkzeug in der Lage, direkt auf das Data Warehouse zugreifen und schreiben. Oft ist der Datenübertragungsprozeß iterativ, die Daten werden extrahiert und Variablen hinzugefügt, gelöscht oder modifiziert. Obwohl viel Wert auf die für die Vorhersage der verwendeten Software gelegt wird, ist der Betrieb der Vorhersagealgorithmen tatsächlich nur ein kleiner Teil des Modellaufbauprozess. Laut der SAP entfallen bei der Erzeugung der Vorhersagemodelle nur 20 Prozent des Zeit-und Arbeitsaufwand, bei der Modellierung an. Datenmanipulation, Erforschung und Umsetzung benötigen mehr Zeit

als das tatsächliche Modell zu erzeugen.[SAP14c]

2.1 Ziele für das Prädiktive Modell

Alle Führungskräfte haben Entscheidungsprobleme die für die Zukunft ihres Unternehmens oder der Branche relevant sind. Das identifizieren der Möglichkeiten, um Predictive Analytics Einblicke in umsetzbare Geschäftsentscheidungen zu ermöglichen, ist oft eine Herausforderung, da Analysten oftmals überfordert sind mit einer Zusammenfassung und Prüfung der verfügbaren Daten, um einen Return on Investment für das Unternehmen zu produzieren. Die Analysten müssen zusammen mit dem Management Ziele für die Analyse und dem gewünschten Ergebnis identifizieren. Diese können sein:

- Welche Kunden reagieren am ehesten auf eine Marketing-Kampagne?
- Welche Kunden könnten ihre Abonnements kündigen oder die Geschäftsbeziehung?
- Welche Angebote, Werbung oder andere Aktionen können einen höheren Kaufbetrag auslösen?
- Welche Kunden haben Lebenereignisse, die einen Kauf auslösen können (Geburtstag, Hochzeitstag)?

Die Beantwortung dieser Fragen liefert umsetzbare Ergebnisse mit einem messbaren Return on Investment.[SAP14c]

Schließlich muss in der Entwicklung der Ziele der Analyse, die Analysten- und Management-Teams eine ausreichende menge Daten zur Verfügung haben, um die Modelle zu entwerfen. So kann beispielsweise ein Unternehmen, das nicht über eine Kundendatenbank verfügt, nicht Entwicklungen oder Kundensegmente ermitteln, wie z.B. welche Kunden sind am profitabelsten. Ein Versicherungsunternehmen, das ein Modell entwerfen will um betrügerische Ansprüche erkennen zu können, muss in der Lage sein zu identifizieren, welche Vorgänge in der Vergangenheit betrügerische Ansprüche waren. Ein Vorhersagemodell ist keine Glaskugel, der Erkenntnisse aus dem Glas entdecken kann, sondern es ist einfach ein System von Regeln oder Gleichungen, die Erfahrungen der Vergangenheit analysieren kann, um die wahrscheinlichsten zukünftigen Ergebnisse zu bestimmen. Leider wird dieser Teil des Prozesses oft ignoriert und Zeit und Mühe werden verschwendet, wenn die Modellentwickler später feststellen, dass es nicht genügend Daten vorhanden sind, um die Analyse abzuschließen.[SAP14c]

Trotzdem kann kein Werkzeug die beste Geschäftsstrategie auswählen und diese kommunizieren, jedoch kann den Analyseteams die Modellierungsanforderungen helfen, die Strategie zu implementieren. Die Datenvisualisierung und BI-Tools helfen, Trends zu erkennen und erste Einblicke in prädiktive Analyse zu geben. Eine gesunde BI Praxis und benutzerfreundliche Abfrage-Tools können Helfen Verbesserungen zu identifizieren und schnell die Angemessenheit der Daten für die Modellierung erkennbar zu machen. Dies beschleunigt den ersten Schritt in Richtung der Modellierung.[SAP14c]

Die Datenquelle

Neben der Berücksichtigung der Unternehmensziele, muss dieser erste Schritt auch einen Plan für die Beschaffung der für die Analyse erforderlichen Daten haben. Die zur prädiktiven Erkenntnisse generieren Datensätze sind entscheidend, für den Erfolg des analytischen Projekt. Die Modellierungsdatensätze müssen nicht nur sorgfältig konstruiert werden, sondern auch eine gemeinsame Kooperation zwischen den Fachexperten die die Daten verstehen, den technischen Teammitglieder, da diese tatsächlich die Daten Extrahieren und damit die Datensätze erstellen und der Analytik Team-Mitglieder, die die Daten benutzen und die Modelle erstellen. In den besten Fällen hat die Organisation ein Data Warehouse mit Daten aus allen Bereichen der Gesellschaft in einer zentraler Datenbank und in einem Standard-Format geladen. Typischerweise wird das unternehmensweite BI-Reporting-System (z. B. Businessobjects) genutzt um Business-Anwender das Reporting und den Datenzugriff zu erleichtern. Um sicherzustellen, dass die Daten für die erforderlich Modellierung genau genug sind, müssen manchmal die Daten als voraggregierten Bericht aus dem Data Warehouse extrahiert werden. [SAP14c]

2.2 Funktionen von Analysewerkzeugen

Bei der Bewertung von vorausschauenden Analysewerkzeugen, sollten Modellentwickler mehrere Funktionsbereiche prüfen um zu gewährleisten, dass das Werkzeug ihren Bedürfnissen entspricht. Die Verbindung zwischen Fähigkeiten des Modellierungswerkzeug, der organisatorischen Anforderungen und dem Budget, bestimmt welche Lösung sich am Besten eignet. Dieser Abschnitt umfasst die wichtigsten Funktionsbereiche, die bei der Auswahl zu berücksichtigen sind. [SAP14c]

Datenzugriff

Wie bereits erläutert, ist die Datenerhebung oft der zeitaufwendigste Teil der Modellierung, so muss sichergestellt werden, dass das Modellierungswerkzeug auf die Daten zugreifen

kann. Dies ist entscheidend, um diesen Prozess zu beschleunigen. So stellen sich bei der Auswahl folgende fragen[SAP14c]:

- Wie werden die Daten in das Tool importiert?
- Kann auf Datenbanken direkt zugegriffen werden, oder müssen Daten ausschließlich über Dateien übertragen werden?
- Unterstützt das Tool das Schreiben von Ergebnisse oder Modellen zurück in die Datenbank?

Daten Manipulation

Datenmanipulation enthält Gebinde, Gruppierung, Wert Änderungen und die Berechnungen auf die bestehenden Datenfelder. Wenn innerhalb des Modellentwicklungsprozess die Bewertung und die Modifizierung der Daten innerhalb der Datenbank erfolgt, kann diese Funktionalität den Modellierungsprozess beschleunigen. So müssen nicht jedes Mal die Daten aus dem Quellsystem entnommen werden. Wenn diese Funktionalität jedoch nicht gegeben ist, müssen diese Daten exportiert und Dokumentiert werden. Außerdem müssen diese neuen Daten in jedem System eingepflegt werden.[SAP14c]

System-Architektur und Verarbeitungskapazität

Einige Vorhersagealgorithmen erfordern erhebliche Rechenleistung, oftmals werden mehrmals die Daten durchlaufen, um ein optimales Modell zu berechnen. Da immer mehr Daten zur Verfügung stehen und wollen Unternehmen diese große Daten analysieren, jedoch sicherzustellen dass das Vorhersagewerkzeug diese große Datenmengen verarbeiten kann, ist kritisch. Daher müssen Unternehmen zwischen prädiktiven Tools, die auf dem lokalen Rechner des Benutzers installiert sind, und jenen, die Daten auf einem Server verarbeiten kann, zu entscheiden. Lokale Client-Tools sind einfach zu installieren und erfordern keine spezielle Hardware, sind jedoch in der Menge der Daten, die sie verarbeiten können, beschränkt. Server-basierte Tools erfordern in der Regel dedizierte Hardware und sind komplexer zu installieren und zu warten, können dafür aber große Datenmengen verarbeiten und ermöglichen viele Nutzer die gleichen Ressourcen zu nutzen.[SAP14c]

Benutzeroberflächen

Predictive Tools haben sehr unterschiedliche Benutzeroberflächen, variierend von benutzerfreundliche Drag-and-Drop-Funktionalität, bis hinzu einem Code-only-Editor. Einige Tools haben nicht einmal eine Benutzeroberflächen und können nur über Batch-Jobs benutzt werden. Werkzeuge, die nur mittels Programmiercode bedienbar sind, bieten oft

mehr Funktionen und umfangreicher prädiktive Bibliotheken, können jedoch die Entwicklungszeit erhöhen und erfordern mehr technisches Verständnis. Benutzeroberflächenbasierte Lösungen können auch durch weniger technischen Mitarbeiter benutzt werden und Beschleunigen die Modellentwicklung.[SAP14c]

Predictive Algorithmen

Die Bibliothek der prädiktiven Algorithmen ist in jedem Werkzeug unterschiedlich. Während zahlreiche Algorithmen existieren, können die meisten Unternehmen eine breite Palette von Analyse mit einer begrenzten Funktionalität bieten, so dass wenige Algorithmen für jede Klassifizierung, Clustering, Regression und Zeitreihenanalyse vorhanden sind. Es ist jedoch wichtig, um die Ziele oder Modelltypen die die Organisation erwartet, vor der Auswahl eines Werkzeugs zu definieren um sicherzustellen, dass das ausgewählte Werkzeug entsprechende Funktionalität aufweisen. Zum Beispiel, wenn eine Organisation eine Vorhersagewerkzeug kaufen möchte, dass ausschließlich Umsatzprognosen erstellt, sollte es ein Tool, dass in diesem Bereich mit besonderen Merkmalen für Quartale und regelmäßige Terminen spezialisiert ist kaufen. Während ein Unternehmen das die Planung bis zur Kundenanalyse durchführen möchte, eine Vielzahl von Werkzeuge benötigt, wie Clustering, Entscheidungsbäumen und möglicherweise Regressionsalgorithmen.[SAP14c]

Modellauswertung Eigenschaften

Auswertung von Modellen und vergleichen von Alternativen ist der Schlüssel für die Auswahl des endgültigen Modells. Werkzeuge, die Analysten im Vergleichen von Alternativen unterstützen, beschleunigen die Entwicklung und den Auswahlprozess. Modell Auswertungstools umfassen automatisierte Visualisierungen wie Liniendiagramm, Restanalysen und Konfidenzintervall von Koeffizienten und vorhergesagten Werten.[SAP14c]

Modell Implementierung und Wartung

Sobald ein Modell gewählt wird, wollen die meisten Organisationen, es so schnell wie möglich bereitstellen. Abhängig von den organisatorischen Anforderungen, kann dies einfach sein durch die Modellauswertung auf Basis einem kleinen Satz von Daten. In vielen Fällen erfordert die Organisation jedoch die Fähigkeit, die Nutzwertanalyse und deren Algorithmus zurück in die Datenbank zuschreiben und nicht nur die Daten. Predictive Tools die die Algorithmen zurück in die Datenbank als eine gespeicherte Prozedur- oder Funktionsaufruf oder als einen abrufbaren Code-Block speichern können, beschleunigen diesen Prozess. Abhängig von der Komplexität des Bewertungsalgorithmus, der Berechnung der Koeffizienten und die Programmierung der Bewertungsfunktion, kann dies unterschied-

lich zeitaufwendig sein. Außerdem, wenn die Daten innerhalb der Modellierungswerkzeug manipuliert wurde und das Tool in der Lage ist, diese Regeln in den Algorithmus zu integrieren, beschleunigt dies die Umsetzung.[SAP14c]

2.3 Werkzeuge für Prädiktive Analysen

Während die Popularität von Predictive Tools explodiert, sind Software-Anbieter gefragt, die erhöhten Anforderungen der Benutzer für die Datenverarbeitungsleistung und erhöhte Funktionalität zu bieten und gleichzeitig die Benutzerfreundlichkeit nicht zu vernachlässigen. Die Forschung zeigt, dass die R-Programmiersprache einer der Top-Statistikpakete ist, von denen die für Durchführung Predictive Analytics verwendet wird. Die Nutzung von R ist schnell gewachsen in den letzten Jahren. Kommerziell verfügbare Software wird häufiger von Wirtschaftsorganisationen eingesetzt. Jedoch durch die Höhe der Lizenzkosten von einigen populären Software-Pakete nutzen Unternehmen auch Open-Source-Tools, wie zum Beispiel R. Weitere Informationen zu R folgen im Abschnitt 3.1. Die folgenden Unterpunkte geben einen Überblick über die Merkmale die bei der Auswahl eines Werkzeugs nützlich sind. Sobald die analytischen Ziele identifiziert sind sollte die Organisation bestimmen, welches Werkzeug an die Bedürfnisse des Projekts und die langfristigen Ziele der Organisation die beste Übereinstimmung liefert.[SAP14c]

Algorithmus-spezifische Werkzeuge

Während viele Tools versuchen, eine vollständige Auswahl von Algorithmen zur Verfügung zu haben, hat dieses Werkzeug mehrere Tools zur Verfügung, mit einem engen Fokus, die ein Algorithmus oder eine Teilmenge von Algorithmen sehr gut durchführen. Diese Tools bieten oft eine gute Bedienbarkeit und Visualisierungsfunktionen, die volle Funktionswerkzeuge übertreffen, jedoch nur für eine Art von Algorithmus, wie z. B. Entscheidungsbäume.[SAP14c]

Vollfunktion Programmierbasierte Werkzeuge

Die umfassendsten Tools, die Zugriff auf die größte Auswahl an Diagnosetools und Modellierungsalgorithmen bieten, sind in der Regel Komplex, so muss sich der Benutzer Kenntnis von Statistiken und Programmierung haben. Diese Werkzeuge sind häufig auch voll Funktionsfähige Programmiersprachen und können daher für alle erforderlichen Datenverarbeitungs-und-manipulation verwendet werden. Und zusätzlich für die Programmierung sämtlicher Algorithmen, die nicht bereits in der Bibliothek enthalten sind. Diese Tools bieten erhebliche Flexibilität bei der Datenaufbereitung, prädiktiven Algorithmen und Modellevaluierung, aber leiden unter einem Mangel an Benutzerfreundlichkeit, eine hohen Vorwissen

bedarf und Schwierigkeit Visualisierungen zu erzeugen.[SAP14c]

Cloud-Lösungen

Die jüngsten Marktteilnehmer bieten Prädiktive Cloud-Lösungen mit Web-basierte Modellierungs Oberflächen, Cloud-basierte Datenspeicherung und-verarbeitung, und einem Pro-Byte-Bezahlsystem oder Pro-Diagramm-Bezahlsystem für die Datenspeicherung, Modellbau und Vorhersage. [SAP14c]

3 SAP Predictive Analysis

SAP entwickelt das neue Predictive Analysis-Tool als Erweiterung von SAP Lumira. Predictive Analysis enthält alle Funktionen von Lumira (z.B. Datenerfassung, Manipulation, Formeln, Visualisierungstools und Metadaten-Anreicherung) mit der Zugabe der Predictive Ansicht. Der Prädiktive Bereich enthält alle Predictive Analysis-Funktionalität und auch die Vorhersage-Algorithmen, Ergebnisse, Visualisierung Analysen und Modell-Management-Tools. SAP sieht Lumira und Predictive Analysis als Visualisierungs-und Analyse-Suite. Diese Tools bieten eine Unternehmenslösung, in der Business-Analytics-Anwender und Wissenschaftler, die Daten zu verwenden, um Predictive Analysis Modelle entwickeln und erstellen zu können und Dateien in aus dem SAP proprietäre *. Svid-Format mit Business-Anwender und Führungskräfte (Diesen Benutzer-und Führungskräften kann der Zugang auf diese Werkzeuge innerhalb von Lumira begrenzt werden) nutzen zu können. Diese Lösung ermöglicht es, das diese Gruppen Einblicke, Informationen und Ergebnisse untereinander austauschen und schnell und einfach zu implementieren zu können, um die diese Erkenntnisse und Modelle, in andere Werkzeuge im SAP Businessobjects nutzen zu können.[SAP14c]

Predictive Analysis kann durch SAP HANA ergänzt werden. Sie können jedoch Predictive Analysis auch ohne HANA nutzen. Predictive Analysis wird lokal auf dem Rechner des Benutzers installiert und greift auf die Daten für die Verarbeitung auf den lokalen Rechner (aus einer CSV-, Microsoft Excel-oder JDBC-Verbindung zu einer Datenbank) oder auf SAP HANA zu. Für die Offline-Verarbeitung unterstützt Predictive Analysis auch eine lokale Installation von SAP Sybase IQ (In-Memory-Datenbank), um die Daten ändern und die Vorhersage speichern zu können. Predictive Analysis ist ein Installationspaket das in wenigen Minuten installiert ist und umfasst ein Installationswerkzeug, um die erforderlichen Komponenten für die R-HANA-Offline-Verarbeitung zu laden. Predictive Analysis kann auf Windows Computern ausgeführt werden und benötigt keine anderen SAP-Tools.

[SAP14b]

Die Zielgruppe für Predictive Analysis ist ein Mitarbeiter, das prädiktive Erkenntnisse aus Daten zu extrahieren muss. Diese Person kann ein Wissenschaftler sein, der in der Regel mit einem Code-basierte Statistik-Tool arbeitet oder ein Basis-oder Business Analyst, der mit der Benutzeroberfläche von Businessobjects-Tools vertraut ist. SAP fördert Predictive Analysis als Prognoseinstrument für die Massen, die Nutzer fühlen sich besser in der Lage zu verstehen und zu interagieren mit den Ergebnissen, wenn sie zumindest einen flüchtigen Hintergrund in der Vorhersagetechniken und statistischer Hinsicht haben. Zukünftige Updates auf das Werkzeug wird wahrscheinlich die Zielgruppe an beiden Enden des Spektrums erhöhen. Wenn zusätzliche Funktionen und Algorithmen hinzugefügt werden, sind mehr Wissenschaftler in der Lage, aus ihren Code-basierte statistische Werkzeuge für die Analyse zu Predictive Analysis zu wechseln. SAP versucht immer mehr Analysepfade zu integrieren, die das Werkzeug besser nutzbar machen für Business-Anwender die kein statistischen Hintergrundwissen haben. Predictive Analysis stützt sich auf mehrere Modellsysteme. Während die Predictive Workbench auf eine Drittanbieter-Verarbeitungs-Engine (SPSS Clementine / PASW) basiert, entschied SAP eine Kombination von intern entwickelten Modellierungsalgorithmen und Open-Source-R Algorithmen für Predictive Analysis zu verwenden. [SAP14c]

3.1 R als Programmiersprache

R ist eine Open-Source-Programmiersprache und Laufzeitumgebung, die stark von den Statistikern und Mathematikern verwendet wird, und ist besonders beliebt in der Lehre und Forschung.[SAP14c] R ist über CRAN(Comprehensive R Archive Network) frei verfügbar und steht unter der General Public License.[RP14] R speichert alle Daten, Objekte und Definitionen im Speicher und hat eine eigene Speicherverwaltung. R wird in der Regel über eine Befehlszeilenschnittstelle abgerufen. Doch mehrere Editoren und integrierte Entwicklungsumgebungen, wie z. B. R Studio, stehen zur Verfügung. R gewinnt immer mehr Popularität in der Geschäftswelt. Da jedoch R eine Programmiersprache ist, bedarf es einem Mitarbeiter und Statistiker mit erheblichen Programmierkenntnisse, um eine prädiktive Analyse durchzuführen. Der Großteil der umfangreichen Funktionalität prädiktive R ist durch Pakete durch das weltweite Netzwerk der R Anwender und Entwickler verfügbar. Die Pakete auf dem CRAN-Netzwerk werden vor der Übermittlung Überprüft und der Großteil der Funktionalität ist weitgehend von Benutzer getestet. Dies führt zu relativ robust und zuverlässig Code für häufig verwendete Algorithmen, aber möglicherweise zu weniger zuverlässig Code für wenig verwendete Algorithmen. Neben der frei Verfügbar-

keit und Open Source, ist ein größte Vorteil von R die Flexibilität, die es bietet.[SAP14c] Da R eine Programmiersprache ist, kann ein erfahrener Programmierer praktisch jeder Algorithmus in R implementieren. So wurden innerhalb von SAP Predictive Analysis viele R-Algorithmen durch SAP entwickelt.

3.2 Die HANA Predictive Analysis-Bibliothek

Die HANA Predictive Analysis Library (PAL) ist eine Reihe von Vorhersagealgorithmen in der HANA Application Function Library (AFL). Diese wurde speziell entwickelt, so dass HANA komplexe prädiktiven Algorithmen durch die Maximierung der Datenbankverarbeitung erfolgen kann [SAP14a], anstatt alle Daten auf dem Anwendungsserver aufbereiten zu müssen. Die PAL stellt Vorhersagefunktionen zur Verfügung, die von SQLScript Code auf HANA aufgerufen werden können. Es stehen sechs Kategorien von Algorithmen in der PAL mit insgesamt 23 unterschiedlichen Algorithmen zur Verfügung. Die sechs Kategorien sind in Tabelle 1 beschrieben. [SAP14c]

Algorithmen Kategorie	**Beschreibung**
Clustering	Es werden Ähnlichkeiten zwischen Daten ermittelt. Es werden nur numerische Daten akzeptieren.
Klassifikation	Objekte werden automatisch Klassifiziert. Es werden Entscheidungsbäume, numerische und logische Daten unterstützt.
Verbindung	Apriori-Algorithmus dieser dient der Auffindung von sinnvollen und nützlichen Zusammenhängen.
Präprozess	Datenaufbereitungsalgorithmen für die Bewertung und Manipulation der Daten einschließlich Normalisierung und Ausreißererkennungs
Zeitreihen	Algorithmen für Einzel-, Doppel-und Dreifach exponentielle Glättung und für die Prognose der zeitabhängigen Daten, mit der Fähigkeit auf Trends und Saisonalität zu reagieren.
Verschiedenes	Gruppenalgorithmen die Alternativen bewerten z.B. über ABC-Analysen.

Tabelle 1: Algorithmen Kategorie [SAP14c]

Während die PAL ist auf jeder HANA Implementierung auch ohne Predictive Analysis verfügbar ist, sind die PAL Algorithmen ausschließlich über Scripting im HANA Studio benutzbar und können meist nur von Programmierer oder Statistiker bedient werden. Zusätzlich erfordert jede Visualisierung oder Auswertung der Modellanpassung erhebliche

Programmieraufwand um die Ergebnisse in ein Visualisierungstool für die Berichterstattung zu exportieren.[SAP14c]

Predictive Analysis Lokale Algorithmen

Während Predictive Analysis sich am stärksten auf der R-Vorhersagenmodelle konzentriert, bietet die PAL im HANA Online-Modus sieben Algorithmen für lokale (offline) Verarbeitung, die nicht R-Basierend sind. Die meisten Algorithmen die Lokal zur Verfügung stehenden, sind R-basierte Algorithmen (Dreifach exponentiellen Glättungszeitreihenmodelle und fünf Sorten von Regressionalgorithmen), jedoch sind diese lokalen Algorithmen die einzige Quelle für die Ausreißererkennung Algorithmen. Die lokalen Vorhersagealgorithmen erlauben Predictive Analysis etwas ähnliche Funktionalität im Offline-Modus, als die PAL im HANA-Online-Modus bietet, aber der Großteil der Predictive Modeling-Funktionalität ist über die R prädiktiven Algorithmen im Offline-Modus verfügbar.
Predictive Analysis arbeitet in zwei Modi: Online-Daten über HANA oder offline mit heruntergeladenen Daten. Wenn die Online-HANA-Datenquelle genutzt wird, aktiviert sich der HANA Online-Verarbeitungsmodus. Der Modus bestimmt, ob Daten Manipulation Funktionen aktiviert sind und der Vorhersagealgorithmen zur Verfügung steht. Der größte Vorteil der direkten Integration von Predictive Analysis mit der Datenbank ist, dass die Datenextrakt-Definition innerhalb der Predictive Analysis Dokumente gespeichert ist und somit bei Bedarf auf einen aktualisierten Datenbestand zugegriffen werden kann. Wenn manuelle Abfragen erstellt werden, um Daten zu extrahieren oder um für den Import von Text-Dateien, können die Feldberechnungen und Auswahlkriterien verloren gehen und damit sehr zeitaufwendig dies neu erstellen.[SAP14c]

Zugreifen auf Daten mit Online-HANA

Neben dem Herunterladen von Daten und dem lokalen ausführen auf dem Client-Rechner, können Predictive Analysis auch in Verbindung mit einer HANA-Server und Linux-Rechner genutzt werden, um die PAL-und R-Algorithmen zu verarbeiten. So können auf die Online-Daten aus HANA wie Tabellen, Berechnung, Ansichten und Analysesichten zugegriffen werden. Im HANA Online-Modus gibt es keine Datenmanipulation möglichkeit, wie jene, die in Predictive Analysis integriert ist, jedoch können die Visualisierungswerkzeuge genutzt werden. Darüber hinaus erhöht HANA die Kapazität Predictive Analysis, da es nicht mehr auf die Verarbeitungsleistung des Client-Maschine beschränkt ist. [SAP14c]

Daten Manipulation

Sobald die Daten aus einer oder mehreren Quellen geladen worden sind, ermöglichen die

Datenmanipulation Komponenten es dem Benutzer schnell Datenelemente zu erstellen und zu ändern. Gruppieren und Transformieren von Daten ist besonders wichtig, für den Modellbauprozess. Viele Modellierungswerkzeuge ermöglichen eine minimale Datenmanipulation, eine völlig neue Modellierung z.B. durch die Änderung der Altersgruppen und so ein neues Modell zu generieren. Die Prädiktive Analyse erleichtert die Berechnungen und Manipulationen an vorhandenen Daten und Erweiterung dieser, um nicht diese Daten außerhalb des Werkzeugs zu manipulieren zu müssen.[SAP14c]

Datenanreicherung

Sobald die Daten innerhalb des Predictive Analysis importiert worden sind, erkennt die Software automatisch potentielle Anreicherungen auf die Attributfelder. Anreicherungen bieten zusätzliche Funktionalität für spezifische Arten von Attributen. Zum Beispielwerden Datumsfelder wie die Zeit Hierarchien bereichert, werden automatisch Teilergebnisse für Jahr, Quartal, Monat und andere Intervalle erstellt. Prädiktive Analyse können diese Daten auf drei verschiedene Arten verbessern:

- Geografische Hierarchie
- Zeit-Hierarchie
- Semantische Anreicherung (Erstellen einer Maßnahme)

[SAP14c]

Daten-Visualisierung

Predictive Analysis hat ein einfach zu bedienendes Daten Findungs-Tool. Durch die einfache Benutzeroberfläche können Anwender Pre-Modellierung Datenexploration Aufgaben schneller durchführen als das Schreiben von Programmiercode oder einer Zusammenfassung der Daten und der Export der Ergebnisse in ein visuelle Werkzeug, wie z.B. Excel. Mehrere Arten von Diagrammen sind nutzbar, Wie Balkendiagramme, Liniendiagramme, Kreisdiagramme, geografische Diagramme, Baum-und Wärme Karten, tabellarische Ansicht.[SAP14c]

3.3 Predictive Analysis Architektur

Predictive Analysis wird lokal installiert und läuft auf dem Client-Rechner. Predictive Analysis hat eine kleine Bibliothek von integrierten Vorhersagefunktionen für die lineare Regression, Zeitreihenanalyse und Erkennung von Ausreißern. Die Software beruht vorwiegend auf dem lokalen R, HANA PAL und der HANA-R prädiktive Bibliotheken für

die meisten seiner Vorhersagefunktionen. Abbildung zeigt die vollen Predictive Analysis Architektur und Interaktion mit den Datenquellen.[SAP14c]

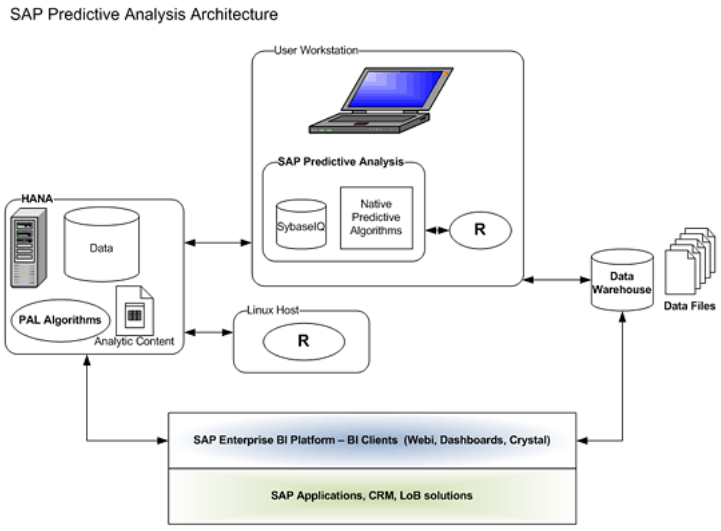


Abbildung 1: SAP Predictive Analysis Architektur [SAP14c]

Predictive Analysis arbeitet in zwei Modi:

- HANA Online-Modus, in dem die Daten und die prädiktive Algorithmen auf HANA gespeichert werden. Sie werden entweder auf dem HANA-System oder einem verbundenen R Linux-Host ausgeführt.
- Offline-Modus, hierbei werden Daten aus einer Datei oder Datenbank auf den Rechner des Benutzers heruntergeladen und auch dort verarbeitet.

