

BY: KENNY LI
STUDENTNR: 1833731

CALLCENTER

CALL VOLUME FORECASTING

MET ARMA(P,Q)
EN HOLT-WINTERSMODELLEN

BWI-WERKSTUK
VRIJE UNIVERSITEIT TE AMSTERDAM
FACULTEIT DER EXACTE
WETENSCHAPPEN
BEDRIJFSWISKUNDE & INFORMATICA
DE BOELELAAN 1081A
1081 HV AMSTERDAM
DECEMBER 2011

INHOUDSOPGAVE

Inhoudsopgave	
Voorwoord	- 1 -
1. Inleiding	- 2 -
2. Voorbereiden van data	- 4 -
2.1 Omgaan met feestdagen	- 5 -
2.2 Omgaan met het aantal operationele dagen	- 6 -
3. Prestatievergelijking	- 7 -
4. Het ARMA(p,q) Model	- 9 -
4.1 Het Auto Regressive proces met lag-operator p [AR(p)]	- 10 -
4.2 Het Moving Average proces met lag-operator q [MA(q)]	- 10 -
4.3 Het Auto Regressive and Moving Average model [ARMA(p, q)]	- 11 -
4.4 Omgaan met trend en seizoenseffecten in het ARMA model	- 11 -
4.4.1 <i>De Small-TrendMethode</i>	- 12 -
4.4.2 <i>Differencing</i>	- 12 -
4.5 Model Selectie	- 13 -
5. Het ES-Model	- 16 -
5.1 Single Exponential Smoothing	- 16 -
5.2 Holt's linear method	- 17 -
5.3 Holt-Winters method	- 18 -
6. Het Onderzoek	- 20 -
6.1 De Data en de voorbereiding	- 20 -
6.3 Uitwerking	- 22 -
6.3.1 <i>ARMA Model</i>	- 22 -
6.3.2 <i>Holt-Winters Model</i>	- 25 -
6.3.3 <i>Een Aangepast regressiemodel</i>	- 27 -
6.4 Resultaat	- 30 -
7. Conclusie en Aanbevelingen	- 33 -
8. Literatuurlijst	- 37 -
9. Bijlage	- 39 -

VOORWOORD

Het BWI werkstuk is een belangrijk onderdeel in de masteropleiding Business Analytics. Het is een kans om je opgedane kennis tijdens de studie toe te passen in de praktijk. Voor dit werkstuk heb ik een onderdeel gekozen uit het onderwerp “Optimization of Business Processes”, de Callcenter Planning.

Voor ik aan de studie begon heb ik lange tijd gewerkt voor een multinational. Het bedrijf is actief in het produceren van consumentenelektronica, die onder een bekende handelsnaam wordt verkocht over de hele wereld. Voor de Europese markt heeft het bedrijf een gecentraliseerd callcenter. Deze wordt uitbesteed aan een bedrijf dat gevestigd is in Barcelona. De centralisatie was nodig om een zo goed mogelijke after sales service te bieden tegen zo laag mogelijke kosten.

Als Customer Information Center (CIC) Data Analyst was ik verantwoordelijk voor de monitoring van het callcenter. Aan het eind van elk maand worden de kosten en de KPI's (Key Performance Indicator) aan de bedrijfsleiding gerapporteerd. De KPI's reflecteren indirect het serviceniveau van een bedrijf. Om de KPI's op niveau te houden is een correcte call-volumevoorspelling heel belangrijk. In het begin werd de maandelijkse call-volumevoorspelling gedaan door het callcenter. Maar omdat ze slechte prestaties leverden, heb ik de taak overgenomen. Met dit werkstuk wil ik daarom een uitleg geven van de bestaande methoden om een goede forecast te maken.

Bij voorspellingen maken in een callcenter zijn de volgende modellen het meest gebruikt, namelijk de ARMA-modellen en de Holt-Winters-modellen. Ter vergelijking van deze modellen wordt een aangepast regressiemodel uitgelegd dat ik destijds heb gebruikt.

Ten slotte wil ik mijn begeleider, prof. Ger Koole, bedanken voor zijn begeleiding bij het schrijven van dit werkstuk.

1. INLEIDING

In de moderne economie betekent een verkocht product nog niet het einde van de productieketen. De after sales service van een bedrijf speelt tegenwoordig een hele belangrijke rol. Consumenten nemen steeds vaker de after sales service als selectie criterium bij de aanschaf van een nieuw product.

Voor de after sales service, is de klantenservice (en voornamelijk het callcenter) het belangrijkste middel. Het is echter vaak voordeliger om een callcenter uit te besteden dan deze intern te beginnen. Een professionele callcenter verschaft daarnaast veel deskundigheid en bovendien wordt er veel efficiënter gewerkt. De vraag naar callcenters die hoge kwaliteit leveren is hierdoor drastische toegenomen.

De doelstelling van een bedrijf is heel simpel. Elk bedrijf wil een zo hoog mogelijke klanttevredenheid, oftewel een hoog serviceniveau, tegen zo laag mogelijke kosten. Een hoog serviceniveau betekent in de branche een hoge answer rate en een korte wachttijd. Het hoge serviceniveau kan alleen worden behaald als er voldoende call agents aanwezig zijn. Maar met te veel call agents zullen de kosten niet meer te overzien zijn. Daarom wordt er geprobeerd om met de beschikbare data, een voorspelling te doen die zo dicht mogelijk bij de werkelijkheid zit. Hiermee kan het juiste aantal call agents worden ingepland.

In de meeste callcenters worden de binnenkomende gesprekken (inbound calls) minstens een maand van te voren voorspeld. Maar het voorspellen gaat lang niet even goed bij alle callcenters. Er zijn namelijk heel veel factoren die het aankomstproces beïnvloeden. Voorbeelden zijn nieuwe marketingcampagnes of de lancering van nieuwe producten. Vaak oefenen deze effecten onbekende invloeden uit op het aantal calls, wat kan leiden tot foutieve voorspellingen.