Jedes Predictive Analysis-Dokument arbeitet nur im HANA online oder offline Modus und kann nicht verändert werden. Im HANA Online-Modus sind die lokalen R-Algorithmen nicht verfügbar, dafür sind im Offline-Modus die HANA PAL-und R-HANA-Algorithmen nicht verfügbar, auch wenn die Daten ursprünglich von HANA bezogen worden sind.

4 Fazit

Predictive Analysis stellt neue Funktionalität mit attraktiver Optik verpackt dar. SAP hat erkannt, dass viele Verbesserungen erforderlich waren, um Predictive Analysis funktionell vergleichbar an anderen gängigen statistischen Werkzeugen zu machen.[SAP14c] So bietet Predictive Analysis nun ein gutes Werkzeug für die Erstellungen von Vorhersagen. Diese

können auch von Benutzern genutzt werden, die bisher wenig mit Statischen verfahren zu tun hatten. Aber es ist auch für Wissenschaftler interessant durch die Nutzbarkeit von der R-Programmiersprache. Durch die Importfunktionalität können Daten importiert und Aufbereitet werden. Mittels der Möglichkeit der Visualisierung bietet es auch eine gute Reporting Möglichkeit für das Management von Unternehmen. Diese können bei der Entscheidungsfindung durch Prädiktive Analysen unterstützt werden. Jedoch muss natürlich abgewogen werden, welches System für das jeweilige Unternehmen passend ist. So ist es nicht Wirtschaftlich für ein klein Unternehmen ein eigenes SAP HANA-System mit Predictive Analysis zu betreiben. Jedoch kann die Cloud-Lösung hierbei eine sinnvolle Alternative sein. Auch ohne SAP HANA kann das Predictive Analysis genutzt werden und bietet viele Funktionialitäten. Durch die Anbindung an SAP HANA und deren Bibliotheken können die Auswertungen schneller ausgeführt werden. Hierbei muss jedoch ein Augenmerk auf die Online und Offline Funktionalitäten gelegt werden.

Literatur

[RP14] R-Projekt. http://www.r-project.org/about.html, Zugriff Juli 2014, What is R?, 2014.

[SAP14a] SAP. http://de.news-sap.com/2013/09/13/sap-predictive-analysis-1-0-11/, Zugriff Juli 2014, SAP Predictive Analysis 1.0.11, 2014.

[SAP14b] SAP. http://www.sap-cio.de/neue-beitrage/losungen/analytics/sap-predictive-analysis/, Zugriff Juli 2014, SAP Predictive Analysis, 2014.

[SAP14c] SAPEXPERTS. http://sapexperts.wispubs.com/An-Introduction-to-SAP-Predictive-Analysis-and-How-It-Integrates-with-SAP-HANA, Zugriff Juli 2014, An Introduction to SAP Predictive Analysis and How It Integrates with SAP HANA, 2014.

Abschließende Erklärung

Ich versichere hiermit, dass ich meine Seminararbeit „Bewertung des Einsatzes von prediktiven Methoden und Werkzeugen im SAP-Umfeld (SAP Predictive Analysis)“ selbständig und ohne fremde Hilfe angefertigt habe, und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Oldenburg, den 1. August 2014 Jonas Schlemminger

Planungs- und Prognosewerkzeuge mit ihren Stärken und Schwächen

am Beispiel der Systeme BW, BPC und SEM der SAP AG

Seminararbeit
im Rahmen der Projektgruppe *in Memory Planung mit SAP HANA*

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dipl.-Kfm. Martin Donauer Deyan Stoyanov, M.Sc.
Vorgelegt von:	Abdulmasih Hadaya, *abdulmasih.hadaya@uni-oldenburg.de* Mariska Janz, *mariska.janz@uni-oldenburg.de*
Abgabetermin:	02. August 2014

Inhaltsverzeichnis

1 Einleitung (M.J.) 4

2 Unternehmensplanung (A.H.) 5
2.1 Definition . . . 5
2.2 Zeitliche Betrachtung . . . 5
2.3 Unternehmensprognose und der Zusammenhang mit der Unternehmensplanung . . . 6

3 Planungssysteme (A.H.) 8
3.1 Warum Planungssysteme? . . . 8
3.2 Definition . . . 8
3.3 Merkmale von Planungssystemen . . . 8
3.4 Marktsegmente für Planungssysteme . . . 9
3.5 Anbieter von Planungssystemen: Marktüberblick Deutschland . . . 9

4 Die Planungssysteme der SAP AG (M.J.) 10
4.1 Die Grenzen der Unternehmensplanung mit SAP R/3 . . . 10
4.2 Unternehmensplanung mit SAP SEM-BPS . . . 12
4.3 Die SAP NetWeaver-Plattform . . . 14
4.4 Der Wandel von SEM-BPS zur BW-integrierten Planung . . . 15
4.5 Unternehmensplanung mit SAP BW-IP . . . 16
4.6 Unternehmensplanung mit SAP BPC . . . 17
4.7 Die Planungssysteme SAP BW-IP und SAP BPC im Vergleich . . . 18

5 Die verschiedenen Aspekte von Planungssystemen (A.H.) 20
5.1 Beispiel: SAP SEM-BPS . . . 20
5.1.1 Der Aufbau . . . 21
5.1.2 Die Offenheit . . . 21
5.1.3 Die Sicherheit . . . 21
5.1.4 Datenimport und -export . . . 22
5.1.5 Modellierung . . . 22
5.1.6 Planungsinstrumente . . . 23
5.1.7 Präsentation . . . 23
5.2 Beispiel: Applix Interactive Planning . . . 24
5.2.1 Der Aufbau . . . 24
5.2.2 Die Offenheit . . . 24
5.2.3 Die Sicherheit . . . 24
5.2.4 Datenimport und -export . . . 25
5.2.5 Modellierung . . . 25
5.2.6 Planungsinstrumente . . . 25
5.2.7 Präsentation . . . 25

6 Ausblick (M.J.) **26**

Literaturverzeichnis **27**

1 Einleitung (M.J.)

Vor dem Hintergrund des steigenden Konkurrenzdrucks, der fortschreitenden Internationalisierung und der zunehmenden Komplexität im Umfeld der Unternehmen werden integrierte Planungsmethoden immer bedeutender. Auf die Planung über die Geschäftsbereiche hinweg kann sowohl aus operativer Sicht als auch aus strategischer Sicht kaum noch verzichtet werden. Als Verbindung zwischen der strategischen Planung und deren operativer Umsetzung in den Geschäftsprozessen in eine leistungsfähige IT-Unterstützung unumgänglich. Den Kern einer solchen Technologie bilden moderne Planungs- und Prognosesysteme. Die Wahl eines zu den Planungsanforderungen des Unternehmens passenden Werkzeugs und die Kenntnis der spezifischen Stärken und Schwächen ist für den erfolgreichen Einsatz von wesentlicher Bedeutung. (Vgl. [Fis05, S. 13])

Die vorliegende Seminararbeit gibt einen Überblick über den Markt der Planungssysteme und geht dabei mit SEM-BPS, BW-IP und BPC insbesondere auf die Lösungen des Softwareherstellers SAP AG ein. Im Anschluss an die Vorstellung der Systeme werden anhand zwei verschiedener Softwarelösungen die wichtigsten Aspekte von Planungssystemen erläutert. Abschließend erfolgt ein Ausblick auf die Entwicklung der Systeme der SAP AG in der nahen Zukunft.

2 Unternehmensplanung (A.H.)

2.1 Definition

„Die Unternehmensplanung ist der Vorgang der Planung in Wirtschaftsbetrieben, wobei unter Planung die gedankliche Vorwegnahme und Gestaltung zukünftiger Strukturen, Prozesse und Ereignisse verstanden wird.“ [Ehr95]

2.2 Zeitliche Betrachtung

Zeitlich betrachtet unterscheidet man zwischen der operativen, taktischen und strategischen Unternehmensplanung. Ziel der strategischen Planung ist die Erkennung von langfristigen (5-10 Jahre) Chancen und Risiken. Die taktische Planung erfolgt mittelfristig. Aufgaben wie Investitionsvorhaben und Sicherung der strukturellen Liquidität gehören unter anderem zur taktischen Planung. Die operative Planung umfasst die kurzfristige Planung von einzelnen Funktionsbereichen. Hier werden kurzfristige Aufgaben mit detaillierten Daten geplant (vgl. [Dah03, S. 13]).

Dabei kann mit der Top-Down-Methode oder mit der Bottom-Up-Methode geplant werden. Bei der Top-Down-Methode werden zuerst die strategischen Ziele des Unternehmens betrachtet, und darauf basierend wird eine Strategie angelegt. Ausgegangen von dieser Strategie wird bei der taktischen Planung in kleineren Zeiträumen geplant und so weiter bis zur operativen Ebene vorgegangen. Die Bottom-Up-Methode erfolgt ähnlich, aber andersherum. Es werden zuerst die operativen Aufgaben, die zu bestimmten operativen Zielen führen, geplant, und darauf basierend weiter in größeren Zeiträumen taktisch und strategisch geplant. Meistens werden in der Praxis die beiden Vorgehensweisen kombiniert (vgl. [Dah03, S. 14]).

Für die Unterstützung der Planung stehen viele Instrumente zur Verfügung. Beispiele dafür sind die ABC-Analyse bei der Materialbedarfsplanung und die Kreativitätstechniken (vgl. [Dah03, S. 20]).

2.3 Unternehmensprognose und der Zusammenhang mit der Unternehmensplanung

Ziel der Unternehmensprognose ist, aus vergangenen Werten kausale Zusammenhänge zu ermitteln, um eine Vorhersage über die Zukunft zu treffen. Die Prognose kann aber nur mit einer bestimmten Wahrscheinlichkeit geliefert werden, deswegen muss immer bei der Planung mit Unsicherheit gerechnet werden (vgl.[Dah03, S. 21]).

Verschiedene Instrumente könnten für die Ermittlung von Prognosen eingesetzt werden. Beispiele dafür sind die Trendextrapolation, die Regressionsanalyse und die Szenario-Technik. Bei der Trendextrapolation wird eine Prognose über die Zukunft einer Variablen ermittelt. Anhand von Beobachtungen in der Vergangenheit kann die Zukunft der Variable „Kosten“ beispielsweise geschätzt werden. Dabei hängt die Prognose stark von der Trendform ab. Der Planungsträger sollte eine Trendform (z.B. lineare oder nicht lineare Funktion) eingeben. Darauf basierend kann der Computer die Prognose anhand von mathematischen Operationen ermitteln (vgl. [Dah03, S. 21]).

Ein weiteres Instrument ist die Regressionsanalyse. Dabei wird der Zusammenhang zwischen 2 Variablen prognostiziert (z.B. Umsatz und Werbeausgaben). Da andere Faktoren auch eine Rolle spielen können, müssen die potenziellen Fehler abgeschätzt werden (vgl. [Dah03, S. 21]).

Verschiedene Möglichkeiten, ein Ziel zu erreichen, prognostiziert die Szenario-Technik. Störfaktoren könnten hier mit einbezogen werden. Der Planungsträger kann somit die Planung auf dem besten „Weg“ zum Ziel weiter verfolgen (vgl. [Dah03, S. 21]).

Die Unternehmensplanung und -prognose sind nicht separat zu sehen. Die Prognose spielt eine wichtige Rolle bei der Planung. Die Planung ohne Prognose reduziert sich auf Wahrsagerei (vgl. [Dah03, S. 21]). Die Phasen der strategischen Planung sind: die strategische Zielsetzung, die Analyse-Phase, die Strategieformulierung und die Strategieimplementierung. Die Prognose ist eine Funktion der Analyse-Phase. Sie unterstützt eine aussagekräftige Analyse, was als Basis für die Strategieformulierung dient (vgl. [Wel07]).

Die folgende Abbildung zeigt die beidseitigen Informationsflüsse zwischen der Unternehmensanalyse, der Umweltanalyse und dem Prognoseinstrument (z.B. Szenario-Technik):

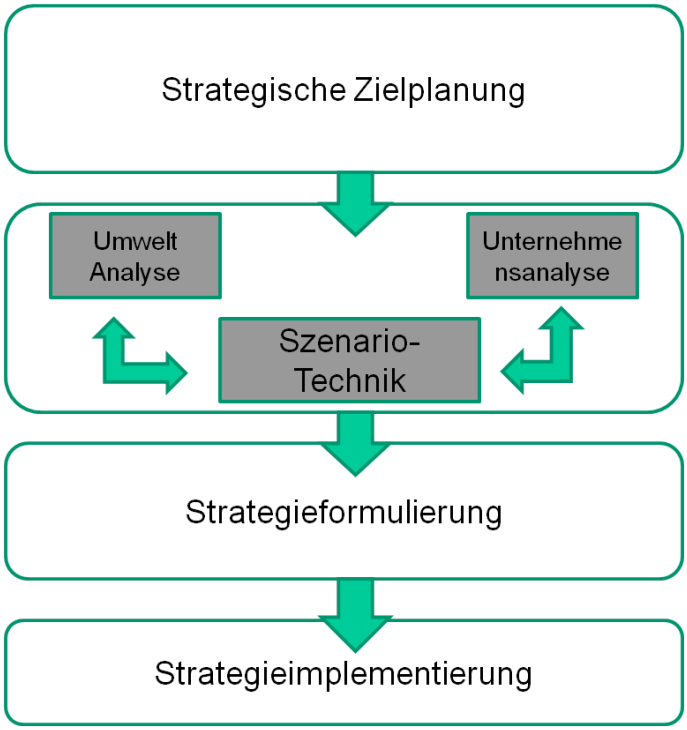


Abbildung 1: Zusammenhang zwischen Unternehmensplanung und-prognose[1]

[1] Quelle: [Wel07]

3 Planungssysteme (A.H.)

3.1 Warum Planungssysteme?

Die Unternehmensplanung ist eine komplexe Aufgabe. Es gibt viele Ziele mit deren Abhängigkeiten. Diese sind für den Planer schwer zu durchschauen. Aus dieser Komplexität ergeben sich viele Entscheidungsvarianten, die für den Entscheidungsträger schwer zu erkennen sind (vgl. [Han, S. 3]).

Aus diesen Gründen wird die IT-Unterstützung benötigt. Die ersten Versuche waren die MIS (Management Information Systems) Anfang der sechziger Jahre, dann folgten die DSS (Decision Support Systems) und danach die EIS (Executive Information Systems). Bei diesen Systemen ging es lediglich um die Auswertung von Ist-Daten. Das Zurückschreiben von Planzahlen auf die Datenbestände war immer noch nicht möglich. Anfang der neunziger Jahre begann die Verbreitung von Business Warehousing und der Online Analytical Processing (OLAP) Methoden. Diese bildeten die Basis für die heutigen Planungssysteme (vgl. [Dah03, S. 3]).

3.2 Definition

„Ein Planungssystem ist ein Anwendungssystem, welches darauf abzielt, dass zum einen die bei der Planung von Unternehmensprozessen verwendeten Daten zuverlässig sind und zum anderen mehr Alternativen innerhalb des Planungsprozesses betrachtet werden können.“ [Mer05]

3.3 Merkmale von Planungssystemen

Planungssysteme ermöglichen sowohl die dezentrale als auch die zentrale Planung. Der Anwender hat einen schreibenden Zugriff, um Plandaten auf der DB zu sichern. Der gesamte Ablauf der Planung kann mit dem Planungssystem unter Berücksichtigung von unterschiedlichen Zeithorizonten koordiniert werden. Planungssysteme bieten die Möglichkeit für ein Forecasting und für die Simulation von Planszenarien. Jedes Planungssystem hat ein Berechtigungskonzept, das die Art und Weise der Einstellung von Benutzerrechten und -rollen bestimmt (vgl. [Dah03, S. 55]).

3.4 Marktsegmente für Planungssysteme

Den Markt für Planungssysteme kann man wie folgt segmentieren (vgl. [Dah03, S. 4]):

1. Einzelplatzlösungen: Diese Software liefert vollständige aber unzureichende Funktionalitäten und eignen sich deswegen für die Planung von Einzel- und Kleinunternehmen. Diese Lösungen sind meistens Excel-basiert. Die Kosten liegen bei wenigen hundert Euro.

2. Standardwerkzeuge: erweitern die Funktionalitäten von Einzelplatzlösungen. Sie enthalten z.B. ein Berechtigungskonzept. Der Preis solcher Lösungen liegt zwischen 1000 und 10000 Euro.

3. Planungsplattformen: Die höchste Flexibilität bieten die Planungsplattformen an. Sie ergänzen ERP-Systeme und bieten umfangreiche Planprozesse an. Der Preis ist ab 10000 Euro anzusetzen und steigt enorm mit den Beratungskosten.

3.5 Anbieter von Planungssystemen: Marktüberblick Deutschland

SAP mit ihren Produkten für Planungssysteme führt den Markt in Deutschland mit ca. 22,6% Marktanteil an. IBM folgt auf dem zweiten Platz mit ca. 12,6% Marktanteil, SAS auf dem dritten Platz mit ca. 11,2%, dann Oracle mit ca. 6,9%, und die restlichen Anbieter haben ca. 41,9% des Marktanteils inne (vgl. [Fuc, S. 9]).

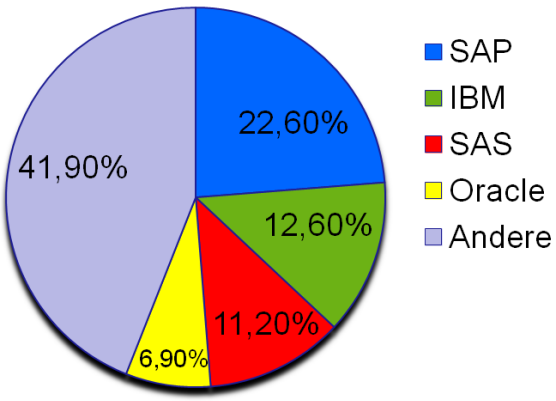


Abbildung 2: Anbieter für Planungssysteme- Marktüberblick: Deutschland

4 Die Planungssysteme der SAP AG (M.J.)

4.1 Die Grenzen der Unternehmensplanung mit SAP R/3

1992 kommt *SAP R/3* auf den Markt (vgl. [SAPb]). R/3 ist ein ERP-System und dient somit „der funktionsbereichsübergreifenden Unterstützung sämtlicher in einem Unternehmen ablaufenden Geschäftsprozesse“ [Vah]. Zusammen mit ihren Nachfolgern *mySAP ERP* und *SAP ERP*[2] gilt die Software R/3 als das wohl bekannteste ERP-System überhaupt (vgl. [Vah]). Wie für ERP-Lösungen üblich, ist R/3 ein transaktionales System. Daher basiert es auf dem *Online-Transaction-Processing-Paradigma* (OLTP), einer „im Dialogbetrieb ablaufende Massendatenverarbeitung in operativen DV-Systemen, bei der betriebswirtschaftliche Transaktionen erfasst und verarbeitet werden“ [Lacb].

Das R/3-System kann für die operative Unternehmensplanung genutzt werden. Im Rahmen des Controllings ist eine Planung im Sinne der Fortschreibung von Istdaten möglich. Beispiele hierfür sind die Planung der Kosten für bestehende Fertigungsmethoden und die Planung der Erlöse in bestimmten Kundensegmenten. (Vgl. [Brü09, S. 25])

Als typisches ERP-System besteht die Software R/3 aus unterschiedlichen Modulen, die sich jeweils auf einen bestimmten Anforderungsbereich im Unternehmen beziehen. Aus der Perspektive des Controllings sind die folgenden Module von besonderer Bedeutung (vgl. [Brü09, S. 35 f.] und [Sch00, S. 372]):

- SD (Sales and Distribution) - Vertrieb
- MM (Material Management) - Materialwirtschaft (Planung der Kostenträger)
- PP (Production Planning) - Produktionsplanung und -steuerung
- FI (Financials) - Finanzwesen
- CO - Controlling (Absatz- und Umsatzplanung; Leistungs- und Kostenplanung der Prozesse, Kostenstellen und Innenaufträge; Planung der Deckungsbeiträge und Ergebnisse)

[2] Aus dem R/3-System wurde 2004 *mySAP ERP*. Seit 2007 heißt das System nur noch schlicht *SAP ERP*. (Vgl. [FGSK08])

Abbildung 3 zeigt eine Übersicht über die verschiedenen Module des R/3-Systems.

Abbildung 3: Übersicht über die Module des SAP R/3-Systems[3]

Bei der Planung mit R/3 sollte der Planungshorizont nicht über das Folgejahr hinausgehen. Das bedeutet, dass das ERP-System zwar für die operative Planung geeignet ist, für die taktische und vor allem strategische Planung aber nicht. Da ein ERP-System aufgrund seines OLTP-Charakters ein transaktionales und kein analytisches System ist, ist es das falsche Werkzeug für eine längerfristige Planung. (Vgl. [Brü09, S. 25])

[3] Quelle: [sap06]

4.2 Unternehmensplanung mit SAP SEM-BPS

Ende der Neunzigerjahre bringt SAP das *Strategic Enterprise Management* (SEM) als Tool zur Unterstützung strategischer Entscheidungen auf den Markt. SEM soll sich von den transaktionalen Prozessen des R3-Systems abheben. Aus diesem Grund ist das Werkzeug auch kein Teil des R/3-Systems, sondern eine Sammlung von Funktionen, die auf dem *SAP Business Information Warehouse* (BW) aufbauen. Das BW ist ein klassisches Data-Warehouse-System und basiert demzufolge auf den Prinzipien des *Online Analytical Processing* (OLAP). OLAP ist ein „Konzept für die im Dialogbetrieb realisierte Verdichtung und Darstellung von managementrelevanten Daten aus einem Data Warehouse“ [Laca] und kann als Gegenstück zum OLTP verstanden werden. Das BW ist für die Nutzung von SEM zwingende Voraussetzung. (Vgl. [Fis05, S. 132/135])

Abbildung 4 zeigt den Aufbau des BW. Zuunterst ist das Vorsystem dargestellt, aus dem die Daten in das Data Warehouse geladen werden. Bei diesem System kann es sich beispielweise, aber nicht zwingend um SAP R/3 handeln. In der ETL-Schicht erfolgt das Extraktion, die Transformation und das Laden der Daten in das Data Warehouse. Im Data Warehouse werden die Daten in sogenannten Datenwürfeln multidimensional gespeichert. Mit Hilfe von OLAP können die Daten auch multidimensional analysiert werden. Die durch OLAP aus dem BW gelesenen Informationen können nach dem Auslesen in SEM verwendet werden.

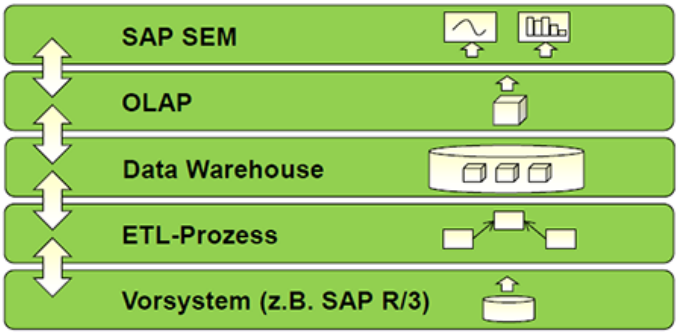


Abbildung 4: Business Information Warehouse (BW)[4]

Einer der fünf Bestandteile des SEM ist *Business Planning and Simulation* (SEM-BPS). Diese Komponente besteht aus einer Reihe an Werkzeugen zur Modellierung von individuellen Planungsszenarien. Die enthaltenen Werkzeuge sind typische Planungsfunktionen (wie zum Beispiel die Top-down-Verteilung) und betriebswirtschaftliche Funktionen (wie

[4]Quelle: In Anlehnung an [Kie10, S. 84]

zum Beispiel die Bestandsrechnung). Außerdem umfasst SEM-BPS Methoden zur Planungssimulation, unter anderem in Form von Prognoseverfahren. Die Modellierung der Planungsprozesse wird dabei durch vorkonfigurierte Planungsanwendungen (beispielsweise zur Investitionsplanung) unterstützt. Die manuelle Eingabe von Plandaten ist auch über eine Excel-Schnittstelle und ein Web-Interface möglich. SEM-BPS beinhaltet allerdings keine Funktionen zur Konsolidierung, daher ist die Vereinheitlichung der Einzelabschlüsse in einem Konzern mit SEM-BPS nicht möglich. Für diesen Anwendungszweck muss auf die SEM-Komponente *Business Consolidation* oder kurz *SEM-BCS* zurückgegriffen werden. (Vgl. [Fis05, S. 133 f.])

Abbildung 5 zeigt die fünf verschiedenen, auf dem BW und dem *Knowledge Warehouse* (KW) aufbauenden SEM-Komponenten:

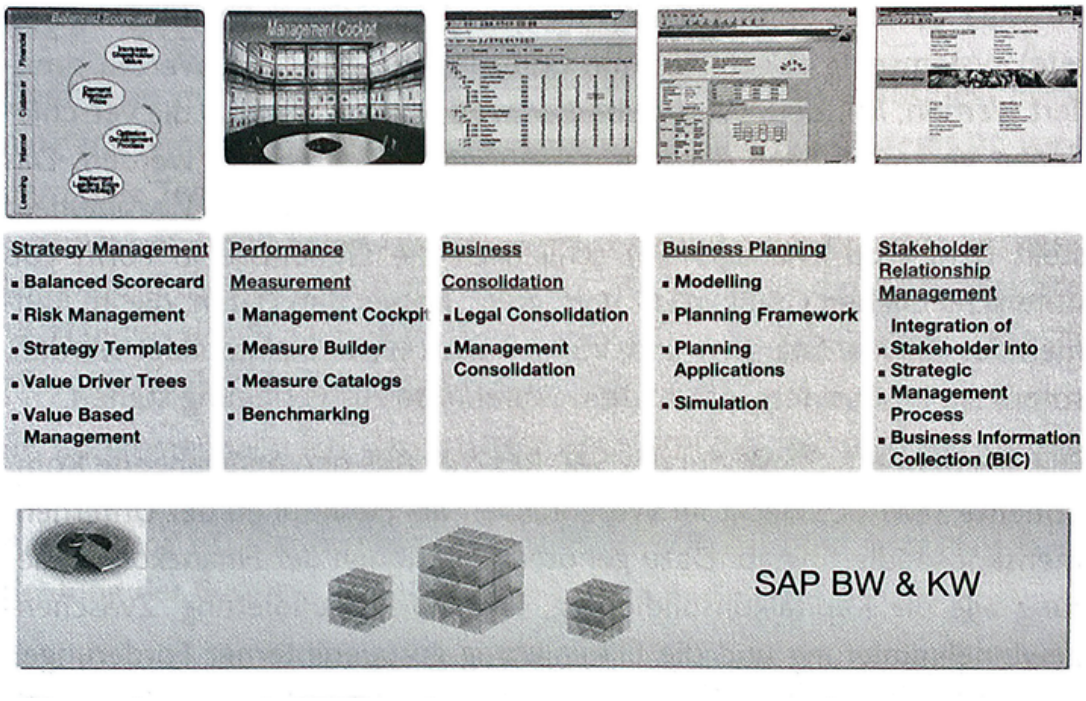


Abbildung 5: Die fünf Komponenten des SAP SEM[5]

Eine genauere Beschreibung der für die Planung mit SEM-BPS relevanten Aspekte erfolgt in Kapitel 5.1.

[5]Quelle: [Fis05, S. 133]

4.3 Die SAP NetWeaver-Plattform

2004 erscheint die erste Version des SAP NetWeaver, einer Technologieplattform, die der Integration aller Prozesse in einem Unternehmen dient und auch als „Universelles Betriebssystem von SAP“ [Kie10, S. 85] bezeichnet wird. (Vgl. [SAPc])

Die NetWeaver-Plattform beinhaltet den Bereich *Business Intelligence* (SAP BI), hinter dem sich das BW mit dem Data Warehouse, einer BI-Plattform, der *BI Suite* sowie speziellen Entwicklertechnologien verbergen (vgl. [SAPa]). Die BI Suite stellt mit dem *Business Explorer* eine Reportinglösung zur Verfügung, die als Frontend genutzt wird (vgl. [Kie10, S. 88]).

Abbildung 6 zeigt den SAP NetWeaver und den darin enthaltenen Bereich *Business Intelligence* mit dem Data Warehouse, der BI-Plattform, der BI Suite sowie den Entwicklertechnologien.

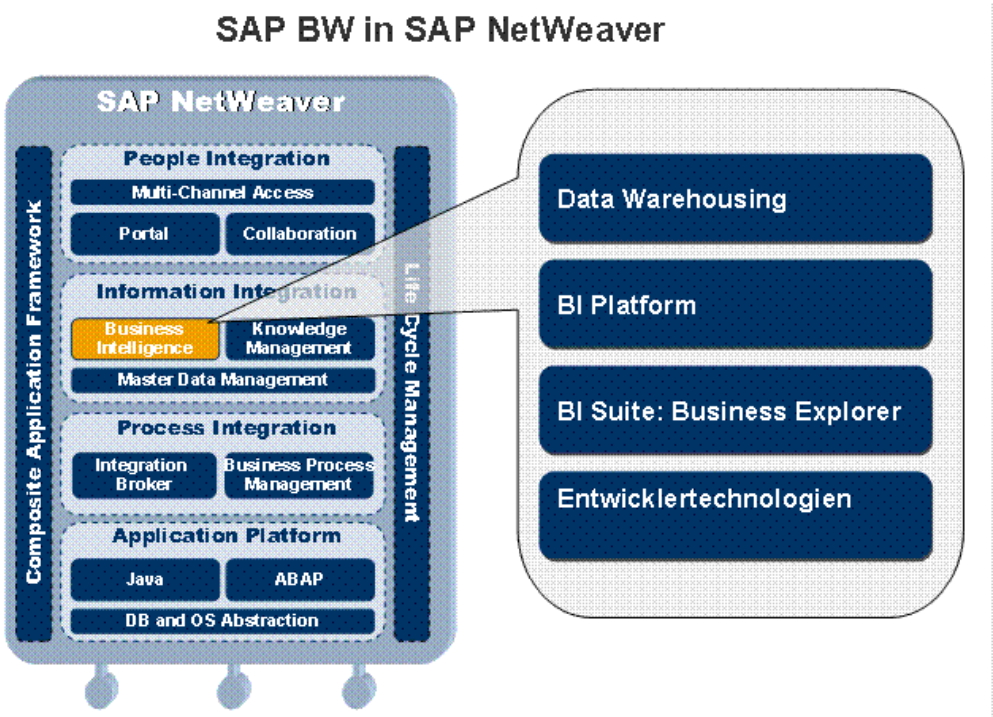


Abbildung 6: SAP NetWeaver mit BI-Bereich[6]

4.4 Der Wandel von SEM-BPS zur BW-integrierten Planung

Da die Planungsfunktionalität nicht mehr nur als Unterstützung der finanzorientierten Planung gesehen werden soll, gibt SAP 2003 bekannt, die zu dem Zeitpunkt noch in SEM angesiedelte Unternehmensplanung in das BW zu verlagern und somit in den BI-Bereich des NetWeaver zu integrieren. So kommt SAP gleichzeitig dem Kundenwunsch nach, ein generisches Planungswerkzeug für alle Planungsbereiche zu stellen. (Vgl. [Fis05, S. 138])

SAP führt in den nachfolgenden Jahren eine vollständige Integration der BPS-Planung in die technische Lösung des BW durch. Die neue Lösung der integrierten Planung ist seitdem als *BW-IP* bekannt.

Abbildung 7 zeigt die in 2004 begonnene Verlagerung. Als Zwischenschritt besteht um das Jahr 2006 die Lösung des *BW-BPS*, die als Übergangslösung dient. BPS ist zu diesem Zeitpunkt noch nicht voll in das BW integriert, sondern hat lediglich die Plattform gewechselt.

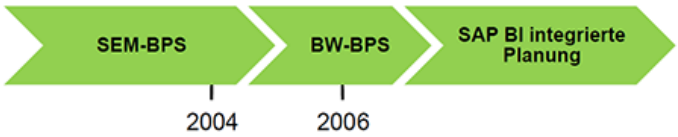


Abbildung 7: Von SEM-BPS zur BI-integrierten Lösung[7]

Im Zuge des Wandels findet auch ein umfassendes Redesign der von BPS gebotenen Planungsfunktionalitäten statt (vgl. [Fis05, S. 142]).

Auch wenn der größte Teil der SEM-BPS-Funktionalitäten in die neue Lösung überführt wird, so verbleibt dennoch alle finanzorientierten Planungsfunktionen im SEM-Modul. Dies ist darin begründet, dass einige dieser Funktionalitäten aus der BCS-Komponente stammen. Daher besitzen sie zum einen eine andere technische Ausprägung und werden zum anderen auch weiterhin für SEM-BCS gebraucht. (Vgl. [Fis05, S. 140])

[6] Quelle: [SAPa]
[7] Quelle: In Anlehnung an [Kie10, S. 87] und [Fis05, S. 139]

4.5 Unternehmensplanung mit SAP BW-IP

In SAP BW-IP sind die Planungs- und BI-Funktionalitäten vollständig integriert. Diese Lösung bietet daher diverse Vorteile. Neben einer einheitlichen und konsistenten Datenbasis gibt es vor allem auch eine einheitliche Prozess- und Modellierungslogik, eine einheitliche Benutzeroberfläche sowie einheitliche Design-Werkzeuge. (Vgl. [MuC09, S. 285 f.] und [Egg07, S. 57 f.])

Insbesondere die Vereinheitlichung der Oberfläche, der Prozesslogik sowie der Tools bedeutet für den Anwender eine große Verbesserung. Zuvor unterschieden sich die Planungsfunktionalitäten deutlich von den Funktionen im BI-Bereich. Die Anwender mussten daher mit zwei Systemen arbeiten, die aufgrund der verschiedenen Bedienkonzepte auf unterschiedliche Weise zu bedienen waren. Im Zuge der Integration in BW-IP ergibt sich durch die Angleichung der Oberflächen und der Prozesslogik für Planungs- und Analysefunktionen ein einheitliches Bedienkonzept und Look and Feel für die Anwender. Die Layouts für die Planung haben nun die gleiche Basis wie die Layouts für die Berichte im BW. So können diese zum einen gleich eingestellt werden und sind zum anderen auch auf die gleiche Art und Weise benutzbar.(Vgl. [Fis05, S. 142])

Durch das Redesign der Planungslösung sind die Benutzerschnittstellen auch insgesamt flexibler und anwendungsfreundlicher geworden. Infolge der Vereinheitlichung der Werkzeuge und der Aufhebung des Bruchs zwischen den Systemen hat sich die Bedienung vereinfacht, die Benutzerfreundlichkeit gesteigert und zudem auch der Schulungsaufwand verringert. (Vgl. [Egg07, S. 58] und [Fis05, S. 142 f.])

Das Konzept der BW-integrierten Planung sorgt zudem für eine kürzere Implementierung, eine bessere Performance und auch geringere Wartungskosten. Durch das umfassende Redesign der Planungsfunktionalitäten bietet BW-IP sowohl Excel- als auch webbasiert eine flexiblere und einfachere Modellierung. Des Weiteren sind auch zusätzliche Planungsfunktionen hinzugekommen. (Vgl. [MuC09, S. 285 f.], [Egg07, S. 57 f.] und [Fis05, S. 143])

Durch die vereinfachten Strukturen können nun auch Fachkräfte aus dem Business-Bereich die Planungsmodellierung vornehmen. Zuvor war dies meist den Mitarbeitern aus der IT-Abteilung vorbehalten. Durch die Vereinfachung der Bedienung verschieben sich auch die Verantwortlichkeiten in Bezug auf die Wartung von der IT-Abteilung zum Bereich Business (vgl. [Fis05, S. 142 f.]).

4.6 Unternehmensplanung mit SAP BPC

Im Jahr 2007 übernimmt die SAP AG die Softwareanbieter OutlookSoft und Business Objects. In der Folge des Zukaufs erweitert SAP sein Portfolio um das Produkt *BusinessObjects Planning and Consolidation* (BPC). (Vgl. [Fol13, S. 26])

SAP bietet BPC in zwei verschiedenen Varianten an, zum einen eine Microsoft-Version und zum anderen eine neu entwickelte NetWeaver-Version. Die Microsoft-basierte Version weist im Gegensatz zur NetWeaver-Variante einen hohen Reifegrad auf, da sie bereits viele Jahre von OutlookSoft vertrieben wurde. (Vgl. [Kie10, S. 96])

In Bezug auf die enthaltenen Funktionalitäten unterscheiden sich die beiden Versionen voneinander. Abbildung 8 stellt die Unterschiede und Gemeinsamkeiten dar (Stand: BPC 7.0). In der Schnittmenge der beiden Kreisflächen sind die gemeinsam genutzten Ressourcen abgebildet. Der Bereich außerhalb der Kreise steht für die allgemeinen Verwaltungsfunktionen des BPC-Systems, die in beiden Versionen gleichermaßen verfügbar sind. (Vgl. [Kie10, S. 99])

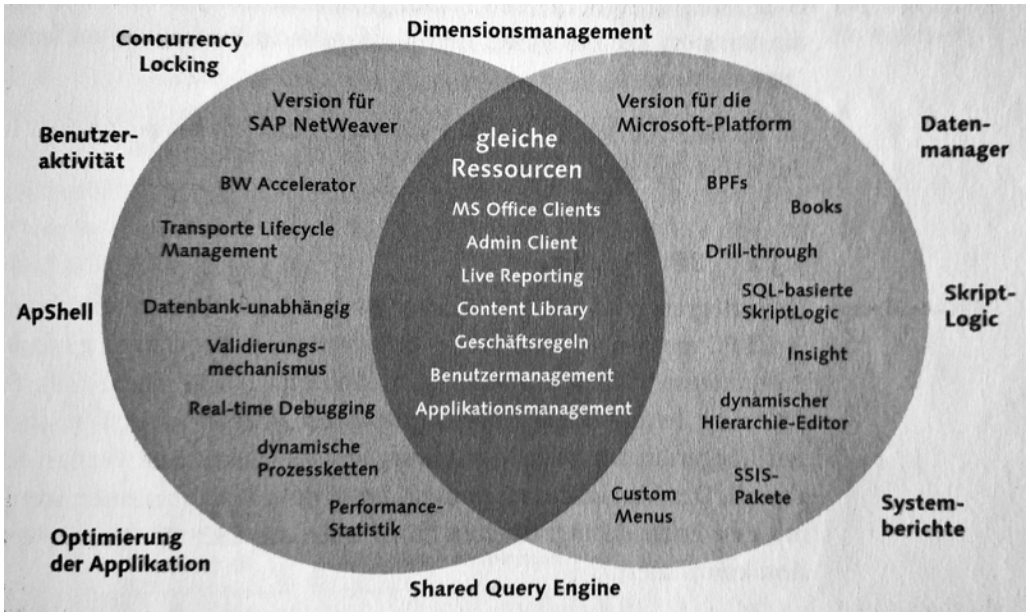


Abbildung 8: Die Funktionalitäten der beiden SAP BPC-Versionen im Vergleich[8]

[8] Quelle: [Kie10, S. 99], Stand: BPC 7.0

4.7 Die Planungssysteme SAP BW-IP und SAP BPC im Vergleich

Wie BW-IP dient auch BPC der Verkürzung von Planungszyklen, der Reduzierung von manuellen Tätigkeiten und der Erhöhung der Planungsqualität (vgl. [Fol13, S. 26]). Jedoch verfolgen beide Systeme unterschiedliche Anwendungsphilosophien. Während sich BW-IP insbesondere für die zentrale unternehmensweite Planung eignet, richtet sich BPC vor allem an Unternehmen, die eine dezentrale Fachbereichs- und Abteilungsplanung bevorzugen. Dies führt dazu, dass sich BW-IP zwar besser für eine einheitliche Planung über die Geschäftsbereiche hinweg eignet, BPC hingegen aber deutlich mehr Flexibilität und Benutzerfreundlichkeit für Fachanwender bereit hält. (Vgl. [Fol13, S. 26])

Die IT- und Backend-getriebene Umsetzung des BW-IP bietet Unternehmen eine vollständige Data-Warehouse- und SAP-Integration sowie ein hohes Maß an Stabilität, Performance, Sicherheit, Konsistenz und Skalierbarkeit. Neben der Administration erfolgt auch die Definition von Planungsprozessen und Reports zentral durch die IT. Die Frontendgetriebene Umsetzung von BPC bietet den Fachbereichen die Integration in Microsoft Office. Außerdem ist ein eigenständiges Reporting und ein flexibles Definieren von Planungsprozessen möglich. Die Administration findet dezentral durch die Endnutzer in den Fachabteilungen statt. Während in BPC auch Funktionen zur Konsolidierung enthalten sind, müssen BW-IP-Nutzer zu diesem Zweck auf andere Systeme (wie beispielsweise SEM-BCS) zurückgreifen. (Vgl. [Fol13, S. 26])

Fällt die Entscheidung für BW-IP, dann müssen Unternehmen mit einer tendenziell langsameren Einführung sowie höheren Kosten für die Implementierung rechnen. Durch die bereits vorhandene SAP-ERP-Lizenz fallen anders als bei BPC jedoch keine Lizenzgebühren an. Dies ist der Grund, warum BPC trotz der geringeren Implementierungskosten als kostenintensivere Wahl gilt. (Vgl. [Fol13, S. 26])

Sowohl BW-IP- als auch BPC-Nutzer haben die Wahl zwischen einer Web-Oberfläche und Excel für Planung und Reporting. Beide Systeme bieten seit diesem Jahr auch die Möglichkeit zur mobilen Nutzung, BW-IP durch das „BusinessObjects Design Studio“ und BPC durch die sogenannte „EPM Unwired Mobile App“. (Vgl. [Fol13, S. 27])

Tabelle 1 stellt die wichtigsten Unterschiede der beiden Systeme noch mal im direkten Vergleich gegenüber.

	BW-IP	**BPC**
Ansatz	Für zentrale unternehmensweite Planung	Für dezentrale Fachbereichs- und Abteilungsplanung
Ausrichtung	Über die Geschäftsbereiche hinweg einheitlich planen	Flexibilität und Benutzerfreundlichkeit für Fachanwender
Umsetzung	IT- und Backend-getriebene Umsetzung: volle Data-Warehouse- und SAP-Integration	Fachbereichs- und Frontendgetriebene Umsetzung: Integration in MS Office
Planungsprozesse	Planungsprozess ist vordefiniert	Planungsprozess ist flexibel
Administration	Administration zentral durch IT	Administration dezentral durch Enduser
Reporting	Vordefinierte Reports	Eigenständiges Reporting
Konsolidierung	Keine Konsolidierung	Konsolidierung implementiert
Einführung	Tendenziell langsamere Einführung	Tendenziell schnellere Einführung
Kosten	Höhere Kosten für Implementierung, aber durch die SAP-ERP-Lizenz keine Lizenzkosten	Geringere Kosten für Implementierung, dafür jedoch Lizenzkosten

Tabelle 1: BW-IP und BPC im Vergleich[9]

Unternehmen müssen nicht zwangsläufig eine Wahl zwischen BW-IP und BPC treffen, denn bei Bedarf ist theoretisch auch eine parallele Nutzung möglich. Werden beide Systeme zur Planung eingesetzt, dann können die Pläne am Ende zusammengeführt werden (vgl. [Fol13, S. 27]).

[9]Quelle: In Anlehnung an [Fol13, S. 26]

5 Die verschiedenen Aspekte von Planungssystemen (A.H.)

5.1 Beispiel: SAP SEM-BPS

Business Planning and Simulation (BPS) wird in diesem Teil anhand von bestimmten Aspekten weiter erläutert. Diese Aspekte sind: Aufbau, Offenheit, Sicherheit, Datenimport und -export, Modellierung, Planungsinstrumente und Präsentation.

Der Aufbau: Das ist die Systemarchitektur. Diese beschreibt, aus welchen Teilen das System besteht. Das System kann im Web geführt werden oder als Client/Server. Der Aufbau sollte auch beschreiben, wo und wie die Daten gespeichert werden und welche Integrationsmöglichkeiten in andere Anwendungen möglich sind.

Die Offenheit: beschreibt die Möglichkeiten, die das System für den externen Zugriff anbietet. Hier wird beschrieben, welche APIs für den Aufbau von individuellen Applikationen zur Verfügung stehen.

Die Sicherheit: beschreibt das Berechtigungskonzept des Systems und welche Möglichkeiten es für die Definition von Rollen bietet.

Der Datenimport und -export: beschreiben, wie und in welchen Formaten Daten ins System importiert oder aus dem System exportiert werden können.

Die Modellierung: Dieser Aspekt beschreibt, wie und mit welchen Tools das Datenmodell die organisatorischen und die betriebliche Strukturen im Unternehmen abbildet und das stellt die Planungsobjekte dar.

Die Planungsinstrumente: Dieser Aspekt umfasst die Instrumente, die das System anbietet und die für die Planung relevant sind, wie z.B. Prognose- oder Simulationstools.

Die Präsentation: Die Präsentation beschreibt, welche Möglichkeiten es im System gibt, die Ergebnisse zu präsentieren, z.B. im Form von Berichten oder Grafiken (vgl. [S. 69]Dahnken.2003).

5.1.1 Der Aufbau

Das Business Warehouse dient als zentraler Applikationsserver für alle SEM Applikationen und darunter auch BPS. BW führt die Kommunikation mit der DB und verwaltet die Meta- und die Bewegungsdaten. Die Datenspeicherung erfolgt ebenfalls im BW.

Das System kann als Client/Server oder als Webanwendung angelegt werden. Bei der Client/Server Architektur werden Frontend-Komponenten auf der Client-Seite benötigt. Die Integration von Microsoft Excel im SAPGUI ermöglicht die Durchführung von Planungsfunktionen, die Microsoft Excel benötigen. Falls das System als Webanwendung eingeführt werden soll, wird auf der Client-Seite keine Software benötigt und die Server-Seite um einen Internetserver ergänzt.

Eingegebene Werte im System werden in InfoCubes gehalten. Diese InfoCubes sind relationale Tabellen, die in Form eines Starschemas im BW gespeichert sind. Mit Hilfe eines geöffneten BPS-Puffers können die Planwerte in den InfoCubes gehalten werden, um es dem Anwender zu ermöglichen, Plansituationen zu simulieren. Sobald der Puffer geschlossen wird, werden die Planwerte endgültig auf der Datenbank gesichert (vgl. [Dah03, S. 238]).

5.1.2 Die Offenheit

Strukturierte Dokumente, wie z.B. Word oder PowerPoint, oder Dokumente in Binärformaten können über die Business Document Service (BDS) im SEM genutzt werden.

Über die Business Application Programming Interface (BAPI) können die Anwender die Funktionalität des Systems um weitere Eigenschaften ergänzen. Die BAPI bieten Schnittstellen, die die Nutzung von Funktionselementen erlauben. Die erfassten Plandaten könnten ins operative R/3 zurückgeschrieben werden (vgl. [Dah03, S. 239]).

5.1.3 Die Sicherheit

Das BW bildet das Sicherheitskonzept für alle SEM-Komponenten. Einschränkungen können über die Rollenzuordnung und die benutzerspezifischen Einstellungen definiert werden. Zusätzlich könnten im BPS weitere Maßnahmen getroffen werden, um die Planungsapplikation abzusichern. Es könnten Merkmalsbeziehungen definiert werden, z.B. (Land und

Währung), so wird die Planung in bestimmten Ländern auf bestimmte Währungen beschränkt. Sowohl die Sicht als auch der Zugriff auf diese Merkmale können im BPS eingeschränkt werden (vgl. [Dah03, S. 239]).

5.1.4 Datenimport und -export

Im BW wird der Datenimport durchgeführt. Das BW bietet Schnittstellen zu SAP R/3, ASCII und XML Dateien an. Der Import aus den unterstützten relationalen Datenbanken erfolgt über DB Connect.

Über die Staging Business Application Programming Interface (Staging BAPI) können Daten aus Fremdsystemen importiert werden. Wie beim Datenimport verfügt BW über eigene Schnittstellen für den Datenexport (vgl. [Dah03, S. 240]).

5.1.5 Modellierung

Während das Datenmodell die Datenstrukturen im Unternehmen abbildet, bildet das Planungsmodell die Planungsobjekte ab. Das Datenmodell wird in Form von InfoCubes in der Administrator Workbench erstellt, wobei das Planungsmodell in der Planning Workbench erstellt wird.

Das Planungsmodell ist hierarchisch aufgebaut: Die erste Ebene ist das Planungsgebiet. Das ist ein abgegrenztes betriebswirtschaftliches Feld, wie z.B. die Vertriebsplanung. Die zweite Ebene ist die Planungsebene, innerhalb derer Planungsgebiete eingeschränkt werden können. Ein Beispiel dafür ist die Einschränkung von Produktgruppenplanung nach Regionen. Auf dieser Ebene können auch Planungsfunktionen definiert werden. Auf der untersten Ebene befinden sich die Planungspakete. Diese schränken die Merkmale und die Kennzahlen der Planungsebene ein. Anpassbare Standardmodelle stehen im BPS zur Verfügung, wie z.B. Vertriebsplanung oder Personalplanung. Diese können vom Anwender angepasst werden für den Aufbau von eigenen Planungsanwendungen. Das Datenmodell kann von Anwendern je nach Benutzerberechtigung geändert werden. Solche Änderungen werden im Planungsmodell automatisch übernommen (vgl. [Dah03, S. 240]).