Voor wie weleens met callcenterdata heeft gewerkt, weet dat het binnenkomen van calls op elk moment van een dag en week varieert. Bijvoorbeeld aan het begin van de week komen aanzienlijk meer calls binnen dan aan het einde van de week (de zogenaamde Inter-day-correlatie, die uitgebreid werd onderzocht door Brown et al. [7]). In de ochtend komen meer calls binnen dan in de middag (de Intra-day-correlatie, zie Avramidis, Deslauriers & L'Ecuyer [2] die een Dirichlet verdeling hiervoor heeft gebruikt). Kortom, het aankomstproces van calls is afhankelijk van het moment. (zie ook Tanir & Booth [19] en Brown et al. [7])

In het verleden werd het aankomstproces vaak verondersteld als een homogeen Poissonproces te zijn waarbij de aankomst-rate constant is. In de bekende Erlang-formule voor callcenters wordt van deze veronderstelling uitgegaan. Recente onderzoeken hebben echter aangetoond (G. Jongbloed & G. Koole [13]) dat de aankomst-rate niet constant is. Daarom is een homogeen aankomst-proces niet adequaat. Om die reden werd het Poisson-Mixture-model geïntroduceerd. In plaats van een constante aankomst-rate in het traditionele Poissonproces, is nu het aankomstproces zelf ook een stochastisch proces.

Het is daarom een grote uitdaging om de juiste voorspelling te maken. Twee bekende modellen die hiervoor geschikt zijn, zijn de Auto Regressive and Moving Average (ARMA) modellen en de Exponential Smoothing (ES) modellen. In dit werkstuk worden deze twee modellen uitgewerkt aan de hand van een real-life dataset.

Het rapport is vooral bedoeld voor callcenter-data-analisten die weinig ervaring hebben met voorspellen. Door een omschrijving met voorbeelden en uitwerkingen te geven, wordt geprobeerd om meer inzicht te geven in het gebruik van wiskundige modellen in de praktijk.

In *Hoofdstuk 2* wordt eerst uitgelegd hoe de dataset voorbereid moet worden en welke veel voorkomende oneffenheden eruit gefilterd moeten worden. Daarna worden in *Hoofdstuk 3* verschillende statistieken uitgelegd die de prestatie van een model kan meten.

In *Hoofdstuk 4* en *5* wordt de theorie van het ARMA-model en het ES-model uitvoerig behandeld.

In *Hoofdstuk 5* wordt een dataset uitgewerkt met de verschillende modellen. De dataset komt van een Nederlandse instantie waarvan de achtergrond minder belangrijk is. S. G. Makridakis [16] benadrukt in zijn boek dat de meest simpele modellen in de praktijk vaak beter presteren dan de ingewikkelde modellen. Maar ook in recente onderzoeken van Taylor [20], waarbij verschillende tijdreeksmodellen werden gebruikt, o.a. de ARMA-modellen en de ES-modellen, is aangetoond dat het meest simpele modellen de ingewikkelde modellen overtreffen. Daarom wordt verder in het hoofdstuk nog een ander model uitgewerkt. Het is een regressie model dat ik voorheen op het werk heb toegepast. Aan het eind van het hoofdstuk worden de resultaten vergeleken.

Tot slot wordt er in *Hoofdstuk 6* een conclusie getrokken en worden er aanbevelingen gegeven.

2. VOORBEREIDEN VAN DATA

Voordat een dataset daadwerkelijk gebruikt kan worden, moet deze eerst worden gecontroleerd op fouten en uitschieters. Bij een fout moet de aard van die fout worden onderzocht, hoe de fout is ontstaan en hoe ermee omgegaan moet worden. Bij een uitschieter moet er bijvoorbeeld onderzocht worden of deze kunstmatig eruit gelaten kan worden, en of deze vervangen moet worden door een meer waarschijnlijke waarde. Men moet er wel op letten dat een uitschieter vaak ook een fout is.

Naast de controle op fouten en uitschieters, zijn er in Nederland vaak speciale feestdagen waarop een callcenter gesloten is. De feestdag kan bijvoorbeeld op een maandag vallen (de dag in een week waarop de meeste calls worden verwacht). Maar de feestdag kan ook op een zondag vallen (de meeste callcenters zijn op zondag gesloten, en het heeft in dit geval weinig invloed op de voorspelling). Op welke weekdag de feestdag valt verschilt van jaar tot jaar. Het aantal calls dat binnenkomt kan hierdoor sterk beïnvloed worden.

Naast de feestdagen verschilt het aantal operationele dagen in een maand ook aanzienlijk. Met andere woorden, het aantal calls van een bepaalde maand kan niet vergeleken worden met een maand eerder, of dezelfde maand van het jaar daarvoor. Om deze redenen moeten de data eerst genormaliseerd worden, voordat het model erop losgelaten kan worden.

Het aantal calls kan ook beïnvloed worden door marketingcampagnes, de lancering van nieuwe producten, het weer of andere gebeurtenissen. Dit zijn de specifieke effecten die moeilijk meetbaar zijn en steeds een andere aanpak vereist. Daarom worden ze hier verder buiten beschouwing gelaten.

Twee belangrijke effecten die gemakkelijk meetbaar zijn, zijn de trend en seasonality (seizoensinvloeden). Deze komen vaak in tijdreeksen voor. Daarom bestaan er verschillende methoden om ze te verwerken. De aanpak is per model verschillend. Deze worden apart in *Hoofdstuk 4* en *5* besproken. Trend en seasonality kunnen gemakkelijk worden geconstateerd met de decompositie van de tijdreeks. Het resultaat van een dergelijke decompositie laat de trend en het seizoenseffect in een aparte figuur zien (zie als voorbeeld *Figuur 9.1* in de bijlage).

2.1 OMGAAN MET FEESTDAGEN

Stel dat de volgende data beschikbaar zijn van een bepaalde maand. Op de eerste rij is de weekdag aangegeven. Op de tweede rij wordt de datum aangegeven. En op de derde rij het aantal ontvangen calls op die dag. Maandag de 17^e is in dit voorbeeld een feestdag:

Weekdag	Za	Zo	Ma	Di	Wo	Do	Vr	Za	Zo	Ma	Di	Wo	Do	Vr
Dag	1	2	3	4	5	6	7	8	9	10	11	12	13	14
#calls	88	0	1061	1061	1015	919	937	75	0	1139	1141	1197	1151	938

Za	Zo	Ma	Di	Wo	Do	Vr	Za	Zo	Ma	Di	Wo	Do	Vr	Za	Zo
15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
67	0	0	1147	1046	1009	845	71	0	1080	993	859	863	874	64	0

Het totaal aantal binnenkomende gesprekken van de maand is daarmee gelijk aan 19.640 calls. Om een feestdag te normaliseren, wordt de feestdag vervangen door het gemiddelde van dezelfde weekdag, uit dezelfde maand:

Maandag 3: 1.061

Maandag 10: 1.139

Maandag 24: 1.080

Gemiddelde van deze 3 dagen = $(1.061+1.139+1.080) / 3 = 1.093$

Het concept van deze methode is: als deze dag geen feestdag was geweest dan zouden er naar verwachting 1.093 calls binnen zijn gekomen. En dan zou het maandtotaal 20.733 calls zijn geweest.

2.2 OMGAAN MET HET AANTAL OPERATIONELE DAGEN

Naast de feestdagen zijn er ook manieren bedacht om het aantal operationele dagen in een maand te normaliseren. Omdat elk maand minstens vier volle weken bevat, wordt 4 weken = 28 dagen als een standaard maand genomen.

In *Tabel 2.2* zijn de datapunten zodanig gerangschikt, dat de som van het aantal calls van elk weekday in een maand gemakkelijk te berekenen is.

Ma	Di	Wo	Do	Vr	Za	Zo
					88	0
1061	1061	1015	919	937	75	0
1139	1141	1197	1151	938	67	0
1093	1147	1046	1009	845	71	0
1080	993	859	863	874	64	0
4373	4342	4117	3942	3594	365	0

In deze maand waren er vier maandagen met een totaal van 4.373 calls. Er waren ook vier dinsdagen, woensdagen, donderdagen en vrijdagen, met een totaal van respectievelijk 4.342, 4.117, 3.942 en 3.594 calls. En als laatste waren er ook 5 zaterdagen geweest met een totaal van 365 calls. Om deze te normaliseren tot het standaard aantal dagen in een maand, wordt één zaterdag eruit gelaten. Dit wordt gedaan met de volgende berekening: $\frac{4}{5} \times 365 = 292$. Het genormaliseerde aantal calls van deze maand is daarmee gelijk aan $4.373 + 4.342 + 4.117 + 3.942 + 3.594 + 292 = 20.660$ calls, ten opzichte van het oorspronkelijke aantal van 20.773 calls.

3. PRESTATIEVERGELIJKING

In dit hoofdstuk wordt aandacht besteed aan de prestatie maatstaven van de modellen, oftewel de nauwkeurigheid van de voorspelling van een model. In de meeste gevallen geldt: een goed model is een model waarbij de voorspelling die gemaakt is, weinig verschilt van de werkelijkheid. Met andere woorden, het model met de laagste (totale) voorspellingsfout verdient de voorkeur. De totale voorspellingsfout is gelijk aan: $\sum_{t=1}^n \varepsilon_t = \sum_{t=1}^n X_t - F_t$, waarbij X_t gelijk aan de observatie op tijdstip t en F_t gelijk aan de voorspelling voor tijdstip t .

Door de diversiteit van data zijn verschillende methoden ontstaan in de literatuur [10], [17]. Onderaan is een lijst te vinden van de meest gebruikte statistieken bij het ARMA-model en het ES-model:

Mean Error:
$$ME = \frac{1}{n} \sum_{t=1}^n \varepsilon_t \quad (3.1)$$

Mean Absolute Error:
$$MAE = \frac{1}{n} \sum_{t=1}^n |\varepsilon_t| \quad (3.2)$$

Mean Squared Error:
$$MSE = \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 \quad (3.3)$$

Root Mean Squared Error:
$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n \varepsilon_t^2} \quad (3.4)$$

Mean Percentage Error:
$$MPE = \frac{1}{n} \sum_{t=1}^n \frac{\varepsilon_t}{X_t} * 100 \quad (3.5)$$

Mean Absolute Percentage Error:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{\varepsilon_t}{X_t} * 100 \right| \quad (3.6)$$

Theil's U-Statistic:

$$U = \sqrt{\frac{\sum_{t=1}^{n-1} (FPE_{t+1} - APE_{t+1})^2}{\sum_{t=1}^{n-1} (APE_{t+1})^2}} \quad (3.7)$$

Met:

Forecast Relative Change:

$$FPE_{t+1} = \frac{F_{t+1} - X_t}{X_t} \quad (3.8)$$

Actual Relative Change:

$$APE_{t+1} = \frac{X_{t+1} - X_t}{X_t} \quad (3.9)$$

$$AIC = 2k - 2\ln(L)$$

Akaike Information Criterion:

$$k = \text{het aantal parameters in het model} \quad (3.10)$$

$L = \text{De optimale waarde van de}$
 $\text{likelihood functie van het model}$

Voor de bovenstaande statistieken gelden ook: de modellen waarvan alle statistieken klein zijn of dicht bij nul zijn verdienen de voorkeur. Maar vaak worden MSE en MAPE als de belangrijkste statistieken beschouwd. In de praktijk hangt de keuze vaak samen met de situatie en de preferentie van de data analyst.