5.1.6 Planungsinstrumente

Im BPS stehen generische und anwendungsspezifische Planungsfunktionen zur Verfügung. Die generischen könnten wie folgt klassifiziert werden: Simulationsfunktionen, manuelle Planungsfunktionen, Allokationsfunktionen, Funktionen für die individuelle Kalkulation von Planwerten und Funktionen für die Datenübernahme. Beispiele dafür sind die Prognose- und die Verteilungsfunktionen. Bei der Prognosefunktion werden aus vorhandenen Daten Plandaten als Forecast generiert. Dabei werden unterschiedliche Prognosestrategien unterstützt, wie z.B. exponentielle Glättung. Bei der Verteilungsfunktion werden nicht zugeordnete Werte oder Mengen auf Basis von Referenzdaten, z.B. der Saisonverteilung, verteilt.

Die vordefinierten anwendungsspezifischen Planungsfunktionen ergänzen die generischen Funktionen für z.B. Abschreibung und Allokation.

Sollten die generischen und die anwendungsspezifischen Funktionen nicht ausreichen, können mit einem Formeleditor algebraische und statistische Formeln definiert werden (vgl. [Dah03, S. 245]).

5.1.7 Präsentation

Die Datenwerte werden in den Planungslayouts innerhalb der Planungsworkbench angezeigt. Der grundsätzliche Aufbau der Layouts wird vom Administrator vorgegeben und kann vom Anwender nicht geändert werden. Falls der Administrator dem Anwender die Berechtigung erteilt, kann der Anwender im Rahmen dieser Berechtigung weitere Selektionen vornehmen, wie z.B. die Einstellung einer Produktgruppe.

Es stehen dem Anwender im BPS Grafikfunktionalitäten zur Verfügung, und es können auch externe Grafiken bezogen werden.

Das Berichtswesen im BPS wird über das BW abgewickelt. Abfragen können in einer Excel- Umgebung getätigt werden. Web Applikationen für das Berichtswesen können über den Business Explorer Web Application Designer veröffentlicht werden (vgl. [Dah03, S. 246]).

5.2 Beispiel: Applix Interactive Planning

Dieses Beispiel dient dazu, andere Möglichkeiten für die Implementierung von Planungssystemen aufzuzeigen und den Vergleich mit den SAP-Lösungen überschaubar zu machen. Applix Interactive Planning ist das Planungswerkzeug der Applix GmbH München. Basierend auf denselben Aspekten wie bei SAP SEM-BPS werden die Eigenschaften dieses Systems behandelt (vgl. [Dah03, S. 113]).

5.2.1 Der Aufbau

Die multidimensionale Datenbank TM1 ist die niedrigste Ebene des Systems. Während die Plandaten in TM1 gespeichert werden, steht eine relationale DB für die Speicherung von Informationen über Planprozesse und Berechtigungen zur Verfügung. Das Anfragen der Clients wird über einen Applikationsserver an die TM1 weitergeleitet.

Auf der zweiten Ebene befindet sich Applix Integra. Dieses bildet die Basis des Systems, denn Applix Integra verfügt über die benötigten Funktionalitäten für die Administration des Systems.

Auf der obersten Ebene steht das Applix Interactive Planning Frontend. Dieses ermöglicht den Zugriff auf die TM1 und die Planungsfunktionalität von Integra (vgl. [Dah03, S. 114]).

5.2.2 Die Offenheit

Die Entwicklungskomponente Developer‘s Studio ermöglicht die freie Definition von Oberflächen. TM1 bietet eine API, die den Zugriff auf den OLAP Engine erlaubt. Somit könnten die Werkzeuge von Applix mit C, C++ und Visual Basic durch individuelle Applikationen ergänzt werden (vgl. [Dah03, S. 115]).

5.2.3 Die Sicherheit

Applix Integra ist die Basis für das Berechtigungskonzept des Systems. Die Administration erfolgt über das Web-Frontend, und die Berechtigungen werden in der relationalen DB gespeichert (vgl. [Dah03, S. 115]).

5.2.4 Datenimport und -export

Der Import erfolgt über die TM1. Daten können aus ASCII Dateien und relationale DB über eine ODBCD-Schnittstelle importiert werden. Die Daten werden durch einen TurboIntegrator transformiert. Über die Planungsoberfläche können die Daten ins Excel-Format exportiert werden (vgl. [Dah03, S. 115]).

5.2.5 Modellierung

TM1 bietet spezielle Benutzeroberfläche für die Modellierung der Unternehmensstrukturen. Die Modellierung der Planung erfolgt auch in TM1. Über die Komponente Developer's Studio können Grafik-Objekte für die Gestaltung von Oberflächen verwendet werden. Der Anwender kann weitere Eigenschaften – die sich an die Gegebenheiten, die im Developer's Studio definiert wurden, orientieren – für die Planungsdurchführung in Interactive Planning modellieren. Das System bietet ein Beispielmodell an, das die Vorgehensweise im Planungswerkzeug aufzeigt. Änderungen der modellierten Unternehmensstrukturen erfolgen in TM1. Das Planmodell kann vom Anwender ja nach Benutzerberechtigung an die Planungsoberfläche angepasst werden (vgl. [Dah03, S. 116]).

5.2.6 Planungsinstrumente

Interactive Planning bietet Simulations- und Prognosefunktionalitäten. Die Simulation wird bei der Plandefinition über Varianten angelegt. Die Prognose erfolgt über die TM1.

Der Anwender kann bei der Dateneingabe bestimmen, ob die Daten als Planzahlen anhand von Wachstumsfaktoren fortgeschrieben werden sollen. Das System bietet auch eine Abweichungsanalyse an. Dabei werden besondere Abweichungen farbig gekennzeichnet (vgl. [Dah03, S. 120]).

5.2.7 Präsentation

Ein Excel-Add-In wird genutzt, um die Daten in Berichten aufzubereiten. Die Berichte können auch mit einer Benachrichtigungsfunktion verteilt werden. Innerhalb der Planungsoberfläche können die Daten in HTML-Dokumente aufbereitet und über Web abgerufen und ausgedruckt werden (vgl. [Dah03, S. 121]).

6 Ausblick (M.J.)

Bereits kurz nach der Übernahme von OutlookSoft und Business Objects gab die SAP AG bekannt, dass BW-IP und BPC mittelfristig zu einer einzigen Plattform verschmelzen sollen. Daher werden bereits seit geraumer Zeit keine funktionalen Weiterentwicklungen mehr in BW-IP durchgeführt. BPC hingegen wird in diesem Zuge immer stärker in SAP NetWeaver BI integriert. Die Microsoft-Version von BPC wird es in Zukunft nicht mehr geben. (Vgl. [Fol13, S. 27])

Mit dem sogenannten *Planning Application Kit* (PAK) sind die Vorteile der HANA-Technologie auch für die Planung nutzbar. Sowohl BW-IP- als auch BPC-Nutzer können das PAK nutzen. Allerdings fallen bei BW-IP-Nutzern doppelte Lizenzgebühren an. Zum einen müssen Lizenzkosten für *BW on HANA* entrichtet werden, was bedeutet, dass die Datenbank in NetWeaver durch SAP HANA gestellt wird. Zudem müssen BW-IP-Kunden, die das PAK verwenden möchten, auch die Lizenzgebühren für BPC bezahlen. Für BPC-Nutzer hingegen fallen keine zusätzlichen Lizenzgebühren an. Diese müssen lediglich die Softwarekomponente *HANA BPC* installieren, um das PAK benutzen zu können. (Vgl. [Fol13, S. 27])

Durch die Nutzung von PAK und HANA liegen die Ergebnisse deutlich schneller vor, wodurch auch die Qualität der Planungen und Simulationen steigt. Ein weiterer Vorteil ist, dass die Planungszyklen und somit auch die Administrationskosten reduziert werden können. Zusätzlich können die Nutzer von den funktionalen Weiterentwicklungen profitieren, die in BW-IP nun nicht mehr durchgeführt werden. (Vgl. [Fol13, S. 27])

Literatur

[Brü09] Brück, Uwe: *Praxishandbuch SAP-Controlling: [Lösungswege für Ihre täglichen Controllingfragen ; CO-OM, CO-PC und CO-PA verständlich dargestellt ; mit Kapiteln zu SAP NetWeaver BI und der BI-integrierten Planung].* SAP Press : SAP betriebswirtschaftlich. Galileo Press, Bonn u.a, 3., aktualisierte Auflage, 2009.

[Dah03] Dahnken, Oliver und Keller, Patrick und Narr Jörg und Bange Carsten: *Planung und Budgetierung: 16 Software-Plattformen für den Aufbau unternehmensweiter Planungsapplikationen.* Oxygon-Verl, München, 1. Auflage, 2003.

[Egg07] Egger, Norbert: *SAP Business Intelligence: [kompakter Gesamtüberblick über alle Neuerungen ; umfassende Informationen zu SAP Visual Composer und BI Accelerator ; aktuell zu Release SAP NetWeaver 2004s BI].* SAP PRESS. Galileo Press, Bonn, 2007. 1., korr. Nachdr. 2007.

[Ehr95] Ehrmann, Harald: *Unternehmensplanung.* Kiehl, Herne, 1995.

[FGSK08] Frick, Detlev, Andreas Gadatsch und UteG. Schäffer-Külz: *Überblick über die SAP-Software-Komponenten.* In: *Grundkurs SAP ERP*, Seiten 5–22. Vieweg+Teubner, 2008.

[Fis05] Fischer, Roland: *Unternehmensplanung mit SAP SEM/SAP BW: Operative und strategische Planung mit SEM-/BW-BPS ; [integrierte Unternehmensplanung: Theorie und Praxis, projektorientierte Anwendung von SAP SEM/SAP BW].* SAP PRESS. Galileo Press, Bonn, 2., aktualisierte Auflage, 2005.

[Fol13] Folberth, Christian: *SAP-Planungswerkzeuge: Welches ist das richtige? BW-IP und BPC von SAP.* Computerwoche, 42:26–27, 2013.

[Fuc] Fuchs, Christian: *Der Markt für Planungs- und Controllingssysteme.*

[Han] Hansl, Rüdiger: *Unternehmensplanung und Prognosetechniken.*

[Kie10] Kiesswetter, Martin und Gulliš, Goran und Vahlkamp Dirk und Arrenbrecht Alex: *Praxishandbuch Unternehmensplanung mit SAP: SAP BusinessObjects Planning and Consolidation.* SAP PRESS. Galileo Press, Bonn, 1. Auflage, 2010.

[Laca] Lackes, Richard: *Gabler Wirtschaftslexikon, Stichwort: Online Analytical Processing (OLAP).*

[Lacb] Lackes, Richard: *Gabler Wirtschaftslexikon, Stichwort: Online Transaction Processing (OLTP).*

[Mer05] Mertens, Peter et al.: *Grundzüge der Wirtschaftsinformatik.* Springer, 9 Auflage, 2005.

[MuC09] Cissek, Peter und Rautenstrauch, Claus Marx Gómez, Jorge Carlos und: *Einführung in Business Intelligence mit SAP NetWeaver 7.0.* Springer Berlin Heidelberg, Berlin und Heidelberg, 2009.

[SAPa] *SAP Business Information Warehouse.*

[SAPb] SAP: *1992-2001: the SAP R/3 era.*

[SAPc] SAP: *2002-present: real-time data where and when you need it.*

[sap06] *SAP - Einführung in MM - Dienstleistung*, 25.04.2006.

[Sch00] Scholz, Gunnar und Svoboda, Michael: *Integrierte Unternehmensplanung mit SAP R/3.* Controlling und Management, 44(6):371–378, 2000.

[Vah] Vahrenkamp, Richard: *Gabler Wirtschaftslexikon, Stichwort: Enterprise-Resource-Planning-System.*

[Wel07] Welge, Eulerich: *Szenario-Technik.* Controlling, 2, 2007.

Abschließende Erklärung

Wir versichern hiermit, dass wir unsere Seminararbeit selbständig und ohne fremde Hilfe angefertigt haben, und dass wir alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen unserer Arbeit besonders gekennzeichnet und die Quellen zitiert haben.

Oldenburg, den 1. August 2014 Abdulmasih Hadaya, Mariska Janz

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Design Thinking

Seminararbeit
im Rahmen der Projektgruppe: inMemory Planung mit SAP HANA

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Nils Giesen

Vorgelegt von:	Ivaylo Ivanov Harlingerstr.13a 26121 Oldenburg Telefonnummer: 017631357464

Abgabetermin:	02. August 2014

Inhaltsverzeichnis

Abbildungsverzeichnis **4**

1. Einleitung **5**
1.1. Problemstellung 5
1.2. Zielsetzung der Seminararbeit 7

2. Design Thinking **7**
2.1. Die innovationsmethode Design Thinking 7
2.1.1. Vorgeschichte 7
2.1.2. Kernelemente 8
2.1.3. Der Prozess Designn Thinking 10
2.1.4. Management und Design Thinking 14
2.1.5. Design Thinking in der Organisation verankern - „The Transformation Pyramid“. 15
2.1.6. Design Thinking vs. Hybrid Thinking 16

3. Anwendung von Design Thinking anhand vom Bespiel von SAP **17**

4. Fazit und Ausblick **20**

A. Anhang **20**

Literaturverzeichnis **21**

Abbildungsverzeichnis

1. Kernelemente Design-Thinking 8
2. Design Thinking 9
3. Iterative Schleife-DT 11
4. Vergleichbare Variante für Design Thinking 14
5. Hybrid Thnker 17
6. Design Thinking macht SAP Innovator 19

1. Einleitung

Aus heutiger Sicht ist es für viele Unternehmen wichtig zu verstehen, woher gute Ideen kommen, die das menschliche Leben bereichern und erleichtern können - Ideen, die wesentliche Neuerungen mit sich bringen und unsere Welt komplett verändern können. Viele sind Teil unseres Alltags und werden als selbstverständlich aufgefasst.

Dabei steht jedoch fest, das alle Ideen von Menschen entwickelt und umgesetzt werden. Das bedeutet, dass jemand entschieden hat, Produkte oder Services aus diesen Ideen entstehen zu lassen, die so funktionieren und sich anfühlen, wie sie es aktuell tun. Sie sind also das Endprodukt von Designprozessen. „Design“ verstehen wir hier aber nicht im Sinne eines Verschönerungsprozesses der Sache, vielmehr bezeichnet der Begriff die Suche nach kreativen Lösungen für komplexe Probleme. Die innovativen Problemlösungen sind dabei selten ein Produkt des Zufalls oder der isolierten Arbeit einzelner Personen, die meisten Erfinder sind Teil einer Gruppe, in dessen Rahmen sich alle Mitglieder gegenseitig inspirieren. Als besonderes Beispiel dafür ist Thomas A. Edisson zu nennen, ein Erfinder, der mit über 1000 Patenten in der Geschichte verewigt ist. Dieses Erfolg ist einer strukturierten Herangehensweise zu verdanken, in der alle Teilnehmer des innovativen Prozesses die Herausforderungen gemeinsam lösen. Es wurde also das kreative Wertschöpfungspotezial durch eine Arbeitsmethode kanalisiert und gesteuert, was wiederum ein Beweis zur bewussten Planung und Förderung von Innovationen ist [GM13].

Design Thinking ist eine neuartige Methode zur Entwicklung innovativer Ideen ist umsetzbar in allen Lebensbereichen. Das Konzept basiert auf der Überzeugung, dass Innovation nur dann möglich sind, wenn starke multidisziplinäre Gruppen kollaborativ arbeiten, eine gemeinschaftliche Kultur bilden und die Schnittstellen der unterschiedlichen Meinungen und Perspektiven erforschen [Ins14b].Bei der Entwicklung der Konzepte, die auch mehrfach geprüft werden, stehen im Mittelpunkt die Bedürfnisse und Motivationen von den Menschen[Bro06].

1.1. Problemstellung

Unsere Welt ist schon längst von Innovationen und Technologieentwicklung erobert. Begriffe wie „Technologischer Wandel“, Hightech-Unternehmen“, „Ubiquitous Computing“, “Digitale Disruption“, „App-Entwicklung“ usw. sind auch Teil dieser komplexen und dynamischen Welt. Hinsichtlich ihres Sinngehalts sind alle Technologien grundverschieden und sehr vielfältig, dennoch haben sie etwas gemeinsam. Es wird immer wieder neue Herausforderungen geben, zu denen dementsprechend die passende Lösung gefunden werden muss. Dies fordert stetig neue Konzepte zur Umsetzung.Die unzählige technische Veräde-

rungen und Neuerungen, paralell mit den ständig wechselnden Ramenbedingungen,stellen alle am Markt tätigen Unternehmen unter enormen Druck. Das bedeutet mit einfachen Worten erklärt: Wer erfolgreich sein will und bleiben möchte,muss einfach innovativ sein.

Innovationen werden am meisten in den Unternehmen mit innovativen Produkten assoziiert und die Produktrevolution beginnt in den Prozessen. Neue Geschäftsmodelle müssen entwickelt werden, die die ausgetretenen Prozessfäden verlassen und zur Verankerung der Digitale-Disruption in den Anwenderunternehmen führen. Aktuell ist die Anpassung an der Digitale-Disruption nicht mehr ausreichend, sondern viel mehr ihre Einführung bei den Endkunden.Und das alles ist verbunden mit der Frage wie man das Unternehmen revolutionieren kann[Com13]

Mit anderen Worten formuliert, die fortgeschrittene Digitalisierung bringt das Ende der Geschäftswelt wie man sie kennt. Um weiterhin erfolgreich zu bleiben ist die Vernetzung der 3 Cs: Communication, Community und Commerce wichtig. Das ist nur mit einem kollaborativen Einsatz von internen und externen Experten, Fans und Kritikern in den Communities sowie des digitalen Handels zu erreichen. Bei reibungsloser Zusammenarbeit dieser drei Bereiche finden nicht nur Kunden schneller ihren Weg durch den Verkaufstrichter, die interne Zusammenarbeit wird zudem produktiver und innovativer[Wei13].

Die neuen Technologien bringen neue Herausforderungen.Vertrackte, schwer zu fassende Probleme, die mit so vielen und tiefgreifenden Dilemmata behaftet sind sind, dass wir uns keine befriedigende Lösung für alle Beteiligten vorstellen können.Im Kontext des Design Thinkings sind dise als “Wicked Problems“bekannt. Das Managemet ist auch vor solchen Probleme machtlos, obwohl es von vielen Menschen als die wichtigste Erfindung der Wirtschaftsgeschichte bezeichnet ist. Denn es beschreibt die Fähigkeit, andere Menschen in einem Wertschöpfungsprozess zu organisieren und anzuleiten. Die zurzeit gängigen Management-und Führungsmethoden sind dagegen teilweise älter als 70 Jahre. Sie fokussieren hauptsächlich den Einfluss der extrinsischen Motivation und sind nicht mehr fähig, bahnbrechende Innovationen hervorzurufen[ER13].

Design Thinking als Form des erfinderischen Denkens mit radikaler Kunden- beziehungsweise Nutzungsorientierung kann mehr neues, schneller in die Welt bringen. Dabei bietet es keine Garantie für eine disruptive Erneuerung, aber in Vergleich mit allen gängigen Innovations- und Managementsmethoden, liefert es mit signifikant höhere Wahrscheinlichkeit nutzerorientierte Problemlösungen[ER13].

Die zahlreichen Definitionen von Design Thinking sind durch die unterschiedlichen Wurzeln des Begriffes, sowie den kontinuierlichen Verbesserung und Anpassung der Methode in verschiedenen Kulturen und Bereichen geprägt. Daher ist der Begriff nicht immer leicht zu verstehen. Des Weiteren fehlen Konzepte, sowie empirische Studien, die uns Erkenntnis-

se und Erfahrungen zum Integrationsprozess von Design Thinking in unternehmerischen Strukturen liefern können. Das gleiche betrifft die Prozesse zur Generalisierung von Innovationen.

1.2. Zielsetzung der Seminararbeit

Hauptziel dieser Seminararbeit ist an erster Stelle zu erklären was sich hinter dem Begriff „Design Thinking“ verbirgt. Des Weiteren sollen die wesentlichen Merkmale und Anforderungen von Design Thinking herausgearbeitet und deren Bedeutung für das Entstehen von sowohl kreativen als auch innovativen Ideen im Rahmen eines Teams erarbeitet werden. Anhand eines Beispiels der SAP AG werden dabei Erfahrungen von Experten/-innen im praktischen Umgang mit Design Thinking in Großunternehmen erhoben und dargelegt.

2. Design Thinking

2.1. Die innovationsmethode Design Thinking

2.1.1. Vorgeschichte

In der Vergangenheit basierten alle Problemlösungskonzepte auf der tayloristischen Arbeitsteilung. Um eine Lösung leichter zu entwickeln, wurden aus den großen und komplexen Problemen, kleinere und überschaubare Aufgabenstellungen abgeleitet. Als nächstes setzte man die Teillösungen zusammen, um eine Gesamtlösung zu erreichen. Diese Vorgehensweise hat sich nicht immer als effektiv erwiesen. Dabei hat man entdeckt, dass die genauere Betrachtung des Problems sowie deren Überprüfung von essentieller Bedeutung ist. Damit ist der erste Schritt zur Methode des Design Thinking gemacht. Erstmalig taucht der aktuelle Begriff 1991 bei einem Symposium an der Technischen Universität Delft auf[Sch10b].

David Kelly, der auch als Mitgründer der Design-Agentur IDEO und als Professor an der Standford Universität bekannt ist, entwickelt Design Thinking als innovativen Ansatz. Außerdem sind die Professoren Terry Winograd und Larry Leifer von der amerikanischen d.school bei der Weiterentwicklung des Ansatzes aktiv beteiligt. Im Jahr 2005 wurde auch Hasso Plattner von IDEO und deren nutzerzentrierter Arbeitsweise zur Entwicklung von Produkten mitgerissen. Der damalige SAP-Vorsitzende ist voll überzeugt, dass die Softwareprodukte der Zukunft nach der Methode des Design Thinkings entwickelt werden müssen. Seitdem er in die sogenannte d.school an der Standford University investierte, trägt die Schule dort seinen Namen: „Hasso Plattner Institute of Design at Stanford“. Mit der Eröffnung der HPI School of Design Thinking in Postdam im Jahr 2007 erechte das

Konzept auch Deutschland. Die Universität in Postdam bietet aktuell Platz für ca. 120 Studierende und kooperiert eng mit der Standford University[Pfe12].

2.1.2. Kernelemente

Der Begriff des Design Thinkings kann auf verschiedene Weise interpretiert werden. Viele innovative und kreative Aktivitäten sind damit eng verbunden. Die verschiedenen Komponenten und Methoden des Konzepts existieren seit Generationen. Der Wert des Design Thinking liegt in ihrer Kombination. Sie schafft einen Prozess, indem es einen Prozess als eine neue, verallgemeinerte und zugänglichere Denkweise schafft [MBN12].

Die Kernelemente von Design Thinking werden von der HPI Universität Postdam folgendermaßen definiert: Prozess, Raum und Mutlidisziplinarität bezeichnet.

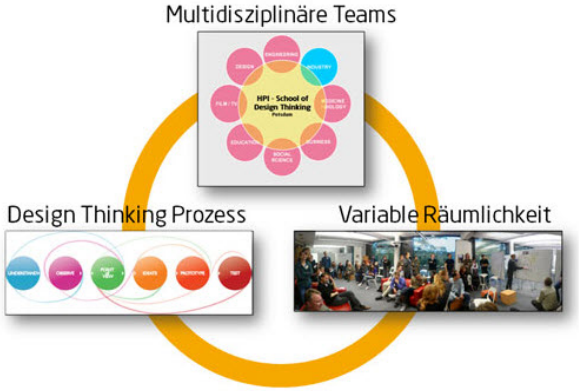


Abbildung 1: Kernelemente Design-Thinking

Im Mittelpunkt des Prozesses steht der Mensch. Daher ist der Prozess ein menschenzentrierter Ansatz, der Methoden und Werkzeuge aus verschiedenen Bereichen wie Design und Ethnographie, dank Technologien und Wirtschaft miteinander verknüpft. Siehe Abbildung 2 [Ins14a] Durch diesen iterativen Prozess werden Nutzerbedürfnisse aufgedeckt und geprüft, ob sie technisch machbar und wirtschaftlich rentabel sind. Vom Endergebnis dieser Arbeitsweise können alle Steakholder profitieren, indem neue Erfahrungen gewonnen in Form von Produkten, Dienstleistungen, Prozesse, Veranstaltungen und sogar Regelwerke in den Wertschöpfungsprozess des Unternehmens eingehen.Die Arbeitsräume, die von der Kultur und Persönlichkeit der Mitarbeiter geprägt sind, müssen passend ausgestattet sein. Im Mittelpunkt stehen Mobilität und Flexibilität, die durch die freie Bewegung der Gegenstände im Raum gesichert sind. Zum Beispiel sollten alle Möbel auf Rollen befes-

tigt sein. Alle Elemente des Raumes inklusiver Wände und Oberflächen sollten für Design Thinking benutzt werden können. Nur starke multidisziplinäre Gruppen, die unterschiedliche Perspektive erforschen, sind durch den Aufbau einer kollaborativen Kultur fähig, neuen Innovation zu kreieren. Viele Prozessteilnehmer sind der Meinung die Teamarbeit als Methode zu kennen und doch wenige sind am Ende mit Erfolg gekrönt. Deswegen ist es besonderes wichtig, gemeinschaftliche Prinzipien in die Gruppe zu etablieren als Grundlage für lange und fruchtbare Arbeit. Dabei ist Design Thinking das Bindeglied zwischen den verschiedenen Disziplinen und führt zum Projekterfolg [Ins14a]. Im Idealfall haben alle Teammitglieder ein „T-shaped-Profil“Ḋas sind Menschen deren Kenntnisstand mit dem Buchstabe „T “ virtualisiert werden kann. Dabei entspricht die fundierte Expertise in einem bestimmten Gebiet des vertikalen Teil des Buchstaben. Der horizontale Teil des „T “ symbolisiert zahlreiche komplementäre Fähigkeiten und Interessen sowie ein breites Allgemeinwissen.[Ins13]

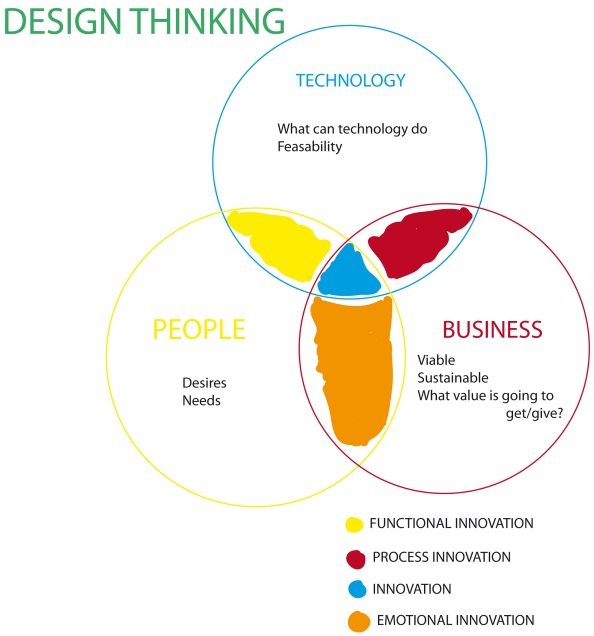


Abbildung 2: Design Thinking

2.1.3. Der Prozess Designn Thinking

Die Zusammensetzung von verknüpften Aufgaben und deren Abhängigkeiten bilden Prozesse. Diese sind Teil unserer Umwelt. Unterschiedliche Aktivitäten vom normalen Alltag

wie zum Beispiel Lebensmitteleinkaufen, ein Spiel spielen oder oder das Binden von Schuhen stellen Prozesse bzw. Prozessabläufe dar. Dennoch lässt sich erst bei einer tieferen Betrachtung der Einzelschritte erkennen, wie sie aufeinander abgestimmt zum Gesamtziel führen. Prozesse sind nicht unbedingt explizit gemacht, aber es gibt sie implizit in Sachen, die schon da sind. Unser Prozess-Bewusstsein ist ein Weg die Welt zu sehen und wenn wir diese Ansicht akzeptieren, dann sind Prozesse allgegenwärtig[Lue11]. Siehe Abbildung **??**[1]

„You cannot hold a design in your hand. It is not a thing. It is a process. A system. A way of thinking“

[Gil]

Design ist ein sich wiederholender Prozess und Design Thinking begleitet jede Phase der Reise, von den Kundenbedürfnissen hinaus bis zum Endprodukt. Dabei kann eine konkrete Aufgabe mit verschiedenen Lösungen verbunden werden, deren Anforderungen an Kreativität, Funktionalität oder Budget variantenreich sein können[2], S5.