Met de Akaike Informatie Criterion (*Formule 3.10*) wordt vaak de “goodness of fits” van een statistisch model gemeten (Zie ook M. C. M. de Gunst [10]). Doorgaans wordt de AIC als een bijproduct bij de uitvoer van een softwarepakket weergegeven. Handmatig berekenen van de AIC is vaak lastig. En ook hier geldt de regel “lager is beter”.

4. HET ARMA(P,Q) MODEL

Het ARMA(p, q) model is geïmplementeerd door G. E. P. Box en G. M. Jenkins in 1970 (zie ook B. H. Andrews & S. M. Cunningham [1]). Zij hebben de autocorrelatieprocessen in tijdreeksen onderzocht. Met autocorrelatieprocessen worden processen bedoeld waarbij de realisatie in verschillende perioden onderling is gecorreleerd. Met andere woorden: de waarde van een fout is statistisch onafhankelijk van de voorgaande of latere waarden. Dit proces wordt ook een “witte-ruisproces” genoemd.

De studie van G. E. P. Box en G. M. Jenkins was vooral gericht op twee vormen van autocorrelatie, het Auto Regressive proces met order p (AR(p)) en het Moving Average proces met order q (MA(q)). Het ARMA(p, q) model is ontstaan door deze twee processen te combineren.

Een belangrijk concept bij het ARMA-model is dat de invoer uit stationaire tijdsreeksen moet bestaan. S. G. Makridakis et al. [17] beschrijft stationariteit als een horizontaal patroon, waarbij de datapunten fluctueren rond een constant gemiddelde, dat over de tijd heen niet toeneemt of afneemt. Een formele definitie van stationariteit kan worden teruggevonden in het boek van G. E. P. Box, G. M. Jenkins en G. C. Reinsel [3].

Om een tijdreeks stationair te maken moeten alle effecten die in *Hoofdstuk 2* zijn besproken eruit gefilterd worden. Een stationaire reeks mag dus ook geen trend- en seizoensinvloeden bevatten. De resulterende reeks is pas stationair als deze voldoet aan de Dickey-Fuller test.

Met de Dickey-Fuller-test wordt de correlatie tussen X_t en X_{t-1} getest. De absolute waarde van de correlatie tussen X_t en X_{t-1} mag niet gelijk zijn aan 1. Indien de correlatie gelijk is aan 1, hangt de waarde van nu volledig af van alle waarnemingen uit het verleden. Deze reeks kan daarom nooit stationair zijn. In wiskundige termen: de H_0 hypothese: $\rho = 1$, moet worden verworpen wil deze reeks stationair zijn. Dit is ook bekend onder de term “unit root”. Meer informatie is te vinden in D. A. Dickey & W. A. Fuller [9].

4.1 HET AUTO REGRESSIVE PROCES MET LAG-OPERATOR P [AR(P)]

Het AR(p) model is een regressiemodel waarbij realisatie uit een eerder tijdstip als invoer wordt gebruikt. Uitgedrukt in een formule:

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + \varepsilon_t \quad (4.1)$$

met $t = p+1, p+2, \dots, p+n$. X_{t-k} is de observatie op tijdstip $t-k$, met $k = 1, \dots, p$. $\alpha_1, \dots, \alpha_p$ zijn de onbekende parameters. De parameters moeten aan bepaalde voorwaarden voldoen. Dit is nodig zodat de waarde van α niet oneindig groot kan worden. Zo moet bij $p = 1$ gelden dat $-1 < \alpha_1 < 1$.

Verder geeft ε_t de fout aan. Het is een witte-ruisproces, met de verwachting dat $E[\varepsilon_t X_{t-k}] = 0$ voor alle $k \geq 1$, met variantie σ^2 . En de parameter c is een constante. Dit model gebruikt observaties uit eerdere tijdstippen als invoer om de realisatie van vandaag te berekenen. Zodra de huidige waarde bekend is, kan ook de toekomstige waarde worden berekend.

4.2 HET MOVING AVERAGE PROCES MET LAG-OPERATOR Q [MA(Q)]

In dit model hangt realisatie X_t geheel af van de fouten ε_{t-k} . Het model heeft de vorm:

$$X_t = c + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q} \quad (4.2)$$

Waarbij X_t de observatie op tijdstip t is. ε_{t-k} is de fout op tijdstip $t-k$. Dit is een witte-ruisproces, met $k = 1, \dots, q$. β_1, \dots, β_q de onbekende parameters. Net zoals bij het autoregressiemodel, moeten de parameters ook aan bepaalde eisen voldoen. Zie het boek van S. G. Makridakis et al. [17], p340, voor een uitgebreide beschrijving.

Verder is het MA(q) proces altijd een stationair proces, met:

$$\text{Gemiddelde } \mu = E[X_t] = c \quad (4.3)$$

$$\text{Variantie } \gamma_0 = \sigma^2 (1 + \sum_{j=1}^q \theta_j^2) \quad (4.4)$$

$$\text{Covariantie } \gamma_k = \sigma^2 (\theta_k + \sum_{j=k+1}^q \theta_j \theta_{j-k}) \quad (4.5)$$

voor $k \leq q$ en $X_k = 0$ voor $k > q$.

4.3 HET AUTO REGRESSIVE AND MOVING AVERAGE MODEL [ARMA(P, Q)]

Het ARMA(p, q) model is een combinatie van de modellen die hiervoor beschreven zijn, dit is weergegeven als:

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q} \quad (4.6)$$

Merk op dat het ARMA(p, 0) model hetzelfde is als het AR(p) model. Neem q = 0 voor de termen die afhangen van q in *Formule 4.6*. Het restant is dan precies het AR(p) model. Om dezelfde reden, geldt ook dat het ARMA(0, q) model gelijk is aan het MA(q) model.