Trotz der kreativen Offenheit, die beim Design Thinking folgen soll, enthält die Methode klar umrissene Schritte, die im Idealfall aufeinander folgen. Verstehen, Beobachten, Synthese, Ideengenerierung, Prototyping und Testen[GMP09].

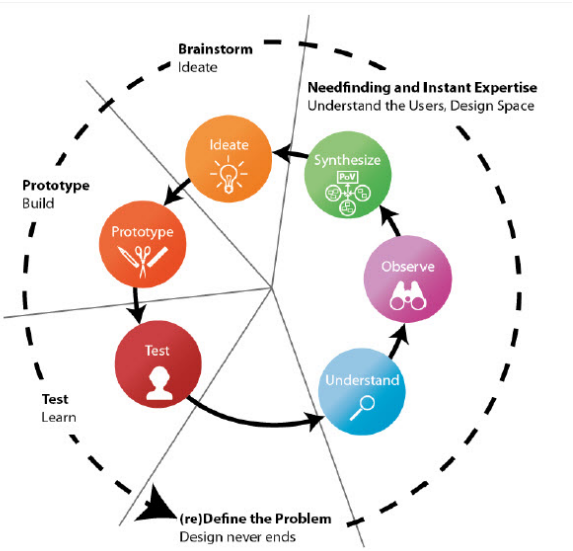


Abbildung 3: Iterative Schleife-DT

[1] Vgl.
[2] AmbroseH

- Verstehen: Zu Beginn geht es darum, die Problemstellung und das damit verbundene Problemfeld zu verstehen. Das zu lösenden Problem muss quasi von allen Perspektiven betrachtet werden. In einem Geschäftsfall sind das Beispielsweise die Persepektiven des Kunden oder des Marktes. Die mit der Problemstellung verbundene Technologie ist hierbei auch zu berücksichtigen, da sie verschiedene Einschränkungen oder Möglichkeiten mit sich bringt.[KL07], S6. Diese Informationen haben verschiedene Quellen und umfassen eine sehr breite Informationsfläche. Die Informationserhebung kann dabei nicht nur viel Zeit in Anspruch nehmen, sondern seine Zielgenauigkeit verlieren. Um ein klares Bild zu bekommen, soll die dazugehörige Recherche sorgfältig geplant werden. Nach der Recherchearbeit kann es sein, dass die Problemstellung neu formuliert, verschoben oder eingegrenzt werden muss. Zum Schluss müssen alle Teilnehmer ein gemeinsames Verständnis von der Problemstellung und Expertenkenntnisse haben.[PMW09].

- Beobachten: Design Thinking ist ein menschenzentrierter Ansatz und als solcher stützt sich überwiegend auf qualitative Forschungsmethoden[Nor13]. Die Methode wendet verschiedene Wissensquellen an. Darunter zählen nicht nur die Zielgruppe, sondern auch die Lead User, Non User sowie auch Situationen. Zusammenfassend formuliert, kann man sagen, dass alle Elemente untersucht werden, die Analogien zum Betrachtungsgegenstand aufweisen. Sehr oft ist die Beobachtung als Teil des Prozesses mit darauf folgenden Interviews verbunden. So wird die Überbrückung zur Realität, die es ermöglicht, menschliche Vorlieben, Abneigungen und latente Bedürfnisse besser wahrzunehmen und zu verstehen[PMW09].

- Synthese: Das Team studiert alle gesammelten Informationen und Erkenntnisse. Wichtig ist hierbei, nicht alle Einzelheiten zu wiederholen, sondern die Erkenntnisse aus bestimmten Situationen und unerwartete Wendungen darzulegen. Auf Post-its werden alle Informationen vermerkt und an die Wände fixiert, wobei man sie zu den relevanten Themen oder sogenannten Cluster zuordnen soll. Dadurch wird alle Information, über die wir verfügen, virtualisieren. Der damit geschaffte Gesamtüberblick macht es möglich alles als ein Ganzes zu betrachten und dabei aus Sicht der Problemstellung, das Wichtige zu selektieren. Passend dafür sind verschiedene Frameworks wie Zwiebeldiagramme, Relationship-Maps, User-Journeys oder Mindmaps. Endergebnis des Vorgehens ist eine Figur der idealtypischen, fiktiven Person. Diese verkörpert unserer Nutzer, samt alle relevanten Eigenschaften und Sichten. Der individuelle Standpunkt dieser Persona ist auf eine Lösung des Problems ausgerichtet, in Form einer Metapher, eines prägnanten Satzes[GMP09].

- Ideengenerierung: Ausgehend von dem Standpunkt der Person sind oftmals alternative Lösungen eines Problems vorhanden. Dieser Lösungsspielraum kann kann auch als Grundlage für Brainstorming-Fragen dienen, die weiterhin als Fundament für die Ideengenerierung fungieren. Die Qualität der zukünftigen Ideen ist abhängig von den Fragen aus denen die entstehen. Voraussetzung für die Durchführung von Brainstorming sind bestimmte Regeln, derer Einhaltung wichtig ist[KL07] S56. Eine davon ist die Abkehr von Kritik, soweit es möglich ist, weil nur so sich kritische Ideenvorschläge weiter entfalten können. Zielführend beim Brainstorming ist die Quantität statt Qualität. Hier ist gewollt, dass alle Teammitglieder sich von ihren Kollegen inspirieren lassen. Hierzu kann man auch schon fertige Ideen weiter modellieren und verbessern. Um transparent zu arbeiten, sollen die Ideen außerdem durch Skizzen virtualisiert werden. So sind alle Ideenvorschläge leichter zu übertragen und präsentieren als Gesamtbild. Chaotische Leitgedanken müssen auch in Betracht gezogen werden. Oftmals Ideen, die zu Beginn als sinnlos oder nicht relevant erscheinen, erweisen sich später als großartige Lösungen. Was noch zu beachten ist, die Entfernung zur Leitfrage. Das Thema darf nicht aus der Sicht verloren werden. Dafür wird am Anfang des Brainstormings, die Leitfrage mit großen Buchstaben über die Arbeitsfläche geschrieben. Wenn Ideen zur Verfügung stehen, werden diese in sortierten Gruppen zusammengestellt. Daraufhin werden die besten aussortiert. Das kann durch Vergabe von Punkten gemacht werden, indem jeder Mitglieder bestimmte Anzahl von Punkte bekommt, mit denen er später seine Favoriten bewerten kann[GMP09].

- Prototyping: Dieses Teil vom Design Thinking konzentriert sich auf schnelles und iteratives Prototyping. Die zahlreichen möglichen Formen variieren hier von sehr rudimentären, über Papier- und Pappmodellen, Rollenspiele bis hin zu voll funktionsfähige Ausarbeitung, die aber schon mehrere Testläufe fordert. Ziel der Prototypen ist eine Antwort für noch offene Fragen zu finden und die parallele Weiterentwicklung der Idee. Mögliche Fragen wären hier zum Beispiel: „Worauf muss sich die Idee konzentrieren um diese am Klarsten darzustellen? Sind mehrere Ideen in einem Ideenkonzept verbunden und muss jede einzeln als Prototyp dargestellt werden? Wie kann die Idee in eine angemessene Form gebracht werden, um sie zu kommunizieren und damit mehr über die Idee selbst zu lernen?“. Dabei ausschlaggebend ist, die erstellten Prototypen vor allem als weitere Ideengeber zu sehen und erst dann als Mittel zur Validierung von Ideen. Eine Idee, die einfach in Worte oder Schrift erfasset ist, kann uns nicht den gleichen Wert für liefern. Die Form der Idee ist das Element, das die Beziehung zu anderen Ideen und Modifikationen aus dem gleichen

Wertschöpfungsdomäne schafft. Jeder weitere Testablauf verstärkt die Aussagekraft des Prototyps. Bei einer intensiven Wiederholungssequenz, wird schnell deutlich, in welchen Varianten Erfolgspotential steckt. Die Iterationsschritte erlauben eine weitere Verbesserung und Verfeinerung der Idee, die Ausdruck in der gemeinsamen Formfindung des Teams finden[GMP09].

- Testen: Test und Feedbacks-Schleifen finden nach dem Prototyping statt. Für sie muss eine konkrete Form vorhanden sein. Die Menschen bei der Befragung haben es leichter etwas zu präzisieren und Alternativen zu finden sowie Vorschläge zu machen, wenn sie einfach über etwas Konkretes sprechen. Zu beachten ist hierbei die Verhaltensweise der Teilnehmer der Befragung, die ein Indiz dafür ist, ob etwas verfolgt werden soll oder nicht. Design Thinking ist ein Human-Centered-Ansatz und daher ist bei den Tests oder Feedbacks, die Aufnahme menschlichen Wissens, Erfahrungen und Intuition wichtig, die später als Fundament für die Inspiration neuer Ideen dienen[GMP09].

Es gibt auch eine vergleichbare Variante für Design Thinking, die einen Zyklus mit fünf Phasen beschreibt: Empaty/Emphatize die „Verstehen“ und „Beobachten“ zusammenfasst, Define enspricht Synthese, Ideate, Prototype und Test.

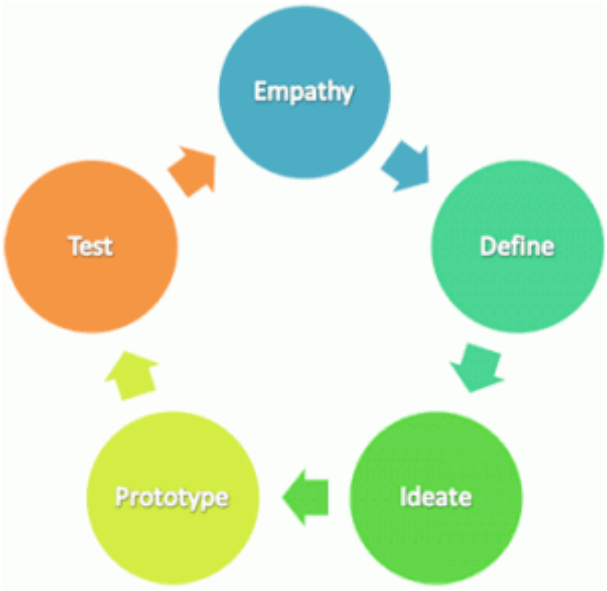


Abbildung 4: Vergleichbare Variante für Design Thinking

Was die Prozessmodelle voneinander unterscheidet, ist eigentlich die geringfügig veränderte Benennung der Phasen. „Beobachten“ und „Verstehen“ werden mit „Emphaty“ bzw. „Emphatize“ ersetzt. Ähnlich verhält es sich bei IDEO Design Thinking, welcher aus drei Phasen besteht: Inspiration-Ideation-Implementation. Trotz verschiedener Prozessdarstellungen, ergibt sich bei genauerer Betrachtung, eine weitgehende und leicht erkenntliche Übereistimmung der Vorgehensweise gegenüber dem hier beschriebenen Prozessmodell für Design Thinking. Eine Lösung wird durch die einzelnen Prozessschritte ausgearbeitet, es ist auch möglich, diese beim Bedarf nichtlinear zu durchlaufen. Im Prinzip, kann man die Reihenfolge der Prozessphasen tauschen und beliebig ausführen. Design Thinking bemüht sich immer zwischen Beobachten, Interpretieren, Aufstellen von Hypothesen sowie Ausprobieren zu agieren und dadurch eine Lösung zu erreichen[Kre14].

2.1.4. Management und Design Thinking

Nach der Geburt des Konzepts in den 60er und 70er, wurde das Thema Design Thinking lange Zeit nicht direkt angesprochen. Erst mit dem Beginn des neuen Jahrhunderts, gewinnt der Begriff wieder an Popularität. Grund dafür ist die dynamische Entwicklung der Wirtschaft. Wirtschaftskrisen, gesättigte Märkte und der kontinuierlich steigende Konkurrenzdruck fordern ständig innovative Lösungen[Bro09]. Das Managementsystem kann aber keine neuen Lösungen anbieten die der neuen Situation entsprechen. Das Management ist eine der wichtigsten Erfindungen in der Geschichte der Wirtschaft und diente lange als Instrument, Menschen innerhalb eines Wertschöpfungsprozesses zu organisieren und anzuleiten. Das Problem dabei ist, dass die verbreitete Management- und Führungsmethoden fast 70 Jahre alt sind. Diese entstanden unter anderen Wirtschaftsbedingungen - starre, hierarchische Strukturen und Belohnungssysteme, die die extrinsische Motivation der Arbeitnehmer lenken sollten, waren die Hauptelemente dieser Systeme[ER13], S15. In der industriell orientierten Ökonomie von damals, hat das Management Organisationen auf Reproduktion programmiert. Diese Methode ist erfolgreich gewesen, weil das wirtschaftliche System relativ berechenbar war. Diese Strategie ist aber nicht für komplexe Problemstellungen vorgesehen. Heutzutage existieren Herausforderungen mit einem unbekannten Grad der Komplexität, sie enthalten oftmals ein Dilemma und sind als „Wicked“-Problemen gekennzeichnet. In der modernen Wirtschaft, sind die Routinen auf den Ebenen der Prozessabläufe, der Organisationsstruktur und der Personalführung, zusammen mit der Größe des Unternehmens, keine Garantie mehr gegen Marktversagen. „Too big, to fail“ gehört der Vergangenheit an, ein gutes Beispiel bietet der Konzern „Nokia“. Heutzutage kann man dank der Technologieentwicklung, die Welt als wirklich flach betrachten, umfassende

ökonomische Paradigmenwechsel und allgegenwärtige Dilemmata zeichnen diese Welt aus, in der ein Hochgeschwindigkeitswettbewerb in kleinen Marktsegmenten herrscht. Das ist eine Weltwirtschaft, in der die Erfolgsmethoden von gestern nicht umgesetzt können und die richtigen Dinge neu erfunden werden müssen[ER13], S172. Design Thinking hat die Kraft, gedankliche Hürden zu überwinden und Dogmen radikal in Frage zu stellen. Diese Methode ist eine neue Richtung der Management-Entwicklung, die menschliche Bedürfnisse und Effizienzsteigerung in den Mittelpunkt stellt[ER13].

2.1.5. Design Thinking in der Organisation verankern - „The Transformation Pyramid“.

Um die maximale Wirkung in einer Organisation zu entfalten, hat das Design Thinking mit den Hürden des alten Managementsystems zu kämpfen. Erst wenn das Management involviert ist, kann das Konzept etwas bewirken. Indem Führungskräfte, Mitarbeiter auf die Wertebene anheben, wird die intrinsische Motivation gestärkt. Das schafft optimale Voraussetzungen für innovative Prozesse. Zudem soll das Unternehmen sein „meaning“ finden. Was grob übersetzt im diesen Kontext als „Sinn und Bedeutung“ zu verstehen ist. Nur solche Organisationen können selbst die richtige Organisationsstruktur finden und innovativ agieren. Das „meaning“ kann auch ein tief im Menschen verwurzeltes, ungestilltes Bedürfnis befriedigen. Wenn dabei auch eine von den durchgesetzten Grundegeln, die alle Markteinehmern berücksichtigen, gebrochen wird, hat man den Weg zu disruptiver Innovation gefunden. Dadurch wird das Geschäftsmodell einzigartig und schwer nachzuahmen[Kob14].

Jürgen Erbeldinger beschreibt in seinem Buch „Durch die Decke denken - Design Thinking in der Praxis“ wie man anhand von 7 Schritten, Design Thinking in der Organisation verankern kann. Diese aufeinander ausgeführt, sind vom Autor als „The Transformation Pyramid “ dargestellt. Der erste Schritt ist „Dont talk! Do“. Dabei werden Tische, Wände und Material in Form eines Raumes zur Verfügung gestellt. Dann kann man mit einem Miniprojekt wie zum Beispiel eines Reports, einer Präsentation oder eines Angebots anfangen, indem man versucht diese mit Design Thinking zu verbessern. „Your Are Not Alone “ ist der zweite Schritt - Hier soll ein Kernteam mit 5 bis 10 Personen gebildet werden, von denen jeder ein Projekt betreut. Im Idealfall verfügt man hier über 5 Räume für Design Thinking. „Change Your Meeting Culture “ ist der dritte Schritt. Die Meetings sollen in den Design-Thinking-Räumen stattfinden. Außerdem, wenn es sinnvoll ist, kann man die Design-Thinking-Techniken einsetzen. Beim nächsten Schritt „Show the Power“ sind kritische Themen anzusprechen. Die Hauptakteure hier sind Vorstände, Bereichslei-

ter und Vorbilder. Zum diesen Zeitpunkt sollte man mit der Ausbildung der nächsten 25-50 Design-Thinking-Moderatoren anfangen. „Dig the Dogma and Invert It“ ist Schritt Nummer fünf. Hier erfolgt Überprüfung der Positionierung des Unternehmens auf dem Markt, mittels Customer Journey und Value Chain. Die Thinker sind auf das jeweilige Geschäftsmodell sowie die Marke fokussiert. Der sechste Schritt nennt sich „Innovate Management Itself “ . Design Thinkig hat hier das Management schon involviert. „Whats is your companys meaning“ ist der letzte Schritt und damit die Spitze der Pyramide[ER13].

2.1.6. Design Thinking vs. Hybrid Thinking

Wenn über Design Thinking gesprochen wird, darf das Thema Hybrid Thinking nicht ignoriert werden. Die Hybrid Thinking-Methode ähnlich wie das Design Thinking hilft dem Unternehmen sich innovativ zu entwickeln. Hinter Hybrid Thinking steht der Name von Dev Patnaik. Er ist Gründer und Direktor von Jump Associates. Die Firma ist in San Mateo, California und New York als ein Beratungsunternehmen bekannt. Zu ihren Kunden zählen Top-Unternehmen wie Nike, Fedex, Samsung and Hewlett-Packard. Was ihre Arbeit wesentlich von allen Konkurrenten unterscheidet, ist eine Kombination von McKinsey & Co Strategie-Beratung verbunden mit kreativer Arbeit wie bei Frog oder Ideo. Laut Patnaik sehen alle Unternehmensherausforderungen anders aus, sind aber eigentlich gleich und können mit einem Ziel erreicht werden.Er denkt, dass Grund dafür ist die Innovation.Sie entsteht aber nicht wie die Menschen annehmen, mit den neuen Produkten und Dienstleistungen. Die Entfesselung der Kreativität der Arbeitskraft,ist der Schlüssel zum Erfolg laut Patniak. Um dies zu schaffen, soll man nicht extra dafür intelligentere Leute einstellen. Potnaiks-Erfolgsrezept ist das Talent, von den bereits angestellten Mitarbeitern, zu inspirieren und zu entfesseln. Hybrid Thinker gehen davon aus, dass jedes Unternehmen große Ziele von Natur aus erreichen will. Dafür soll man alte und neue Ideen in der Verfolgung des Wachstums zusammen mischen. Dabei profitiert das Unternehmen von drei Vorteilen: die Angestellten werden erstmal motivierter oder härter arbeiten, dann werden sie auch innovative Pläne schaffen und diese Pläne werden letztendlich dem Unternehmen helfen zu wachsen[Upb11].

Weil beide Methoden das gleiche Ziel verfolgen ist es unvermeidlich, dass auch eine Art Konkurrenz dazwischen entstanden ist. Bei der Ausarbeitung dieses Themas bin ich zum Rückschluss gekommen,dass die Methoden, in den wesentlichen Deteils sehr ahnlich sind. Die Befürworter beider innovativen Konzepte haben auch Meinungsdifferenzen. Design Thinker sind überzeugt, dass die konkurrierende Methode, die Grundidee des Design Thinking enthält[Sch10a]. Hybrid Thinker dagegen sind der Meinung, dass ihre Metho-

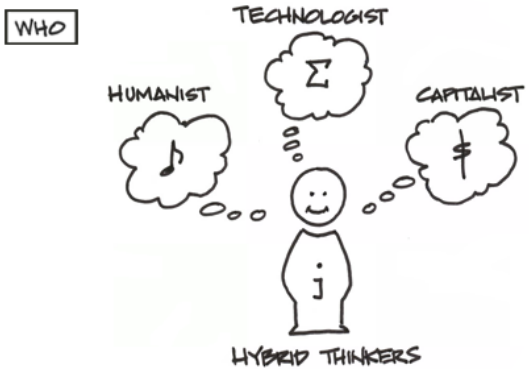


Abbildung 5: Hybrid Thnker

de im Vergleich zu Design Thinking zusätzliche und andere Denkweisen anbietet[Bra10]. Guenther[Gue12], beschreibt in seinem Buch, Design Thinking als Kern von Hybrid Thinking.

3. Anwendung von Design Thinking anhand vom Bespiel von SAP

Laut der Internet-Seite des Unternehmens ist SAP, mit Stammsitz in Walldorf, führender Anbieter von Unternehmenssoftware. Seit seiner Gründung im Jahr 1972 hat sich SAP durch Innovationen und Wachstum zum führenden Anbieter von Unternehmenssoftware entwickelt. Mehr als 253.500 Kunden weltweit sind dank der Anwendungen und Services von SAP in der Lage, rentabel zu wirtschaften, sich ständig neuen Anforderungen anzupassen und nachhaltig zu wachsen[SAP].

Das Walldorfer-Unternehmen hat auch die Design Thinking–Methode für sich entdeckt und nutzt sie als Lösung, um geschäftliche Herausforderungen zu bewältigen. Dadurch wurden Innovationen und Unternehmenserfolg miteinander verknüpft. SAP findet den Ansatz besonderes gut geeignet für Fragestellungen aus ihrer Roadmap[Kul13]. Unter anderem auch um Mobile Apps kreativ zu entwickeln[SAP13]. Für Personen, die den Design Thinking Ansatz in Ihrem Unternehmen anwenden möchten, bietet SAP zusätzlich verschiedene Workshops[SAP14].

SAP erklärt, dass bei der Arbeit mit Design Thinking, die schnelle Entwicklung und der Bedienkomfort vorteilhaft sind. Zudem sind die Anforderungen und Wünsche den Anwender immer im Blick zu behalten. Teams, die aus Entwicklern bestehen, können innovative

Lösungen in kürzen Zeiträumen, durch Zugriff auf breit gefächertes Fachwissen, entwickeln. Dafür müssen sie zunächst die gemeinsamen Ziele unter Berücksichtigung der Unternehmensstrategie identifizieren. Wenn dies abgeschlossen ist, konzentrieren sich die beteiligten Akteure auf die Definition des Problems und damit auf die Aufdeckung der unausgesprochenen Bedürfnisse der Zielgruppen. Dabei werden mit Hilfe von Design Thinking als ein wiederholbarer Prozess, verschiedene Einsichten geschafft und neue Lösungen gefunden, von denen nicht nur die Anwender, sondern auch die produzierenden Unternehmen begeistert sind. Eine Win-Win-Situation. Die Kunden bekommen das Produkt, das sie benötigen und die Unternehmen haben eine umsetzbare und dauerhafte Lösung, die zielgenau auf die Kundenwünsche angepasst.

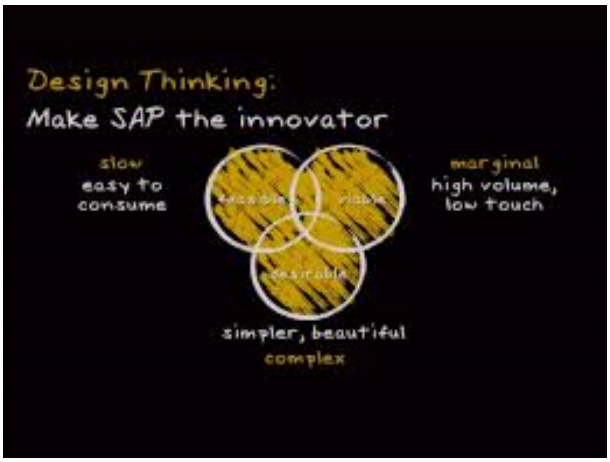


Abbildung 6: Design Thinking macht SAP Innovator

Speziell für Design Thinking vorgesehen, wurde am 14.11.2013 das AppHaus in Heidelberg eröffnet. Die kreative Büroumgebung ist optimal auf Design Thinking angepasst. In diesen Räumen, in denen Software-Designer, Geschäftsexperten und Softaware-Entwickler, kooperativ mit Kunden arbeiten, können bestehenden Produkte ausgebaut oder neue Produktideen ins Leben zu gerufen werden.