4.4 OMGAAN MET TREND EN SEIZOENSEFFECTEN IN HET ARMA MODEL

De meeste tijdreeksen bestaan vaak uit niet-stationaire componenten zoals trend en seasonality. Dergelijke effecten moeten daarom eerst gefilterd worden voordat deze gebruikt kunnen worden. Twee methoden die vaak worden gebruikt zijn (zie P. J. Brockwell & R. A. Davis [5] en M. C. M. de Gunst [10]):

- 1) De small-trendmethode
- 2) Differencing

Differencing wordt vaak gecombineerd met het ARMA(p, q) model. Samen vormen deze het Autoregressive Integrated Moving Average model met order (p, d, q). Oftewel het ARIMA(p, d, q) model met lag operator d voor differencing ($d \in \mathbb{R}$).

4.4.1 DE SMALL-TRENDMETHODE

Bij de small-trendmethode, worden de stationaire component als volg berekend:

$$\hat{Y}_{j,k} = X_{j,k} - \hat{m}_j - \hat{s}_k \text{ en } j, k \in \mathbb{R} \quad (4.7)$$

Waarbij $\hat{Y}_{j,k}$ de stationaire component, \hat{m}_j de trend component, en \hat{s}_k de seasonality component. Uitgaande van een periode van 1 jaar, is j het aantal beschikbare data in jaren, en k het aantal maanden in een jaar. Voor de parameter k worden ook het aantal kwartalen of weken in een jaar gebruikt. De keuze hangt geheel af van de dataset en de data analist. Zo geeft $X_{2,6}$ de observatie uit juni van jaar 2 uit de data, indien k het aantal maanden voorstelt.

\hat{m}_j en \hat{s}_k kan worden geschat met het volgende:

$$\hat{m}_j = \frac{1}{12} \sum_{k=1}^{12} X_{j,k}, j = 1, \dots, \text{aantal jaren} \quad (4.8)$$

$$\hat{s}_k = \frac{1}{j} \sum_{j=1}^j (X_{j,k} - \hat{m}_j), k = 1, \dots, 12 \quad (4.9)$$

Uit de formule is te zien dat \hat{m}_j het jaar gemiddelde is. Terwijl \hat{s}_k het maand gemiddelde is, verminderd met de trend.

4.4.2 DIFFERENCING

Bij differencing wordt het verschil tussen 2 observaties in de tijd genomen. De lag-operator d geeft de tijd aan waarmee het verschil genomen moet worden. In formule vorm:

$$X'_t = X_t - X_{t-d} \quad (4.10)$$

X'_t is de stationaire reeks die gevraagd wordt. X_t en X_{t-d} zijn realisaties uit tijdstip t en $t-d$.

In de financiële wereld wordt bijvoorbeeld vaak gebruik gemaakt van het “random-walkmodel”, waarbij $X_t - X_{t-1} = \varepsilon_t$, met $d = 1$.

Voor seizoensinvloeden wordt het ARMA(p, q) vaak uitgebreid tot het ARIMA(p, d, q)(P, D, Q)_s model. In dit model geeft de parameters (p, d, q) (in kleine letters) het non-seasonal deel aan. Terwijl het seasonal deel met de parameters (P, D, Q)_s (in hoofdletters) wordt aangegeven. Hierbij is s gelijk is aan het aantal perioden in een seizoen. Voor jaarlijkse seizoensfluctuaties wordt vaak s = 12 gebruikt.

Een uitwerking van het ARIMA(1, 1, 1)(1, 1, 1)₄ model, is in het boek van S. G. Makridakis et al. [17], p346, te vinden. Zodra de parameters geschat zijn, kan de formule meteen gebruikt worden voor de voorspelling.

4.5 MODEL SELECTIE

Bij voorspellingen met het ARMA model, wordt vaak gebruikt gemaakt van de ACF en PACF plot. ACF staat voor autocorrelatiefunctie. Deze wordt gegeven door:

$$\rho_x(h) = Cor(X_{t+h}, X_t) = \frac{\gamma_x(h)}{\gamma_x(0)} \quad (3411)$$

Waarbij $\gamma_x(h)$ de autocovariantiefunctie is met lag h, en is gegeven door:

$$\begin{aligned} \gamma_x(h) &= \gamma_x(h, 0) = Cov(X_{t+h}, X_t) \\ &= E[(X_{t+h} - EX_{t+h})(X_t - EX_t)] \end{aligned} \quad (4.12)$$

De algemene formule voor PACF (Partiële Autocorrelatie functie) is:

$$\alpha(1,1) = \gamma_x(1), \text{ voor } k = 1 \quad (4.13)$$

$$\alpha(k, k) = \frac{\gamma_x(k) - \sum_{j=1}^{k-1} \alpha(k-1, j) \gamma_x(k-j)}{1 - \sum_{j=1}^{k-1} \alpha(k-1, j) \gamma_x(j)} \quad (4.14)$$

, voor $k > 1$

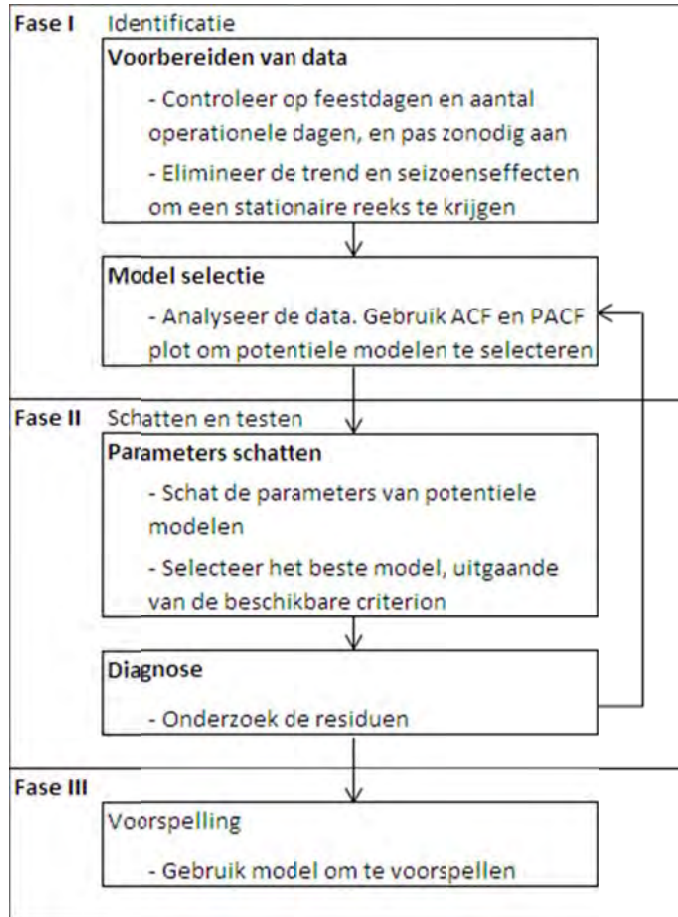
De ACF en PACF plots worden vaak gebruikt om de parameters in het ARMA model te herkennen. Hieronder is een beschrijving van de verschillende modellen van de plot:

Proces	ACF	PACF
AR(1)	Exponentieel dalend aan de positieve kant als $\alpha_1 > 0$ Exponentieel dalend maar alterneert tussen de positieve en negatieve kant, en beginnend aan de negatieve kant, dan $\alpha_1 < 0$	Piek op lag 1 en daarna 0. De piek is positief als $\alpha_1 > 0$ en negatief als $\alpha_1 < 0$
AR(p)	Naarmate de lag groter wordt, daalt de pieken Exponentieel of als een sinus functie. Het exacte patroon hangt uiteindelijk af van de waarde van $\alpha_1, \dots, \alpha_p$	Pieken lopen van lag 1 tot lag p, daarna heeft de functie een waarde van 0
MA(1)	Piek op lag 1 en daarna 0. De piek is positief als $\beta_1 < 0$, en negatief als $\beta_1 > 0$	Exponentieel dalend aan de positieve kant als $\beta_1 > 0$ Exponentieel dalend maar alterneert tussen de positieve en negatieve kant, en beginnend aan de negatieve kant, dan $\beta_1 < 0$
MA(q)	Pieken lopen van lag 1 tot lag q, daarna heeft de functie een waarde van 0	Naarmate de lag groter wordt, daalt de pieken Exponentieel of als een sinus functie. Het exacte patroon hangt uiteindelijk af van de waarde van β_1, \dots, β_q

Nadat de data zijn genormaliseerd, de parameters zijn geschat en het model is gekozen, moeten de residuen onderzocht worden. De test die aangeeft of er bij de residuen inderdaad sprake is van een witte-ruisproces, heet de Portmanteau-test. In deze test wordt de correlatie coëfficiënten tussen elk residu getoetst. Als de teststatistieken significantie vertonen, dan wordt er aangenomen dat de correlatiecoëfficiënten niet gelijk zijn aan nul. Zie ook G. E. P. Box & D. A. Pierce [4] en G. M. Ljung & G. E. P. Box [14].

Indien het bij de residuen niet om een witte ruis proces gaat, betekent dat er mogelijk nog andere onbekende, maar relevante factoren spelen. In dat geval moet een ander model worden gekozen. Pas als alles voldoet, kan een goed voorspelling worden gemaakt.

Globaal kunnen we het voorspellingsproces in het volgende schema uitdrukken:



Figuur 4.2 Schematisch representatie van de Box-Jenkins methode voor tijdreeksen

5. HET ES-MODEL

Het ES-model werd in de late jaren vijftig geïntroduceerd door C. C. Holt [11], R. G. Brown [6], en J. F. Magee [15]. Het is niet zeker wie van de drie het model als eerste gebruikt. Maar het model is zonder twijfels één van het meest gebruikte voorspellingsmodellen, mede door de gemakkelijke uitvoering, maar ook door de lage implementatiekosten.

Bij het ES model spreekt men vaak over een uitbreiding van de moving-averagemethode. Bij voorspellingen met de moving-averagemethode, wordt het gemiddelde genomen over de laatste k waarnemingen. De wegingsfactor voor elke k is in dit geval gelijk aan $1/k$. Bij exponential smoothing worden de wegingsfactoren op een speciale manier berekend. Vaak heeft de meest recente waarnemingen veel meer invloed op de toekomstige waarden dan de waarnemingen uit het verleden. Daarom krijgen de meest recente waarnemingen zwaardere wegingsfactoren.

Er bestaat een aantal varianten van het model: het Single Exponential Smoothing (SES) model, het Holt's lineair model en het Holt-Winters trend en seasonality model.

5.1 SINGLE EXPONENTIAL SMOOTHING

Het Single Exponential Smoothing model (SES) is de simpelste vorm uit de serie. De methode is vooral geschikt voor data zonder enige trend- of seizoenseffecten.