Zitat:„Unsere AppHäuser sind ein wichtiger Bestandteil der Innovationskultur von SAP und finden großen Anklang bei unseren Kunden. Sie helfen uns bei der Verjüngung unseres Produktportfolios mit einfach zu bedienenden, innovativen Anwendungen.“... „In dem AppHaus in Heidelberg können bis zu 50 Teilnehmer unter Anleitung von Experten für Design Thinking und Oberflächendesign von Software in Kundenworkshops an neuen, kreativen Ideen arbeiten“[Len13].

Das AppHaus ist abgeleitet vom Konzept des Design-Thinkings, nach dem Leitbild vom „Hasso Plattner Institut of Design Thinking“ in Stanford. So gelingt es SAP, anhand von Design Thinking und gängigen Methoden bzw. Technologien für Softwareentwicklung, die Schaffung von benutzerfreundlicher Software zu beschleunigen. Im Mittelpunkt dieses Prozesses steht immer der Benutzer. Von großem Vorteil ist die Zusammenarbeit mit den Kunden, was die Möglichkeit bietet, alle gefertigten Prototypen des Produktes zu prüfen. Dank des schnellen Kunden-Feedbacks kann die Zeitspanne zwischen dem Entstehen der Produktidee und ihrer Markteinführung wesentlich reduziert werden. Die SAP-Mitarbeiter im AppHause, die projektbezogen arbeiten, können bis zu 30 Personen umfassen. Die Arbeitsgruppen involvieren Mitarbeiter verschiedener Abteilungen, die vor Ort direkten Kontakt mit dem Kunden haben, um mit diesen neue Benutzeroberflächen und Applikationen zu entwickeln.

Design Thinking kann auch in Richtung SAP-HANA eingesetzt werden. Ein hervorragendes und für unsere Projektgruppenarbeit relevantes Beispiel dafür ist die Zusammenarbeit von SAP -Design-Experten mit dem Nationalen Zentrum für Tumorerkrankungen in Heidelberg (NCT). Die HANA-basierte Anwendung nennt sich „Patientendaten Explorer“. Das Produkt erlaubt den Ärzten, einen schnellen Zugriff auf therapierelevante Patientendaten, die aus verschiedenen Quellen stammen, sowie deren Analyse. Auf diese Weise können lebensrettende Entscheidungen in kürzester Zeit getroffen werden und die Patienten können von einer erhöhten Lebensqualität profitieren. Dank der innovativen Arbeitsweise des „Design Thinking“- Konzepts sind diese tollen Ergebnisse zustande gekommen[CO13].

4. Fazit und Ausblick

Menschen und Unternehmen, haben heutzutage mit vielen Probleme zu kämpfen. Der Klimawandel, die Technologieentwicklung und die Knappheit der Ressourcen bringen Herausforderungen mit sich, deren Lösungen noch zu finden sind. Der einzige Weg, eine nachhaltige Lösung in dieser komplexen Situation zu finden, ist eine neue, kreative und ungewöhnliche Denkweise einzusetzen. Dafür muss man die gewohnten Denkpfaden verlassen und mentale Schranken überwinden, von konvergentem zu divergentem Denken wechseln. Wie bereits der berühmte Physiker Alber Einstein sagte: „Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind.“

In diesem Kontext ist Design Thinking kein Wundermittel und kann nicht für Erfolg garantieren. Wichtig ist allerdings, dass diese auf die Bedürfnisse der Menschen bzw. der Kunden ausgerichtete Methode eine signifikant höhere Wahrscheinlichkeit für eine erfolg-

reiche Lösung bietet. Viele etablierten Unternehmen wie zu Beispiel SAP, BMW, Deutsche Bank, Google usw. profitieren schon von den Vorteilen des Design-Thinkings-Konzeptes. Es existiert bereits auch einen Studiengang zu Design Thinking am Hasso-Plattner –Institut. Wenn man dies alles berücksichtigt, spricht nichts dagegen es einmal selber zu probieren.

A. Anhang

Literatur

[Bra10] Anthony J. Bradley. A new way of thinking for enterprise architects. http://blogs.gartner.com/anthony$_b$*radley*/2010/04/21/$a - new - way - of - thinking - for - enterprise - architects/, April2010.

[Bro06] Tim Brown. Innovation through design thinking - video des vortrages von tim brown am massachusetts institute of technology. http://video.mit.edu/watch/innovation-through-design-thinking-9138/, Maerz 2006.

[Bro09] Tim Brown. *Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation.* HarperCoollins, 2009.

[CO13] Gloria Costa and Charlotte Otter. Sap holt kunden ins apphaus. http://de.news-sap.com/2013/11/26/sap-apphaus-heidelberg/, November 2013.

[Com13] Computerwoche. Die produktrevolution beginnt in den prozessen. http://www.computerwoche.de/a/die-produktrevolution-beginnt-in-den-prozessen, März 2013.

[ER13] Juergen Erbeldinger and Thomas Ramge. *Durch die Decke denken: Design Thinking in der Praxis.* Redline Verlag, 2013.

[Gil] Bob Gill. collaborative designers. http://collaborativedesigners.com/about/.

[GM13] Jochen Guertler and Johannes Meyer. *30 Minuten Design Thinking.* GABAL Verlag GmbH, 2013.

[GMP09] Alexander Grots and M.A. Margarete Pratschke. Design thinking – kreativität als methode. *Marketing Review St. Gallen*, pages 18–22, 2009.

[Gue12] Milan Guenther. *Intersection: How Enterprise Design Bridges the Gap between Business, Technology, and Peoplen.* MorganKaufmann, 2012.

[Ins13] Hasso Plattner Institut. Aktuell. http://www.hpi.uni-potsdam.de/d$_s$*chool*/*news*/*beitrag*/$hpi - school - of - design - thinking - bewerbt - euch - jetzt - fuer - das - wintersemester - 20132014.html$, 2013.

[Ins14a] Hasso Plattner Institut. Kernelemente. http://www.hpi.uni-potsdam.de/d$_s$*chool*/*designthinking*/*kernelemente.html*, 2014.

[Ins14b] Hasso Plattner Institut. Was ist design thinking? http://www.hpi.uni-potsdam.de/d$_s$*chool*/*designthinking.html*, 2014.

[KL07] Tom Kelly and Jonathan Littmann. *The Art of Innovation: Lessons in Creativity from IDEO, America's Leading Design Firm.* Random House LLC, 2007.

[Kob14] Joachim Kobuss. Juergen erbeldinger: Durch die decke denken. http://www.designersbusiness.de/info/literatur/juergen-erbeldinger, Mai 2014.

[Kre14] Kreativitätstechniken.info. Design thinking. http://xn–kreativittstechniken-jzb.info/kreativitaetsframeworks/design-thinking/, 2014.

[Kul13] Sabine Kulhanek. Geschäftliche herausforderungen mit design thinking bewältigen. http://de.news-sap.com/2013/07/15/geschaeftliche-herausforderungen-mit-design-thinking-bewaeltigen/, Juli 2013.

[Len13] Alicia Lenze. Apphaus in heidelberg eröffnet. http://de.news-sap.com/2013/11/15/apphaus-in-heidelberg-eroffnet/, November 2013.

[Lue11] Alexander Luebbe). *Tangible Business Process Modeling, Design and Evaluation of a Process Model Elicitation Technique.* PhD thesis, 2011.

[MBN12] Alexander Maedche, Achim Botzenhardt, and Ludwig Neer. *Software for People: Fundamentals, Trends and Best Practices.* Springer Science and Business Media, 2012.

[Nor13] Don Norman. *The Design of everyday: Revised and expandet edition.* Basic Book, 2013.

[Pfe12] Sabine Pfeiffer. *Smarte Innovation.* VS Verlag fuer Sozialwissenschaften and Springer Fachmedien Wiesbaden GmbH, 2012.

[PMW09] Hasso Plattner, Christoph Meinel, and Ulrich Weinberg. *Design Thinking: Innovation lernen, Ideenwelten öffnen.* Mi-Wirtschaftsbuch, FinanzBuch-Verlag, 2009.

[SAP] SAP. Das unternehmen sap. http://www.sap.com/germany/about.html.

[SAP13] SAP. Design thinking: Mobile apps kreativ entwickeln. http://news.sap-im-dialog.com/design-thinking-mobile-apps-kreativ-entwickeln/, Juni 2013.

[SAP14] SAP. Design thinking for business innovation. https://training.sap.com/shop/course/wdt100-design-thinking-for-business-innovation-classroom-010-de-de/, 2014.

[Sch10a] Peter J. Schmitt. Forget design thinking and try hybrid thinking. http://ps.identitaetsarchitekten.de/?tag=hybrid-thinking, Februar 2010.

[Sch10b] Friedhelm Schwarz. Die geschichte des design thinking. http://www.dtn.brain-server.de/?p=79, April 2010.

[Upb11] Bruce Upbin. The power of hybrid thinking. http://www.forbes.com/sites/bruceupbin/2011/12/05/the-power-of-hybrid-thinking/, Mai 2011.

[Wei13] Stephan Weiland. Digital disruption: Das geheimnis der drei cs und fÜnf tipps. https://www.bitkom-trendkongress.de/news/digital-disruption-das-geheimnis-der-drei-cs-und-fNovember 2013.

VERY LARGE
BUSINESS APPLICATIONS
Carl von Ossietzky Universität Oldenburg

Klassische vs. agile Softwareentwicklung: Einsatz von Scrum in der Projektgruppe

Seminararbeit
im Rahmen der Projektgruppe VLBA inMemory Planung mit SAP HANA

Themensteller:	Prof. Dr.-Ing. Jorge Marx Gómez
Betreuer:	Dipl.-Inform. Nils Giesen
Vorgelegt von:	Benjamin Hemken benjamin.hemken@uni-oldenburg.de
Abgabetermin:	28.7.2014

Inhaltsverzeichnis

Abbildungsverzeichnis **3**

1 Aufgabenstellung und Zielsetzung **4**

2 Charakteristika **4**
2.1 Schwergewichtig und leichtgewichtig 5
2.2 Sequenz 5
2.3 Inkrementell und iterativ 5
2.4 Agil 6
2.5 Bewertung 7

3 Allgemeine agile Methoden **8**
3.1 Visionsermittlung 8
3.2 User Stories 8
3.3 Priorisierungsmethoden 9
3.4 Schätzverfahren 9
3.5 Code Reviews und Testautomatisierung 10
3.6 Continuous Integration und Delivery 10

4 Scrum **10**
4.1 Einleitung 11
4.2 Rollen 11
4.3 Artefakte 14
4.4 Strategische Planung 15
4.5 Sprint 17
4.6 Reporting 18

5 Software Kanban **20**
5.1 Einleitung 21
5.2 Messgrößen 21
5.3 Techniken 21
5.4 Kanban Board (Artefakt) 22

6 Auswahl **22**

7 Anwendung von Scrum **23**

Literaturverzeichnis **25**

Abbildungsverzeichnis

1 Der Scrum Prozess [Glo13, S. 9] 12
2 Strategischer Planungsprozess[Glo13, S. 112] 15
3 Scrum Taskboard[Glo13, S. 167] 19
4 Burndown Chart [Glo13, S. 210] 19
5 Parking-Lot-Chart [Glo13, S. 214] 20
6 Velocity-Chart [Glo13, S. 214] 20
7 Beispiel Kanban Signalkarte [Epp11, S. 117] 22
8 Beispiel Kanban Board [Epp11, S. 121] 23

1 Aufgabenstellung und Zielsetzung

Der Chaos-Studie der Standish Group zufolge werden sehr viele IT-Projekte nicht erfolgreich abgeschlossen. Nicht erfolgreich bedeutet, dass sie entweder nicht im gesetzten Zeitrahmen umgesetzt werden, das Budget überschreiten oder die Anforderungen nicht erfüllen. Die Studie von 2012 zeigt, dass mittlerweile 37% aller Projekte erfolgreich, 42% mit Mängeln und 21% nicht zum Abschluss kommen. [Gro12, Vgl.] Das IT-Projektmanagement beschäftigt sich mit dem permanenten Interessenskonflikt zwischen den projektbezogenen Bestimmungsgrößen Leistung/Funktionalität, Qualität, Projektdauer und Projektressourcen, um ein Projekt erfolgreich abzuschließen. Beispiele für Projektressourcen sind Budget, Personal und Betriebsmittel. [WR08, Vgl. S. 15f.] Zu den Aufgaben des IT-Projektmanagements gehört die Vorgehensplanung zur Auswahl und Anpassung von Vorgehensmodellen. [WR08, Vgl. S. 23] Vorgehensmodelle stellen den Ablauf der Entwicklung von Software-Systemen abstrakt dar. Häufig werden diese Modelle auch als Software-Prozessmodelle bezeichnet. Die Unterteilung des Entwicklungsprozesses in einzelne Prozess-Schritte wird mit Phasen beschrieben. Diese Phasen befassen sich beispielsweise jeweils mit der Planung des Durchführungs-Prozesses, Spezifikation der Produkt-Anforderungen, Design des Produkts, Implementierung (Umsetzung) und Testen des Produkts. [Han10, Vgl. S. 1 ff.] Die Unterteilung dient der Reduktion der Komplexität zur Verbesserung der Beherrschbarkeit des Entwicklungs-Prozesses. In welcher Art- und Weise die Phasen durchgeführt werden, bestimmt das konkrete Vorgehensmodell. [WR08, Vgl. Seite 25] In der Projektgruppe wird ein Projekt bearbeitet, das die oben genannten Merkmale eines IT-Projekts erfüllt. [Bol09, Vgl. S. 1ff.] Ziel dieser Seminararbeit ist ein geeignetes, modernes Vorgehensmodell für den Einsatz in der Projektgruppe zu finden und beispielhaft darzustellen.

2 Charakteristika

Im Laufe der letzten Jahre und Jahrzehnte entstanden diverse Vorgehensmodelle, die sich durch bestimmte Charakteristika zusammenfassen lassen. Jedes der Modelle hat seine individuelle Eignung für verschiedene Anwendungszwecke. [Han10, Vgl. S. 3ff.] Zuerst werden die Merkmale schwergewichtig und leichtgewichtig definiert. Anschließend Sequenz, Inkrementell und iterativ sowie Agil. Zuletzt wird das Projekt im Kapitel Bewertung kurz typisiert und die Merkmale hinsichtlich ihrer Eignung bewertet.

2.1 Schwergewichtig und leichtgewichtig

Vorgehensmodelle werden in schwergewichtige und leichtgewichtige Prozessmodelle unterteilt. Schwergewichtige Modelle sind formal definiert und stark dokumentengestützt. Jede Phase wird ausführlich dokumentiert und der Ablauf ist klar beschrieben. Schwergewichtige Vorgehensmodelle sind dann einzusetzen, wenn Softwarequalität von großer Bedeutung ist. Allerdings gelten diese Modelle als unflexibel - insbesondere dann, wenn die Anforderungen sich im Projektverlauf ändern. Leichtgewichtige Vorgehensmodelle sind insbesondere für kleine Teams geeignet, bei denen die Anforderungen nicht vollständig definiert sind. Kommunikation im Team und mit dem Kunden spielt dabei eine wichtige Rolle. Informationen müssen nicht explizit schriftlich fixiert werden. Die funktionierende Software steht anstelle ihrer Dokumentation im Vordergrund. [Han10, Vgl. S. 2f.]

2.2 Sequenz

In den 1970er Jahren wird mit dem Phasenmodell eine Struktur geschaffen, die den Ablauf des Entwicklungs-Prozesses in eine klar definierte Sequenz unterteilt. Bei rein sequentiellen Modellen ist der Ausgangspunkt, dass die Anforderungen an das Software-Produkt anfangs bereits feststehen und sich diese bis zum Schluss nicht ändern. [Han10, Vgl. S. 3f.] Dadurch gelten sequentielle Modelle als starr. Die Einbindung des späteren Anwenders erfolgt nur zu Beginn und am Ende des Modells. Die Dauer zwischen Projektidee und Inbetriebnahme ist in der Regel sehr lang. [WR08, Vgl. S. 31]

2.3 Inkrementell und iterativ

Eine inkrementelle Vorgehensweise unterteilt ein komplexes IT-System in sinnvolle, selbstständig entwickelbare Teile (Inkremente), die nacheinander oder parallel erstellt werden. Der große Vorteil einer inkrementellen Vorgehensweise ist, dass frühzeitig lauffähige Teilsysteme entstehen. Dadurch lassen sich Risiken früh erkennen und gewonnene Erfahrungen können in weitere Inkremente einfließen. Der Anwender bekommt durch die Vorgehensweise früh einen Eindruck des späteren Endprodukts. Eine Gefahr bei einer inkrementellen Vorgehensweise ist, dass die System-Architektur frühzeitig festgelegt wird. Möglicherweise wird später festgestellt, dass diese den Anforderungen nicht mehr genügt. Da der Anwender frühzeitig Einsicht in das Produkt bekommt, ist mit neuen oder geänderten Anforderungen zu rechnen. Dadurch werden Aufwandsschätzungen erschwert. [WR08, Vgl. S. 32f.] Ein iterativ, inkrementelles Vorgehensmodell erweitert diese Vorgehensweise. Der Entwicklungsprozess ist dabei als evolutionärer Prozess beschrieben. Ausgehend von einem Prototypen erfolgt über Verbesserungen eine Annäherung an das Endprodukt. Es werden

also ganze Phasen und deren Schritte mehrfach wiederholt (iterativ) und das Produkt Stück für Stück (inkrementell) erweitert. [WR08, Vgl. S. 34.]

2.4 Agil

Agile Methoden bauen auf dem in Kapitel 2.3 genannten Konzept der iterativen und inkrementellen Vorgehensweise auf. [WR08, Vgl. Seite 36f.] 2001 wird das Agile Manifest veröffentlicht. Die Kernaussagen des Manifests sind, dass (1) Individuen und Interaktionen statt Prozesse und Werkzeuge, (2) funktionierende Software statt Dokumentation, (3) Zusammenarbeit mit dem Kunden statt Verträge und (4) Flexibilität bei Veränderungen statt Realisierung eines Plans im Fokus stehen.[ea01] Modelle, die auf diesem Konzept aufbauen gelten als leichtgewichtig.[Han10, Vgl. S. 6.] Agile Methoden stehen zu den traditionellen Methoden der Softwareentwicklung in Konkurrenz. Bei der traditionellen Softwareentwicklung wird früh versucht, alle Anforderungen und Entwicklungsziele zu erfassen. Dabei wird der Entwicklungsprozess in Verbindung mit umfangreicher Dokumentation nachvollziehbar gestaltet und vorangetrieben. Mit der agilen Vorgehensweise werden schnell Entwicklungsergebnisse erzielt und die Zeitspanne bis zur Markteinführung (auch bekannt als „time-to-market“) verkürzt. Während der Entwicklung werden sich ändernde Anforderungen (auch genannt „moving targets“) berücksichtigt. [WR08, Vgl. S. 34ff.]

Viele moderne, agile Vorgehensmodelle arbeiten nach dem Prinzip des Lean Product Developments (in der Softwareentwicklung auch Lean Software Development genannt). Zu Deutsch *schlanke* Produktentwicklung. Es stammt ursprünglich aus dem Lean Manufacturing. Lean Manufacturing arbeitet nach dem Pull-Prinzip. Das bedeutet, dass immer nur dann produziert wird, wenn eine Nachfrage besteht. Ein weiterer wichtiger Aspekt des Lean Manufacturing ist die Übertragung von Verantwortung. Dadurch wird die Kreativität und Mitarbeit der Projektmitglieder gesteigert. Autonome Teams koordinieren sich selbst. Nicht zuletzt basiert das System auf kontinuierlicher Verbesserung. Das bedeutet, dass den Teammitgliedern regelmäßig Freiraum zur Erarbeitung von Verbesserungen gegeben wird.[Glo13, Vgl. S. 28ff.] Zudem werden diese Aspekte hervorgehoben: Erstens ist eine vollständige Teamauslastung nicht optimal. Es wird mit einer steigenden Teamauslastung ein exponentielles Wachstum der wartenden Aktivitäten beschrieben. [Glo13, Vgl. S. 34] Zweitens ist Variabilität (Ungewissheit) normal bei der Produktentwicklung und nicht planbar. Es gilt dabei nachträgliche Änderungen bei den Anforderungen nicht zu verhindern sondern diese als Chancen zu verstehen. *Gerade der Faktor Ungewissheit ist die Größe, durch die Innovation entsteht.* [Glo13, Vgl. S. 36f.] Drittens ist sogenannter Ballast (*Waste*) zu eliminieren. Waste verhindert oder verlangsamt den gleichmäßigen

Arbeitsfortschritt. Einige Beispiele für Waste in der Softwareentwicklung: Nur teilweise erledigte Arbeiten, unnötige Zusatzprozesse, unnötige Zusatzfeatures, ständiges Wechseln der bearbeiteten Aufgabe, Wartezeiten, unnötige Dokumentenflüsse und unerkannte Programmfehler. Viertens sind Entscheidungen so spät wie möglich zu treffen, da die Informationslage sich im Verlauf verbessert. Fünftens ist die Software so schnell wie möglich auszuliefern. Zuletzt steht die Akzeptanz des Anwenders im Fokus. [Epp11, Vgl. S. 44ff.]

2.5 Bewertung

In diesem Abschnitt werden die verschiedenen Charakteristika hinsichtlich ihrer Eignung für die Projektgruppe bewertet. Zuerst eine kurze Risikoanalyse des Projekts. Bei dem Projekt handelt sich eine Forschungsarbeit, bei der eine komplexe Aufgabenstellung auf Basis von SAP Hana gelöst werden soll. Da diese Technologie relativ neu am Markt ist (seit 2011[AG11]), besteht ein technisches Risko. Es äußert sich darin, dass bisher wenig Erfahrung mit der verwendeten Technologie in Verbindung mit möglichen Anwendungsbereichen existiert. Außerdem ist wegen der hohen Komplexität, die tendenziell bei der kommenden Projektaufgabe gegeben ist, ein Produktrisiko vorhanden.[WR08, Vgl. S. 247] Jetzt werden die Merkmale von Vorgehensmodellen hinsichtlich der Eignung für das Projekt diskutiert. An erster Stelle ist zu prüfen, ob ein leichtgewichtiges oder schwergewichtiges Vorgehensmodell zu wählen ist. Da es sich tendenziell um ein *Semidetached Mode*-Projekt handelt (unterschiedlicher Wissensstand der Mitglieder, räumlich verteilte Arbeit, mittlere Teamgröße)[Han10, Vgl. S. 2], sind grundsätzlich beide Varianten denkbar. Maßgeblich für die Auswahl ist die Tatsache, dass bei dieser Projektaufgabe die Anforderungen im Vorfeld nicht vollständig definiert sind. Im Gegensatz dazu werden später in Zusammenarbeit zwischen dem Auftraggeber und der Projektgruppe die Anforderungen kooperativ ermittelt und je nach Entwicklung des Projektverlaufs angepasst. Bei Verwendung eines schwergewichtigen Vorgehensmodells würde dies einen großen Änderungsaufwand bedeuten. Die Entscheidung fällt auf ein *leichtgewichtiges Vorgehensmodell.*[Han10, Vgl. S. 2f.] Es ist zu ermitteln, ob ein inkrementell- iteratives Vorgehen bei der Projektgruppe sinnvoll ist. Wie bereits beschrieben sind mit dem Projekt Risiken verbunden. Um die Komplexität beherrschbar zu machen und um sich ändernde Anforderungen flexibel abzubilden, wird ein *inkrementell- iteratives* Vorgehensmodell bevorzugt. [Han10, Vgl. S. 5.] Im Rahmen dieses Forschungsprojekts sollen sichtbare Ergebnisse geliefert werden, die am Ende einen Nutzen darstellen. Dabei stehen Kreativität und Flexibilität vor strikter Planung. Es besteht eine Tendenz zu einem *agilen* Vorgehensmodell. Dies zielt darauf ab, das gesamte kreative Potential aller Projektgruppenmitglieder auszuschöpfen.

3 Allgemeine agile Methoden

In diesem Kapitel werden vorab allgemeine agilen Methoden vorgestellt, die unabhängig vom konkreten Vorgehensmodell genutzt werden können. Zuerst werden Methoden zur Ermittlung und Ausformulierung einer Vision dargestellt. User Stories unterstützen die Anforderungsanalyse. Priorisierungsmethoden helfen bei der Bestimmung der Priorität von Teilen eines Softwarepodukts. Code Reviews decken Schwachstellen im Quelltext und in der Organisation aus Entwicklersicht auf. Testautomatisierung erlaubt es mit hoher Frequenz zu testen. Zuletzt werden die beiden Methoden Continuous Integration und Continuous Delivery erklärt.

3.1 Visionsermittlung

Freewriting ist eine Methode zur erstellung einer Vision. Beim Freewriting werden alle Ideen zu einer Produktidee ohne jegliche Kritik und Unterbrechung aufgeschrieben. Durch Analyse, Reflexion und weiterer Detaillierung relevanter Aspekte wird die Vision gebildet. [Glo13, Vgl. S. 120] Ein *Elevator Pitch* ist der Versuch in 30 Sekunden auszudrücken, was man besonderes leistet. Je spezifischer und klarer die verwendeten Worte sind, desto aussagekräftiger ist die Vision. Er kann diesen Aufbau haben: Für *(Kunden)*, die *(Beschreibung des Bedarfs oder der Gelegenheit)*, ist das *(Produktname)* eine *(Produktkategorie)*, die *(Hauptvorteil, Grund das Produkt zu kaufen)*; anders als *(Alternative des Wettbewerbs)* kann unser Produkt *(Beschreibung des Hauptunterschieds)* [Glo13, Vgl. S. 121] Hier ein konkretes Beispiel: Für Motorradhändler, die eCommerce betreiben, ist der CSB Shop eine Full-Service-Lösung, die Ihre Prozesse vollständig in einem System integriert; anders als bei herkömmlichen Shoplösungen können Sie in Echtzeit Ihr Präsenzgeschäft mit dem Onlinehandel vereinen.