Om een voorspelling maken, nemen we het gewogen gemiddelde van de voorgaande waarnemingen:

$$\begin{aligned} F_{t+1} &= \hat{X}_{t+1} = w_0 X_t + w_1 X_{t-1} + w_2 X_{t-2} + \dots \\ &= \sum_{i=0}^{\infty} w_i X_{t-i} \end{aligned} \tag{5.1}$$

Met F_{t+1} de voorspelling voor $t + 1$. En w_i het gewicht voor de i^e waarneming. En de som van w_i , waar $i = 0, \dots, \infty$, is gelijk aan 1. Het model suggereert dat de waarnemingen die verder in het verleden zijn gedaan minder belangrijk zijn. Dit wordt opgelost met een wegingsfactor die afneemt, naarmate i groter wordt:

$$w_i = \alpha(1 - \alpha)^i \tag{5.2}$$

In *Formule 5.2* is α een constante tussen 0 en 1. Als *Formule 5.1* en *Formule 5.2* gecombineerd wordt, ontstaat het volgende:

$$\begin{aligned} F_{t+1} &= \alpha X_t + \alpha(1-\alpha)^1 X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \dots \\ &= \alpha X_t + (1-\alpha)[\alpha X_{t-1} + \alpha(1-\alpha)X_{t-2} + \dots] \end{aligned} \quad (5.3)$$

Hier is te zien dat de opsomming tussen de rechthoekige haakjes gelijk is aan F_t . Vandaar:

Forecast:
$$F_{t+1} = \alpha X_t + (1-\alpha)F_t \quad (5.4)$$

Formule (5.4) is anders te schrijven als:

$$F_{t+1} = F_t + \alpha(X_t - F_t) \quad (5.5)$$

Hier is te zien dat de voorspelling op tijdstip $t+1$ gelijk is aan de voorspelling op tijdstip t , plus een correctie op de fout uit de voorspelling op tijdstip t . Om een voorspelling te maken moet alleen nog α worden gekozen. Meestal wordt α zodanig gekozen dat MSE geminimaliseerd is. Maar in principe kan elk statistiek besproken in *Hoofdstuk 3* als criterium dienen. De keuze ligt bij de data analist.

5.2 HOLT'S LINEAR METHOD

Een uitbreiding op de Single Exponential Smoothing methode is de Holt's Linear Methode. Met dit model wordt een voorspelling met trendeffecten mogelijk gemaakt.

Anders dan bij SES zijn hier twee smoothing constanten, α en β die beide tussen 0 en 1 zitten met drie vergelijkingen in plaats van één:

Forecast:
$$F_{t+m} = L_t + mb_t \quad (5.6)$$

Level:
$$L_t = \alpha X_t + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (5.7)$$

Trend:
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (5.8)$$

Formule (5.6) voor de voorspelling op $t+m$ is een lineaire functie, waarbij $m=1,2,\dots$ *Formule (5.7)* L_t is een schatting van het niveau op de waarnemingen op tijdstip t . Deze is een gewogen gemiddelde tussen X_t en de voorspelling van de afgelopen periode ($L_{t-1} + b_{t-1} = F_{t-1}$). *Formule (5.8)* is een schatting van de trend op tijdstip t . Het is het

gewogen gemiddelde tussen de vroegere trend (b_{t-1}) en de trend van de laatste periode ($L_t - L_{t-1}$). Het model werkt echter alleen bij een datareeks waar geen seasonality aanwezig is.

5.3 HOLT-WINTERS METHOD

De hiervoor besproken ES modellen zijn bruikbaar zolang er geen seasonality aanwezig zijn. Daarom heeft P. R. Winters [21] het lineaire model uitgebreid zodat ook seasonality gemeten kan worden.

In het Holt's Linear Model wordt een extra component toegevoegd die een schatting is van de seizoenseffecten. S. G. Makridakis et al. [17] maakt onderscheid tussen twee verschillende methoden, de multiplicatieve methode en de additieve methode. De keuze hangt echter af van de karakteristieken van de beschikbare data. Maar over het algemeen komt de additieve methode minder vaak voor.

Het multiplicatieve model is als volgt gespecificeerd:

$$\text{Forecast:} \quad F_{t+m} = (L_t + mb_t)S_{t-s+m} \quad (5.9)$$

$$\text{Niveau:} \quad L_t = \alpha \frac{X_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad (5.10)$$

$$\text{Trend:} \quad b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \quad (5.11)$$

$$\text{Seizoen:} \quad S_t = \gamma \frac{X_t}{L_t} + (1 - \gamma)S_{t-s} \quad (5.12)$$

Waarbij α , β en γ tussen 0 en 1 zitten. $m = 1, 2, \dots$ zijn de perioden die voorspeld moeten worden. En $s = 1, 2, \dots$ gelijk aan de lengte van een seizoensperiode. Bij een maandelijks voorspelling met een seizoen periode van een jaar is s gelijk aan 12.

Het additieve model omvat de volgende vergelijkingen:

$$\text{Forecast:} \quad F_{t+m} = L_t + mb_t + S_{t-s+m} \quad (5.13)$$

$$\text{Niveau:} \quad L_t = \alpha(X_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad (5.14)$$

$$\text{Trend:} \quad b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \quad (5.15)$$

$$\text{Seizoen:} \quad S_t = \gamma(X_t - L_t) + (1 - \gamma)S_{t-s} \quad (5.16)$$

De specificaties zijn verder gelijk aan het multiplicatieve model.

De verschillende ES-modellen hebben aparte manieren om de beginwaarden te initialiseren. Deze worden in *Hoofdstuk 6* behandeld.

Ook bij de exponential smoothing is het neutraliseren van feestdagen en het aantal operationele dagen zeer belangrijk. De in *Hoofdstuk 2* besproken manieren komen ook hier van pas.

6. HET ONDERZOEK

6.1 DE DATA EN DE VOORBEREIDING

De data die worden gebruikt voor het onderzoek zijn afkomstig van een Nederlands callcenter. Het betreft alle calls die ontvangen zijn tussen de periode januari 2011 – augustus 2011. De samengevatte dataset is terug te vinden in de bijlage (*Tabel 9.1*).

Het doel van dit onderzoek is een geschikt model vinden om voorspellingen te doen voor de komende perioden. Daarbij worden de ARMA-modellen, de ES-modellen en de regressiemodellen gebruikt. Gezien de beperkte grootte van de dataset, is een dagelijkse voorspelling de beste optie.