3.2 User Stories

User Stories beschreiben Funktionalitäten, die für den Benutzer eines Systems wertvoll sind. Sie haben typischerweise diese Form: Als *Anwender mit der Rolle* benötige ich *eine Funktionalität*, damit ich *den Nutzen bekomme*. Ein Beispiel: Als Mechaniker benötige ich einen Teilekatalog, damit ich weiß, welches Teil ich für ein Fahrzeug verbauen kann.[Glo13, Vgl. S. 133]

3.3 Priorisierungsmethoden

Ein Softwareprodukt besteht aus vielen möglichen Teilen, die entwickelt werden können. Eine zentrale Frage ist, welche Teile welche Priorität aufweisen. Mögliche Priorisierungsmethoden für Teile einer Applikation sind MusCoW, 1000 Pingpong-Bälle, Kano und Relatives Gewicht. MusCoW teilt die Teile und Must, Could und Wish-Funktionalitäten ein, wobei die Priorität von Must zu Wish absteigend ist. Bei 1000 Pingpong-Bälle wird eine Basismenge (z.B. 1000 Bälle) auf die einzelnen Teile verteilt. Je wertvoller eine Funktionalität geschätzt ist, desto mehr Bälle erhält es. Kano unterschiedet Produktmerkmale in sechs verschiedene Kategorien mit dem Ziel die Kundenzufriedenheit zu adressieren. Dabei wird direkt mit dem Anwender zusammengearbeitet. Relatives Gewicht führt in acht Schritten zur Priorisierung. Dabei werden relativer Vorteil (1-9) und die Strafe (1-9), wenn ein Teil fehlt aufsummiert, der Aufwand in Storypoints und das Risiko bewertet und anschließend in einer Prioritätszahl dargestellt.[Glo13, Vgl. S. 134ff.]

3.4 Schätzverfahren

Im agilen Umfeld werden statt der Aufwände (Dauer für eine Umsetzung) die Größen der Teile einer Applikation geschätzt. Mit Größe ist der individuelle Grad des Verständnisses eines jeden Entwicklers für eine Aufgabe gemeint. Entgegen der Möglichkeit den Aufwand bei normierter Akkordarbeit zu schätzen, gilt dies nicht für Software-Entwicklung. Ein Grund dafür ist beispielsweise, dass ein erfahrener Programmierer bis zu 25 mal produktiver ist als ein unerfahrener. Außerdem entstehen bahnbrechende und effizienzsteigernde Ideen und Lösungen nicht unter Zeitdruck sondern in einem kreativen Umfeld (zum Beispiel auf dem Nachhauseweg). Es ist im innovativen Umfeld unmöglich vorhersagbar, wie lange eine Programmierleistung dauert. Für eine Bewertung wird eine Referenz, eine Maßeinheit und eine Skala benötigt. Als Referenz dient die kleinste User Story. Als Maßeinheit werden Storypoints verwendet. Die agile Community hat sich auf die Cohnsche Unreine-Fibonacci-Reihe als Skala geeinigt. Je weniger Verständnis für eine User Story seitens des Teams besteht, desto mehr Storypoints werden ihr zugeordnet. Anstatt der populären Schätzmethode Planning Poker wird ein schnelleres Schätzverfahren empfohlen: *Magic Estimation*. Erst schreibt der Verwalter der Produkt-Teile (in Scrum Product Owner) alle User Stories auf Karten und erstellt eine Skala von 1 bis 100 (Beschriftung nach zuvor genannter Skala). Jedes Teammitglied bekommt die gleiche Anzahl Karten. Jetzt liest jedes Teammitglied seine Karte und ordnet sie den Zahlen auf der Skala zu - ohne sich dabei mit anderen auszutauschen. Je größer die zugewiesene Zahl, desto geringer ist das Verständnis. Wenn die eigenen Karten verteilt sind, werden die zugewiesenen

Karten der anderen durchgelesen und gegebenenfalls nach eigenem Ermessen umgelegt. Dadurch werden Meinungsverschiedenheiten deutlich. Wenn es nur noch springende Karten gibt oder sich keine Karte mehr bewegt, ist das Spiel beendet. Die Teammitglieder schreiben jetzt auf jede Karte die Bewertung - es liegt eine Schätzung vor. Die Schätzung wirkt sich dann auch auf die Priorisierung der Teile der Applikation aus. Items, die den gleichen Geschäftswert aufweisen, jedoch weniger Storypoints (Umfang) besitzen, sind zu bevorzugen. [Glo13, Vgl. S. 142ff.]

3.5 Code Reviews und Testautomatisierung

Beim *Code Review* tauscht sich das Entwicklungsteam über technische Lösungen und organisatorische Regelungen aus. Oftmals sind Refactoring-Maßnahmen (Umstrukturierung des Quellcodes) Ergebnis dieser Reviews. Diese führen zu neuen User Stories. [Epp11, Vgl. S. 97ff.] Bei der *Testautomatisierung* werden Anforderungen für den Auftraggeber automatisch validiert und verifiziert. Der Vorteil dabei: Das Testen kann in hoher Frequenz durchgeführt werden. Die Entwicklung einer Anforderung kann somit in kleinen Schritten durchgeführt werden. Beim parallelen, automatisierten Testen wird sichergestellt, dass die Entwicklung einer Anforderung mit dem Entwicklungsstand anderer Anforderungen vereinbar ist.[Epp11, Vgl. S. 103]

3.6 Continuous Integration und Delivery

Continuous Integration verlangt, dass die Entwickler ihre Arbeit häufig integrieren. Dies sollte wenigstens einmal am Tag erfolgen. Jede Integration wird auf ihre Qualität geprüft. Das kann beispielsweise durch automatisierte Tests (siehe Kapitel 3.5) erfolgen. Ziel dieser Methode ist es, Integrationsprobleme abzuschwächen und zusammenhängende Software schneller zu entwickeln. [Fow06] *Continuous Delivery* baut auf Continuous Integration auf. Diese Methode bewerkstelligt, dass zu jeder Zeit die Software in den Produktivbetrieb genommen werden kann. [Fow13]

4 Scrum

Als erstes agiles und schlankes Vorgehensmodell wird Scrum vorgestellt. Zuerst eine Einleitung, dann folgt die Definition der Rollen in Scrum. Anschließend wird der Prozess der strategischen Planung in Scrum dargestellt. Im weiteren Verlauf erfolgt die Erläuterung des Sprints mit seinen verschiedenen Meetings und ein kurzer Einblick in das Reporting.

4.1 Einleitung

Der Begriff Scrum stammt ursprünglich aus dem Rugby. Zu Deutsch kann er mit *Gedränge* übersetzt werden. Dieser Begriff ist in Anlehnung an den Zusammenhalt des Entwicklerteams, der bei der Projektarbeit entsteht, gewählt. Er hebt hervor, dass wie beim Rugby wenige Regeln vorhanden sind. Ursprünglich wurde der Begriff Scrum 1986 Nonaka und Takeuchi aufgegriffen und von DeGrace und Hulet-Stahl 1990 weitergeführt.[Glo13, Vgl. S. 7] 1996 wird Scrum auf der OOPSLA als Methode vorgestellt. 2001 veröffentlichen Schwaber, Souther und Beedle das Buch *Agile Software Development with Scrum* mit dem Ziel die Software-Entwicklung zu verbessern. [Glo13, Vgl. S. 24] Scrum wurde ursprünglich nur als agiles Vorgehensmodell gesehen. Allgemeingültige Bereiche, wie die Personalführung führen dazu, dass Scrum ab 2003 auch als Projektmanagementverfahren eingesetzt wird. Es hilft Mitarbeitern fokussiert und konzentriert zu arbeiten. In Wahrheit wird Scrum als Produktentwicklungsmethode gesehen. Es gilt folglich dabei sich von dem klassischen Projektgedanken zu befreien und stattdessen einen kontinuierlichen Strom an Produktteilen auszuliefern (ohne zeitliche Limitierung). Scrum stellt den gesamten Produktlebenszyklus von der Idee bis zur Eliminierung dar.[Glo13, Vgl. S. 15ff.] 2011 wird von Nonaka und Takeuchi beschrieben, dass jedes Organisationsmitglied weise handeln soll, um in jeder Situation kreativ und flexibel zu reagieren. Dies ist angelehnt an den Begriff des Lean Product Developments (siehe Kapitel 2.4). Scrum ist also die Antwort auf die Herausforderungen von moderner Produktentwicklung. Es wird mittlerweile als Managementframework abseits der eigentlichen Software-Entwicklung verstanden und angewandt. [Glo13, Vgl. S. 26] Abbildung 1 stellt den Scrum-Prozess grafisch dar. [Glo13, Vgl. S. 9]

4.2 Rollen

Scrum fasst Rollen nicht als hierarchische Positionen innerhalb einer Firma oder Organisation auf. Eine Rolle ist vielmehr als ein Bündel Aufgaben zu verstehen, wofür ein Teammitglied aus eigener Überzeugung Verantwortung übernimmt. Scrum gibt jedem Einzelnen die Verantwortung für das eigene Handeln zurück. Konkret äußert sich dies so, dass ein Teammitglied ein *Commitment* abgibt, mit dem es sich freiwillig selbst verpflichtet für die Aufgabe einzustehen. [Glo13, Vgl. S. 64ff.] Folgend werden die Rollen von Scrum definiert. Das *Entwicklungsteam* besteht aus Spezialisten (Team-Mitgliedern), die gemeinsam das Produkt ausliefern. Es ist für die Qualität des Produkts verantwortlich. Die Aufgaben des Entwicklungsteams sind: Anforderungsermittlung- und Analyse, Design, Implementation und Testen. [Glo13, Vgl. S. 66ff.] Für den Erfolg eines Projekts ist ein Commitment seitens des Entwicklungsteams ausschlaggebend. Maßnahmen, die ein Commitment

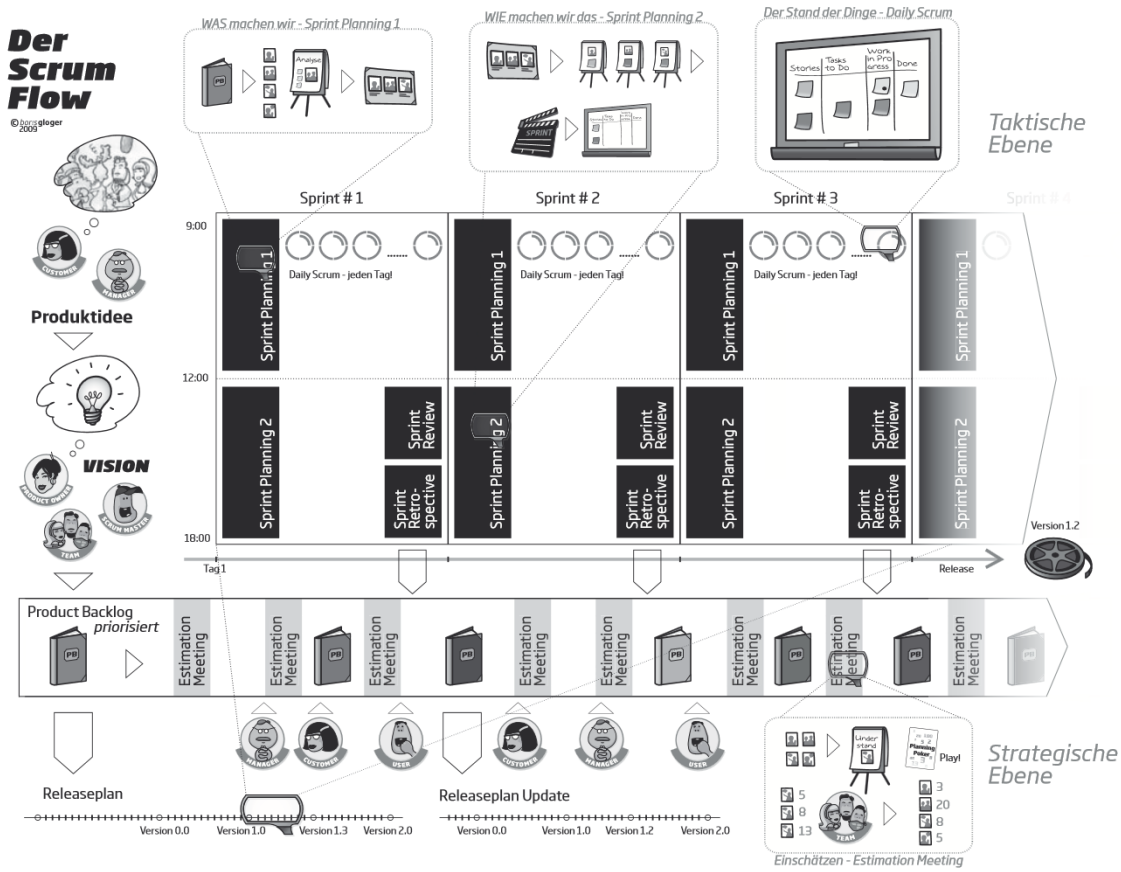


Abbildung 1: Der Scrum Prozess [Glo13, S. 9]

fördern: Erstens muss ein Teammitglied gefragt werden, ob es an dem Projekt teilnehmen möchte. Zweitens müssen realistische aber herausfordernde Ziele gesetzt werden. Drittens sollten Teammitglieder, die einen guten Job machen, Anerkennung erhalten. Anerkennung kann beispielsweise durch ein Feedback am Ende des Sprints, messbare Steigerung der Leistung durch Burn-Down-Charts oder gegenseitige Anerkennung der Teammitglieder erreicht werden. Das führt dazu, dass Scrum-Teams sich ständig verbessern wollen. Gerade bei kognitiv schwierigen Aufgaben sind finanzielle Anreize weniger sinnvoll. Stattdessen ist ein autonomes Arbeitsumfeld leistungsfördernd. [Glo13, Vgl. S. 69] Das Team arbeitet cross-funktional, was soviel bedeutet, dass jeder auch die Aufgaben des anderen übernimmt. Es ist also somit beispielsweise üblich, dass ein Tester auch entwickelt und ein Entwickler testet. [Glo13, Vgl. S. 76] Der *Product Owner* kommuniziert eine klare Vision des Produkts, legt die Eigenschaften fest und bewertet am Ende die erreichten Ergebnisse. Er ist dafür verantwortlich, dass das Entwicklungsteam an den wertvollsten

Aspekten des Produkts arbeitet. Die wertvollsten Aspekte sind beispielsweise jene, die am profitabelsten für das Unternehmen sind und einen Return on Investment (ROI) sichern. Er verfolgt die gleichen Ziele wie das Entwicklungsteam. [Glo13, Vgl. S. 78ff] Der *ScrumMaster* schützt das Team vor äußeren Störungen und räumt Blockaden, auch Impediments genannt, schnellstmöglich aus dem Weg. Diese Impediments hindern das Team am produktiven Arbeiten. Sie treten beispielsweise in diesen Bereichen auf: Im Software-Entwicklungsprozess, in der Kommunikation und Abstimmung, im Scrum-Prozess, in der organisationalen Einbettung des Teams, in der Zusammenarbeit mit dem Business, im persönlichen Bereich, bei den Priorisierungen und bei Störungen während der eigentlichen Arbeit. [Glo13, Vgl. S. 89f.] Er vermittelt dem Entwicklungsteam und Product Owner, was Scrum ist und wie es angewandt wird. Er setzt sich dafür ein, dass die Kommunikation zwischen den beiden Rollen optimal funktioniert. Er lässt Scrum Wirklichkeit werden.[Glo13, Vgl. S. 87] Der ScrumMaster ist kein Teamleiter im klassischen Sinne. In der klassischen Sichtweise treffen Teamleiter möglicherweise Entscheidungen, ohne das Team zu befragen. Oft arbeitet im dem Fall ein Teamleiter auch bei dem Projekt aktiv mit. Der ScrumMaster ist nur dann zu Entscheidungen befugt, wenn in dringenden Fällen der Product Owner oder das Entwicklungsteam diese nicht treffen kann. Der ScrumMaster sollte immer nur diese Rolle gleichzeitig wahrnehmen, damit er diese Tätigkeit sinnvoll ausfüllen kann.[Glo13, Vgl. S. 97f.] Gerade am Anfang muss der ScrumMaster klare Handlungsanweisungen geben. [Glo13, Vgl. S. 76] Der *Customer (Kunde)* ist der Auftraggeber des Projekts. Er bezahlt die Produktentwicklung und ist derjenige, der das Produkt am Ende der Laufzeit besitzen möchte, um seine Anwender zufriedenzustellen. In der ursprünglichen Scrum-Formulierung sind Customer und Product Owner nicht voneinander differenziert. Dabei sollte diese kombinierte Rolle einerseits den Gewinn des Projekts maximieren und andererseits den Markt und die Bedürfnisse seiner Kunden vollständig kennen. Es stellt sich heraus, dass dies nicht immer von einer Person abgebildet werden kann. Der Product Owner hat jetzt die Aufgabe seinen Customer von dem Produkt zu begeistern. [Glo13, Vgl. S. 101f.] Mit der Rolle des *Users* werden die Benutzer des Produkts abgebildet. Nicht immer ist er auch gleichzeitig der Customer. User kennen oder bewerten die Anforderungen. Auch diese Rolle ist in der ursprünglichen Formulierung von Scrum nicht enthalten. Die Besonderheit beim User ist, dass er keine Verantwortung trägt. Dennoch muss das Team den User stetig im Blick behalten, da das Produkt einzig und allein für ihn hergestellt wird. [Glo13, Vgl. S. 102f.] Der *Manager* treibt als helfende und unterstützende Führungskraft die Veränderung an. Er sorgt dafür, dass das Team alles bereitgestellt bekommt, damit es seine Mission erledigen kann. Darunter fallen beispielsweise eine Aufgabe, Räumlichkeiten, Kunden, Ressourcen, Arbeitsmittel und ein Kontext. Es gilt dabei immer

zu hinterfragen, ob typische Standardvorschriften auch für das Projektteam gelten sollten. Gegebenenfalls werden Sonderregelungen getroffen. Das Management hebt zusammen mit dem ScrumMaster Impediments auf. [Glo13, Vgl. S. 104f.]

4.3 Artefakte

Scrum enthält viele verschiedene Artefakte. In diesem Kapitel wird auf die wesentlichen eingegangen. Die *Vision* fasst die gesamte Leidenschaft, die Werte und Vorstellungen des Produkt Owners in Worte. Eine gute Version betont die Besonderheit und Wichtigkeit des Projekts. Die Projektmitglieder müssen das Gefühl vermittelt bekommen, sich auf einer Mission zu befinden. Sie muss ständig neu erklärt werden und bietet die Grundlage für die Motivation des Teams. Methoden, die beim Erstellen einer Vision helfen, werden in Kapitel 3.1 vorgestellt. [Glo13, Vgl. S. 119ff] Zur Konkretisierung der Vision wird das *Product Backlog* zusammengestellt. [Glo13, Vgl. S. 78ff] Das Product Backlog beinhaltet die gewünschten Eigenschaften und Merkmale von Produkten, jedoch keine konkreten Anforderungen. Damit ist gemeint, dass dem Entwicklerteam keine Designentscheidungen vorweggenommen werden. Product Backlog Items sind Teile der zu entwickelnden Applikation. Diese Teile sind so zu wählen, dass sie beschreibbar und verständlich sind. Auch sie stellen keine konrekten Anforderungen oder Spezifikationen dar - sondern eher Repräsentationen von Ideen. Die Ideen können allein vom Product Owner, aber auch gemeinsam mit dem Team zusammengetragen werden. Eine häufige Repräsentationsform der Backlog Items sind User Stories (siehe Kapitel 3.2). Nach der gemeinsamen Erstellung des Product Backlogs priorisiert der Product Owner die Product Backlog Items anhand ihrer Wichtigkeit (zum Beispiel nach ROI). [Glo13, Vgl. S. 81] Das *Sprint Goal* konkretisiert die Zielsetzung eines Sprints (siehe auch Kapitel 4.5). Ein gutes Ziel erfüllt SMARTe Kriterien (spezifisch, messbar, attraktiv, realistisch und terminiert). Die Definition des Ziels erfolgt durch den Product Owner und jedes weitere Teammitglied. Dadurch ensteht ein beiderseitiges Commitment. [Glo13, Vgl. S. 161ff.] Das *Selected Product Backlog* ist eine für einen Sprint vom Team ausgewählte Liste von Product Backlog Items, die zugesagt ist. Es orientiert sich am Sprint Goal. Diese Liste entsteht im Sprint Planning Meeting 1. [Glo13, Vgl. S. 165] Der *Releaseplan* stellt dar, in welchem Sprint welches Backlog Item je nach Velocity (Geschwindigkeit des Teams) voraussichtlich geliefert wird. Es handelt sich dabei um kein Planungsinstrument sondern lediglich um ein Informationsinstrument. [Glo13, Vgl. S. 149] Im *Impediment Backlog* werden die vom ScrumMaster beobachteten Impediments vermerkt. Diese Vermerkung dient dazu, dass Impediments strukturiert vom ScrumMaster abgearbeitet werden können. [Glo13, Vgl. S. 89f.] Ein *Produkt-Inkrement* ist

ein Ergebnis, das mit wenig oder keinem Aufwand an den Kunden auslieferbar ist. Eine Übergabe an ein anderes Team (wie zum Beispiel die Qualitätssicherung) ist nicht erlaubt. Das Scrum-Team übernimmt die volle Verantwortung für das entwickelte Inkrement und steht für dessen Qualität ein. [Max13, Vgl. S. 177] Die *Definition of Done* ist eine Vereinbarung zwischen dem Entwicklungsteam und dem Product Owner. Sie legt fest, was erfüllt sein muss, damit eine umgesetzte Anforderung als fertig (done) angesehen wird. Sie dient der Sicherstellung von Produktqualität. Beispiele für möglicherweise enthaltene Aspekte: Durchführung nach dem Vier-Augen-Prinzip, die Dokumentation ist angepasst, die Coderichtlinien wurden eingehalten etc. ... [Max13, Vgl. S. 179]

4.4 Strategische Planung

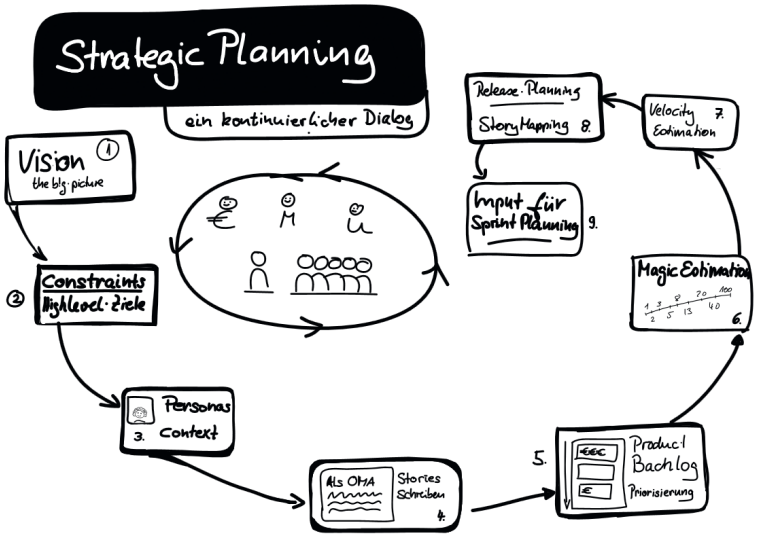


Abbildung 2: Strategischer Planungsprozess[Glo13, S. 112]

Nachdem im vorherigen Kapiteln die Erläuterung von Rollen und einige Artefakten erfolgte, werden diese jetzt durch den Planungsprozess in Scrum in den richtigen Zusammenhang gebracht. Es wird zwischen strategischer und taktischer Planung unterschieden. Die strategische Planung legt die Ziele fest, die erreicht werden sollen. Bei der taktischen Planung werden die Aktionen, die nötig sind, geplant, damit die Ziele erreicht werden. Inhalt dieses Kapitels ist die strategische Planung. Abbildung 2 zeigt den Prozess der strategischen Planung in Scrum. [Glo13, Vgl. S. 111f.] Am Anfang des Planungsprozesses steht die Erstellung der bereits in Kapitel 4.3 beschriebene Vision durch den Product Owner. Zweitens werden die Rahmenbedingungen (Constraints) festgelegt. Das sind Bedingungen,

die unbedingt bei der Produktentwicklung eingehalten werden müssen. Beispiele für eine solche Bedingung sind, dass ein bestimmter Nutzerkreis angesprochen werden soll, oder, dass ein Corporate Design eingehalten werden muss. Es sollten Constraints aus diesen Feldern Berücksichtigt werden: Wichtigste Features, verwendete Technologie, Usability (Kundengruppe), Design, rechtliche Anforderungen, Standards, Kosten und Budget sowie Lieferantenbeziehungen. [Glo13, Vgl. S. 123f.] Drittens werden die relevanten Personas identifiziert. Mit einer Persona ist eine echte Anwenderrolle gemeint, mit der gearbeitet wird (nicht nur eine User-Rolle). Sie sind am lebenden Menschen orientierte, imaginäre Personen, die deren Ziele, persönliche Eigenschaften, Erwartungen und Motivationen repräsentieren. Aus der Rolle Mechaniker wird dann beispielsweise Heinz Müller, 44 Jahre, Zweiradmechaniker, der in Augustfehn wohnt. Er hat zwei Kinder, spielt gerne in der Freizeit mit Ihnen Lego und liebt Metal-Festivals. Diese Vorgehensweise ist sinnvoll, damit die Usability-Bedürfnisse und der mögliche Nutzen des Produktes für die Persona verstanden werden. Viertens werden die User Stories vom Product Owner alleine oder mit dem Projektteam im Product Backlog (siehe auch Kapitel 4.3) mit Hilfe der identifizierten Personas zusammengetragen. [Glo13, Vgl. S. 125f.] Im fünften Schritt wird das Product Backlog vom Product Owner priorisiert. Einige mögliche Priorisierungsmethoden werden in Kapitel 3.3 vorgestellt. [Glo13, Vgl. S. 134ff.] Im sechsten Schritt schätzt das Projetteam die Größe der einzelnen priorisierten Backlog Items. Schätzverfahren sind in Kapitel 3.4 nachzulesen. [Glo13, Vgl. S. 142ff.] Siebtens erfolgt die Bestimmung der Velocity. Mit Velocity wird die Kapazität eines Teams in einem gewissen Zeitraum bezeichnet. Sie gibt an, wie schnell ein Team im Sprint (siehe Kapitel 4.5) ist. Wenn das Projekt bereits begonnen hat, kann die Velocity anhand der bereits geleisteten Backlog Items gemessen werden. Steht das Team vor dem ersten Sprint, wird das Team befragt, wie viele Backlog Items es in dem Sprint liefern möchte. Die Summe der Storypoints jener Backlog Items bestimmt die Velocity. [Glo13, Vgl. S. 148f.] Nach Bestimmung der Velocity wird im achten Schritt der Releaseplan aufgestellt. Prinzipiell kann die Anzahl der nötigen Sprints ermittelt werden, indem die Gesamt-Storypoints durch die Velocity geteilt wird. Es gibt aber Einflüsse, die die Velocity eines Teams schwanken lassen: Schwankungen zu Beginn, Urlaube, Feiertage und Skalierungseffekte (bei der Vergrößerung des Teams sinkt die Velocity für drei Sprints). Teams schätzen im Verlauf immer besser. Es gilt also für den Release-Plan Sicherheitspuffer einzubauen. Diese sollen in Form von Puffer-Funktionalitäten (Funktionalitäten, die nicht wichtig sind) eingeplant werden. [Glo13, Vgl. S. 149f.] Das Ergebnis strategischen Planungsaktivitäten ist Input für die Planung der Sprints, die im folgenden Kapitel erläutert wird. Da während der Projektdurchführung deutlich wird, dass Planungsfehler- und Ungenauigkeiten stattfinden, ist die strategische Planung als sich wiederholender Pla-

nungskreislauf zu verstehen. Damit wird die Planung auf Veränderungen im Projekt angepasst. [Glo13, Vgl. S. 151]

4.5 Sprint

Der Sprint stellt die Durchführungsphase des Projekts dar. Er dauert typischerweise etwa vier Wochen. Für die erfolgreiche Durchführung eines Sprints müssen die beteiligten drei bedeutende Prinzipien einhalten: Erstens arbeitet das Team nach dem Pull-Prinzip. Das bedeutet, dass das Team die Anzahl der Backlog Items bestimmt, die es in einem Sprint bearbeiten möchte. Nur so kann der Product Owner erwarten, dass termingerecht ausgeliefert wird. Zweitens werden Meetings immer absolut pünktlich begonnen und beendet. So wird erzwungen, dass die Beteiligten konzentriert die Themen des Meetings bearbeiten. Sprints werden nicht verlängert. Drittens organisiert sich das Team selbst. Sie bestimmen ihre Arbeitsweise selbstverantwortlich mit dem Willen zu liefern (Stichwort Commitment - siehe auch Kapitel 4.2). [Glo13, Vgl. S. 155ff.] Jeder Sprint enthält eine Planungs-, eine Durchführungs-, eine Evaluierungs- und eine Verbesserungsphase. Die Phasen werden durch diese Meetings begleitet: Estimation Meeting, Sprint Planning I + II, Daily Scrum, Sprint Review und Sprint Retrospektive. [Glo13, Vgl. S. 155ff.] Zum Estimation Meeting erscheint der Product Owner mit einem geschätzten und priorisierten Product Backlog. In diesem Meeting wird der Dialog zwischen Product Owner und dem restlichen Team aufrechterhalten. Neue Backlog Items werden besprochen und ergänzt, bestehende Schätzungen für Backlog Items werden aktualisiert und oft werden Backlog Items in mehrere neue Items aufgebrochen. Dadurch entstehen weitere, neue User Stories. Die strategische Planung wird angepasst. Dieses Meeting sollte mindestens einmal, besser zweimal pro Sprint stattfinden. Gerade zu Beginn sollte es zweimal oder öfter durchgeführt werden. Die Dauer des Estimation Meetings darf nicht länger als 35 Minuten sein. Der Product Owner hat die Aufgabe dieses Meeting einzuberufen. [Glo13, Vgl. S. 157 ff.] Im Bereich der taktischen Planung findet das Sprint Planning statt. Das Sprint Planning gliedert sich in zwei Meetings. Das erste Meeting (Sprint Planning I) beschäftigt sich mit dem Briefing und der Analyse. Anwesend sind das Team, der ScrumMaster, der Product Owner, der User und der Manager. Der Product Owner stellt seine Vorstellungen von dem geplanten Sprint vor. Anschließend definieren alle beteiligten gemeinsam ein klares Ziel, wie die Funktionalitäten gestaltet werden. Es wird eine detaillierte Anforderungsanalyse vorgenommen. Ergebnis dieses Meetings ist das Selected Product Backlog, das das Goal für den Sprint definiert. Die Dauer beträgt etwa 60 Minuten pro Woche des Sprints. [Glo13, Vgl. S. 164f.] Im dem typischerweise nachfolgenden Sprint Planning 2 Meeting sind der Product Owner und der

Anwender nur noch in einer passiven Rolle anwesend. In dieser Sitzung legt das Team das Design, die Spezifikation und die Architektur des Produktes fest. Als Resultat weiß jetzt jeder, wie gemeinsam die gewählten Aufgaben bewältigt werden können. Es erfolgt keine Festlegung, wer genau sie durchführt. Am Ende des Meetings wird eine Liste von Aufgaben (Tasks) erstellt, die nicht größer als acht Stunden sind. Diese sollten optimalerweise zu dem Selected Product Backlog abgelegt werden. Die Dauer beträgt etwa 60 Minuten pro Woche des Sprints. [Glo13, Vgl. S. 166ff.] Beim Daily Scrum plant und koordiniert sich das Team für einen Tag. Die vier Fragestellungen, die jedes Teammitglied beantworten muss: Was habe ich seit dem letzten Meeting erreicht? Was will ich bis zum nächsten Meeting erreichen? Was steht mir dabei im Weg? Wie kann ich jemandem aus dem Team heute helfen, schneller fertig zu werden? Wird an diesem Punkt festgestellt, dass eine Aufgabe länger dauert als geplant, kann sie in kleinere Aufgaben unterteilt werden. Der ScrumMaster ist bei diesem Meeting anwesend, um Impediments aufzudecken. Die Dauer darf 15 Minuten nicht überschreiten. [Glo13, Vgl. S. 171ff.] Die tägliche Arbeit wird mit Hilfe eines Taskboards unterstützt. Es hilft dem Team seine Aufgaben anzuordnen und den Überblick zu behalten. Abbildung 3 beispielhaft ein Scrum Taskboard. [Glo13, Vgl. S. 167] Im Sprint Review werden die Ergebnisse in Form des Produkt-Inkrements (siehe Kapitel 4.3) vom Team vorgestellt. Das Produkt-Inkrement wird analysiert und Verbesserungspotentiale werden gesucht. Es wird auch deutlich, ob das Sprint Goal erreicht wurde. Wichtig dabei ist, dass alle Probleme offengelegt werden, damit in zukünftigen Sprints darauf reagiert werden kann. Sollten Qualitätsprobleme existieren, kann eine Definition of Done hinzugezogen werden (siehe Kapitel 4.3). Typischerweise werden alle beteiligten bis auf den Kunden eingeladen. Die Dauer des Meetings beträgt 90 Minuten. [Glo13, Vgl. S. 177ff.] An der Sprint-Retroperspektive nehmen das Team, der ScrumMaster und der Manager teil. Alle Teammitglieder reden offen über Verbesserungspotentiale für den gemeinsamen Arbeitsprozess, um ihn effektiver und effizienter zu gestalten. Resultat dieses Meetings sind zwei Impediment-Listen. Eine Liste mit Impediments, die das Team lösen kann und eine Liste der Impediments, die der ScrumMaster lösen muss. Diese Impediments müssen jetzt behoben werden. Dauer etwa 90 Minuten. [Glo13, Vgl. S. 181]

4.6 Reporting

Scrum-Reporting findet primär am Ende eines Sprints statt. Als Darstellungsmethode wird das Burn-Down-Chart empfohlen. Ein Burn-Down-Chart stellt den noch verbleibenden Entwicklungsaufwand im Sprint in Stunden dar. Wenn beispielsweise der geschätzte Gesamtaufwand 900 Stunden beträgt, zeigt der Burn-Down-Chart am ersten Tag 900

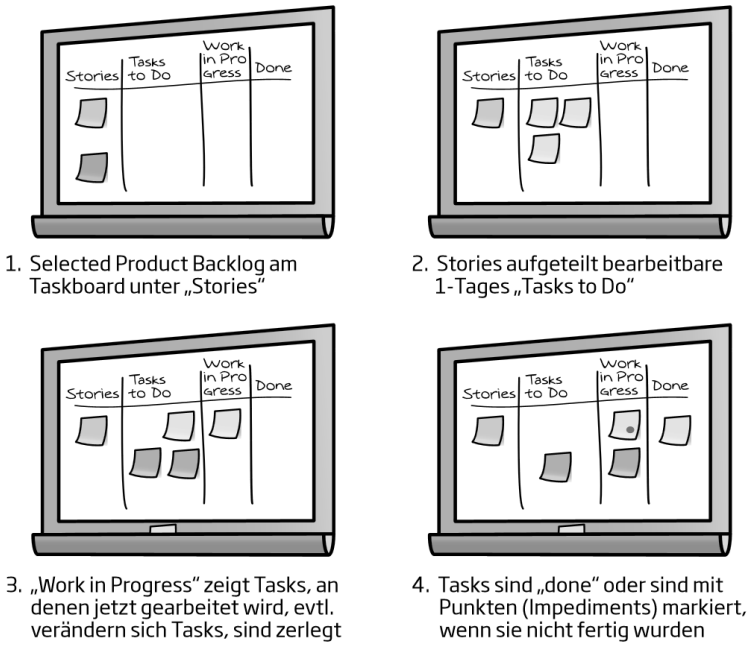


Abbildung 3: Scrum Taskboard[Glo13, S. 167]

Stunden an. Pro Tag wird die Restarbeitszeit neu geschätzt und eingetragen. Im Laufe der Zeit ist eine Trendlinie wie in Abbildung 4 zu erkennen. [Glo13, Vgl. S. 210]

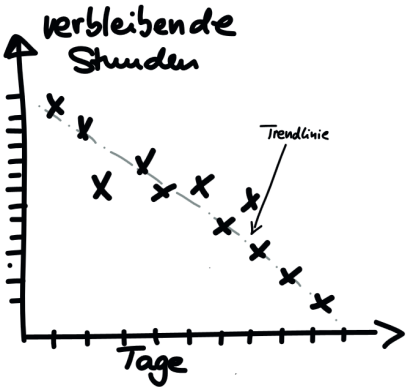


Abbildung 4: Burndown Chart [Glo13, S. 210]

Beim Burn-Down-Chart ist zu kritisieren, dass teilweise erledigte Tasks in die Ermittlung der geleisteten Stunden einfließen. Dem wird Einhalt geboten, indem statt des noch offenen Aufwandes die noch offenen Tasks gezählt werden. Möglicherweise erhöht sich im Laufe der Zeit die Anzahl der offenen Tasks. In dem Moment wird aus einem Burn-Down-

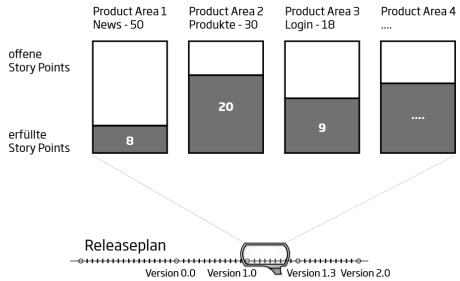


Abbildung 5: Parking-Lot-Chart [Glo13, S. 214]

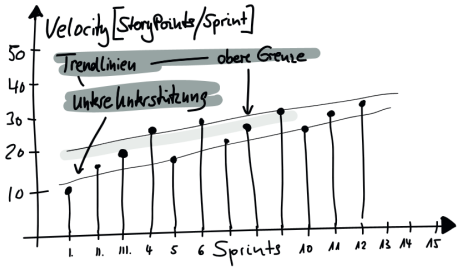


Abbildung 6: Velocity-Chart [Glo13, S. 214]

Chart ein Burn-Up-Chart. [Glo13, S. 210f.] In einem Parking-Lot-Chart (siehe Abbildung 5) werden die Backlog Items nach ihrer Kategorie gezählt und aufgezeigt, wie viele abgearbeitet sind. Damit wird dargestellt, wie sich die Applikation in ihren verschiedenen Funktionsbereichen entwickelt. [Glo13, Vgl. S. 213] Das Velocity-Chart (siehe Abbildung 6) stellt die Entwicklung der Velocity des Teams (siehe Kapitel 4.4) im Laufe der durchgeführten Sprints dar. [Glo13, Vgl. S. 214] Neben den Charts ist es auch sinnvoll ein Logbuch zu führen, das beispielsweise diese Angaben enthält: Geplante Abwesenheiten, wichtige Events, Teilnahmen an Trainings, Ergebnisse aus Reviews und Retrospektiven, Impediments und Entscheidungen. Auch persönliche Notizen können eingebracht werden können. [Glo13, Vgl. S. 215]

5 Software Kanban

Als weiteres agiles und schlankes Vorgehensmodell wird Software Kanban vorgestellt. Zuerst erfolgt eine Einleitung und anschließend werden Messgrößen für Kanban definiert. Im dritten Unterkapitel erfolgt die Vorstellung von den Rollen und empfohlenen Techniken in Verbindung mit Kanban. Zuletzt wird das einzige wirkliche Artefakt - das Kanban Board - vorgestellt.

5.1 Einleitung

Wie Scrum ist Software Kanban ein Vorgehensmodell der schlanken und agilen Softwareentwicklung (Lean Software Development), das von dem Toyota Production System inspiriert wurde. Das Vorgehensmodell wurde erstmals 2004 von Dragos Dumitriu bei Microsoft eingesetzt und 2007 von David J. Anderson der Öffentlichkeit vorgestellt. Kanban hat die Eigenart, dass es durch das Team individuell gestaltbar ist. Diese vier Elemente charaktarisieren Kanban: Erstens wird Arbeit genommen, nicht gegeben (Pull-Prinzip). Zweitens werden Mengen limitiert. Drittens sind Informationen zu veröffentlichen. Viertens sind Arbeitsabläufe kontinuierlich zu verbessern. [Epp11, Vgl. S. 1ff] David J. Anderson betont, dass der generierte Geschäftswert aus Sicht des Auftraggebers an erster Stelle steht. Nur wenn das Ende der Wertschöpfungskette erreicht ist, wird Geschäftswert generiert. An zweiter Stelle liegt der gleichmäßige Arbeitsfortschritt. Damit wird erreicht, dass der Projektverlauf vorhersagbar wird. An letzter stelle steht die Elimination von *Waste*. Im Kern werden die Eigenschaften von Lean Software Development, die in Kapitel 2.4 beschrieben werden, umgesetzt. Im Vergleich zu Scrum ist Kanban deutlich weniger in Form von Rollen, Meetings oder Artefakten reguliert. [Epp11, Vgl. S. 25ff.]

5.2 Messgrößen

Diese Messgrößen haben sich bei Kanban etabliert: $WorkInProgess$ (WIP) bezeichnet die Anzahl der Tasks, die zeitgleich in Arbeit sind. $CycleTime$ misst die Dauer zwischen Beginn und Ende der Arbeit an einem Task. $AverageCompletionRate$ beschreibt die durchschnittliche Anzahl fertiggestellter Tasks innerhalb eines festgelegten Zeitraums. Bei den Messgrößen besteht ein Zusammenhang, der auch Little's Law genannt wird: $CycleTime = \frac{WorkInProgess}{AverageCompletionRate}$ [Epp11, Vgl. S. 28ff.]

5.3 Techniken

Software Kanban sieht nur exemplarisch unterstützende Techniken für die klassischen Rollen Requirements Engineering, Entwicklung, Qualitätssicherung und Projektmanagement vor. [Epp11, Vgl. S. 85] Im Requirements Engineering werden User Stories (siehe Kapitel 3.2) und die allgemeinen Schätzmethoden (siehe Kapitel 3.4) empfohlen. Für die Entwicklung können Code Reviews (siehe Kapitel 3.5) hinzugezogen werden. Die Qualitätssicherung profitiert von Testautomatisierung (siehe Kapitel 3.5) und das Projektmanagement von den in Kapitel 5.2 definierten Messgrößen. [Epp11, Vgl. S. 89ff.] Tägliche Stand-Ups dienen dazu, dass diese drei Fragen von jedem Teilnehmer beantwortet werden: Was habe ich seit dem letzten Stand-Up getan? Was werde ich bis zum nächsten Stand-Up tun? Was

behindert mich derzeit bei meiner aktuellen Arbeit? Die Dauer eines solchen Meetings darf maximal 15 Minuten betragen. [Epp11, Vgl. S. 123] Bei einer Retrospektive treffen sich alle Personen eines Teams. Das Ziel dieses Meetings ist die Verbesserung der Organisation durch Identifikation von Waste in der Wertschöpfungskette. [Epp11, Vgl. S. 125]

5.4 Kanban Board (Artefakt)

Das einzige wirkliche Artefakt in Kanban ist das Kanban Board. Das Board visualisiert die Phasen der Wertschöpfungskette und die limitierten Mengen der abzuarbeitenden Aufgaben (Tasks) in transparenter Art- und Weise. Die Anforderungen werden in Form von Signalkarten (siehe Abbildung 7) visualisiert. Die Wertschöpfungskette kann beispielsweise in diese Phasen aufgeteilt werden: 1. Bereit zur Umsetzung, 2. Wird umgesetzt, 3. Bereit zur Qualitätssicherung, 4. Wird qualitätsgesichert und 5. Umgesetzt und qualitätsgesichert. Jeder Phase wird eine limitierte Menge (Anzahl der maximal zu bearbeitenden Aufgaben) zugewiesen. Abbildung 8 zeigt ein Kanban Board.[Epp11, S. 115ff.]

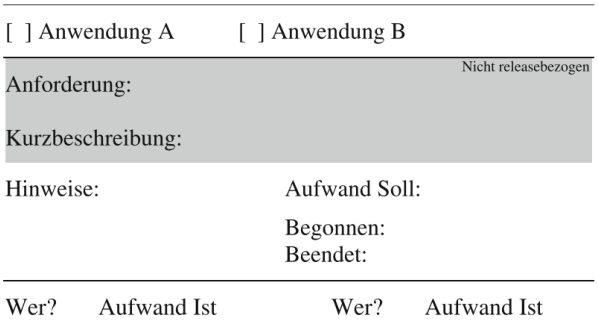


Abbildung 7: Beispiel Kanban Signalkarte [Epp11, S. 117]

6 Auswahl

Bei dem direkten Vergleich der beiden Vorgehensmodelle Scrum und Kanban wird deutlich, dass Scrum mehr Vorgaben zum Prozess macht. Scrum enthält eigene Rollen, Artefakte, definierte Planungsläufe auf strategischer und taktischer Ebene und eine Vielzahl von Meetings, die schnell Routine in den Ablauf bringen. Software Kanban hingegen ist unpräziser definiert. Dafür bietet Software Kanban mehr Flexibilität und Anpassungsfähigkeit. Die Mitglieder der Projektgruppe weisen sehr unterschiedliche Erfahrungswerte im Umgang von Projektarbeiten im Bereich der Softwareentwicklung auf. Das bedeutet, dass Flexibilität auch dazu führen kann, dass zu wenig Organisation zur Unproduktivität führt. Um

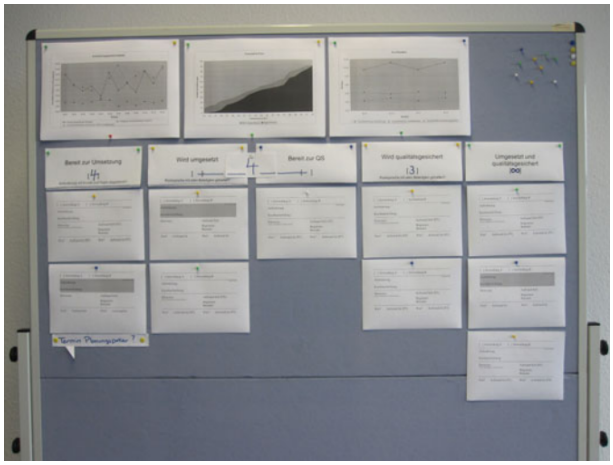


Abbildung 8: Beispiel Kanban Board [Epp11, S. 121]

von einem klar definierten Prozessmodell zu profitieren, fällt die Auswahl auf Scrum.

7 Anwendung von Scrum

Es folgt ein Vorschlag zur Anwendung von Scrum in der Projektgruppe. Zuerst stellt sich die Frage der Rollenverteilung. Die Aufgabe des Entwicklungsteams wird vom Projektgruppenteam übernommen. Die Rolle der Product Owner können die Projektgruppen-Mitglieder oder die Betreuer belegen. Die Betreuer sind tendenziell besser dazu geeignet, da sie spezifisches Business-Know-How für die Fachdomäne der Projektaufgabe aufweisen. Die Rolle des Customers übernimmt ein potentieller Kunde der eXin AG. Dadurch ergibt sich auch die Besetzung der Rolle des Users. ScrumMaster können Projektteam-Mitglieder oder Betreuer werden. Wenn ein Projektgruppenmitglied diese Rolle übernimmt ist abzuwägen, ob es diese als einzige Aufgabe wahrnimmt oder auch parallel eine andere Aufgabe wahrgenommen werden soll. Wenn beispielsweise der Teamleiter parallel die Aufgabe des ScrumMasters übernimmt, läuft er Gefahr seinen alten Führungsstil beizubehalten und die Rolle falsch wahrzunehmen. Wenn ein Software-Entwickler auch zum ScrumMaster wird, leidet die Produktivität darunter. In dem Fall muss die Rolle des ScrumMasters rotieren. [Glo13, Vgl. S. 99] Sollte ein Betreuer ScrumMaster werden, führt dies zu einer sehr intensiven Zusammenarbeit mit dem Projektteam. Für das Projektteam bedeutet es die größtmögliche Produktivität, jedoch ist fraglich, ob damit der Aufgabenbereich des Betreuers überschritten wird. Jetzt werden die Teamgrößen ermittelt. Ein Scrum-Team sollte aus maximal 7 Personen bestehen. [Glo13, Vgl. S.97] Die Projektgruppe besteht

aus 11 Mitgliedern. Angenommen, dass der ScrumMaster aus dem Projektgruppenteam stammt und jedes Team einen Product Owner seitens der Betreuer gestellt bekommt, sind diese Konstellationen möglich: Entweder drei Teams bestehend aus 4-5 Mitgliedern oder zwei Teams bestehend aus 6-7 Mitgliedern. Bei den genannten Teamgrößen sind ein Product Owner und ein ScrumMaster enthalten. Die Rolle des ScrumMasters wird vom Entwickler als rotierende Teilzeit-Rolle wahrgenommen werden. Bei mehreren Teams ist Kommunikaton unverzichtbar. Vision und Sprint Goal sind teamübergreifend zu verstehen. Deshalb werden drei weitere Meetings eingeführt: Scrum of Scrums, Product Owner Daily Scrum und Weekly Scrum der ScrumMaster Group. Beim Scrum of Scrums treffen sich Teammitglieder der unterschiedlichen Teams zur Besprechung von Abhängigkeiten (mindestens aus jedem Team eines). Beim Product Owner Daily Scrum koordinieren sich die Product Owner. Bei der ScrumMaster Group synchronisieren sich die ScrumMaster, damit die Qualität des Scrum-Prozesses gleichbleibend sichergestellt wird. [Glo13, S. 238ff.] Da das Team räumlich verteilt arbeitet, lautet die Empfehlung die Artefakte digital zu verwalten. Denkbar wäre die Nutzung des bisher eingeführten Projektmanagementsystems Atlassian - Jira. [Glo13, Vgl. S. 308] Grundsätzlich sollten alle Meetings vor Ort in Oldenburg durchgeführt werden. Da es jedoch aus organisatorischen Gründen nur schwer möglich ist ein Daily Scrum auf die Weise durchzuführen, wird empfohlen, es via Video- oder Audio-Konferenz stattfinden zu lassen. Die Timebox (Start und Dauer des Meetings) muss dabei strikt eingehalten werden. [Glo13, Vgl. S. 175] Jetzt wird die Sprint-Länge bestimmt. Ein Vorschlag: Für einen schnellen Lernprozess wird anfangs die Sprint-Länge auf 1-2 Wochen festgelegt. Auch sollten die ersten selektierten Backlog Items von Größe und Anzahl überschaubar sein. Dadurch kann schnelles Feedback direkt berücksichtigt werden. Anschließend können bei Bedarf die Turnaround-Zeiten sowie der Umfang pro Sprint je nach Velocity erhöht werden. Unerfahrene Entwickler sollten zuerst nur kleinere Product Backlog Items Bearbeiten (entsprechend Pull-Prinzip). Offenheit ist für eine gute Zusammenarbeit wichtig, Impediments müssen direkt angesprochen werden.

Literatur

[AG11] SAP AG. http://de.news-sap.com/2011/06/21/sap-hana-ist-ab-sofort-fur-kunden-weltweit-verfugbar/#more-1553, Zugriff Juni 2014, SAP HANA ist ab sofort für Kunden weltweit verfügbar, 2011.

[Bol09] Dietrich Boles. http://www-is.informatik.uni-oldenburg.de/~dibo/teaching/pg_fb10/Leitfaden-PG.pdf, leitfaden zur durchführung von projektgruppen, 2009.

[ea01] Beck K et al. http://agilemanifesto.org/iso/de/, Zugriff Mai 2014, Manifest für agile Softwareentwicklung, 2001.

[Epp11] Thomas Epping. *Kanban für die Softwareentwicklung*. Springer, Köln, 2011.

[Fow06] Martin Fowler. http://martinfowler.com/articles/continuousIntegration.htm, Zugriff Juni 2014, Continuous Integration, 2006.

[Fow13] Martin Fowler. http://martinfowler.com/bliki/ContinuousDelivery.html, Zugriff Juni 2014, Continuous Delivery, 2013.

[Glo13] Boris Gloger. *Scrum - Produkte zuverlässig und schnell entwickeln*. Hanser, Wien, 2013.

[Gro12] Standish Group. *CHAOS MANIFESTO 2012*. Standish Group, Boston, 2012.

[Han10] Eckhart Hanserb. *Agile Prozesse: Von XP über Scrum bis MAP*. Springer, Lörrach, 2010.

[Max13] Dominik Maximini. *Scrum - Einführung in der Unternehmenspraxis*. Springer Gabler, Berlin, 2013.

[WR08] Thomas Fittkau Walter Ruf. *Ganzheitliches IT-Projektmanagement - Wissen, Praxis, Anwendungen*. Oldenbourg, München, 2008.

Abschließende Erklärung

Ich versichere hiermit, dass ich meine Seminararbeit Klassische vs. agile Softwareentwicklung: Einsatz von SCRUM in der Projektgruppe selbständig und ohne fremde Hilfe angefertigt habe, und dass ich alle von anderen Autoren wörtlich übernommenen Stellen wie auch die sich an die Gedankengänge anderer Autoren eng anlegenden Ausführungen meiner Arbeit besonders gekennzeichnet und die Quellen zitiert habe.

Oldenburg, den 28. Juli 2014 Benjamin Hemken