Om de prestatie van de modellen te kunnen meten wordt de dataset verdeeld in een trainingset en een evaluatieset. De trainingset bevat de data van de periode 1 januari 2011 – 31 juli 2011, en komt overeen met 212 datapunten. Deze wordt alleen gebruikt voor het schatten van de parameters. De evaluatieset bevat de data van de periode 1 augustus 2011 – 31 augustus 2011 (31 data punten), en wordt gebruikt om het model te evalueren.

De ruwe dataset bevat 286.570 datapunten. Elk datapunt representeert een call, met een specificatie op datum, tijd en status. De status geeft aan of een call beantwoord of onbeantwoord is. Een samenvatting van de maandelijkse inbound calls en de answer rate daarvan wordt in *Tabel 6.1* weergegeven:

Maand	Trainingset							Evaluatieset	Overall
	1	2	3	4	5	6	7	8	
Answered	31982	25939	27335	22590	24083	23901	21135	24641	201606
Abandoned	8361	10283	6385	10672	13318	14257	9390	12298	84964
Total	40343	36222	33720	33262	37401	38158	30525	36939	286570
Answer Rate	79.3%	71.6%	81.1%	67.9%	64.4%	62.6%	69.2%	66.7%	70.4%

Tabel 6.1

De meeste callcenters hanteren een answer rate van 95%. In *Tabel 6.1* is te zien dat de answer rate van dit callcenter langdurig onder niveau is. In de praktijk kan een lage answer rate direct leiden tot hogere inbound calls. Als een klant het drukke callcenter telefonisch probeert te bereiken, zijn hoogstwaarschijnlijk alle lijnen bezet. In de meeste gevallen probeert de klant het direct of op een later tijdstip opnieuw. Hiermee wordt extra traffic gecreëerd voor het callcenter. Het effect is veel duidelijker als het callcenter langdurig bezet is. Dit zijn de zogenaamde repeated calls. De data-analist zal er rekening mee moeten houden dat er geprobeerd wordt om het aantal potentiële klanten te

voorspellen dat contact wil opnemen met het callcenter, en niet wat er daadwerkelijk ontvangen of beantwoord is. Door gebrek aan gegevens zullen deze verder buiten beschouwing worden gelaten.

Het callcenter is op alle dagen operationeel, ook op alle feestdagen en in het weekeinde. In 2011 vielen de feestdagen op de volgende data:

- 01 januari 2011 (zaterdag) – Nieuwjaarsdag
- 22 april 2011 (vrijdag) – Goede Vrijdag
- 24 april 2011 (zondag) – 1^e paasdag
- 25 april 2011 (maandag) – 2^e paasdag
- 30 april 2011 (zaterdag) – Koninginnedag
- 05 mei 2011 (donderdag) – Bevrijdingsdag
- 02 juni 2011 (donderdag) – hemelvaartsdag
- 12 juni 2011 (zondag) – 1^e pinksterdag
- 13 juni 2011 (maandag) – 2^e pinksterdag

Zoals besproken in *Hoofdstuk 2*, zullen deze datapunten aangepast moeten worden. Een ander belangrijk punt is het vóór- en na-effect van een feestdag. Ook hier worden de datapunten aangepast. In totaal zijn 9 datapunten vervangen door het gemiddelde. Het resultaat is terug te vinden in de bijlage (*Tabel 9.2*).

Omdat de voorspelling op dagbasis wordt gemaakt, is het aantal operationele dagen binnen een maand minder belangrijk. Daarnaast worden de voorspellingen op zaterdag en zondag apart berekend. Dit is vooral omdat het aantal calls tussen een normale werkdag en in het weekeinde zo groot is, dat deze veel vertroebeling kan veroorzaken. Bovendien is het aantal te verwachten calls in het weekeinde beperkt. In de meeste gevallen zullen één á twee telefonisten voldoende zijn. In het vervolg worden de voorspellingen in het weekeinde genomen die gelijk zijn aan het gemiddelde van die weekdag.

In de bijlage is de decompositie van de data te vinden (*Figuur 9.1*). Dit laat duidelijke maandelijkse seizoenfluctuaties zien. In ons context worden de maandelijkse fluctuaties als trend genoemd. Naast de trend zijn er ook duidelijke wekelijkse seizoenseffecten aanwezig.

De constatering van de trend en seasonality worden nogmaals bevestigd door de ACF plot, PACF plot en het resultaat van de Dicky-Fuller-test. De ACF en PACF plot laat duidelijke terugkerende patronen zien op lag = 5. Terwijl de teststatistiek uit de Dickey-Fuller-test niet significant is bij een significantieniveau van 5%. Daarmee wordt aangetoond dat de reeks niet stationair is. De resultaten zijn terug te vinden in de bijlage.

6.3 UITWERKING

6.3.1 ARMA MODEL

Bij het ARMA model wordt het stappenschema (*Figuur 4.2*) gebruikt. Verder wordt het softwarepakket R gebruikt bij het onderzoek. De keuze is hierop gevallen vanwege de beschikbaarheid, de gebruiksvriendelijkheid en zijn populariteit bij verschillende universiteiten.

Het softwarepakket is gratis te downloaden via de homepage van R Project:

<http://cran.r-project.org>

Voor het onderzoek is gebruik gemaakt van versie 2.14.0. Naast de standaard functiepakketten zijn ook de uitbreidingspakketten “stat”, “quadprog”, “tseries”, “zoo”, “forecast”, “fracdiff”, “Rcpp” en “RcppArmadillo” in gebruik genomen. De uitbreidingspakketten zijn te downloaden via de R Project Homepage (zie ook de links in de bijlage [22]). In de bijlage zijn ook de codes te vinden die gebruikt zijn.

Stationariteit van de reeks en Model selectie

Uit de ACF plot in *Figuur 9.2* in de bijlage is te zien dat de significantie om de vijf dagen weer terugkeert. Dit is herkenbaar als een typisch seizoenseffect met een periode van vijf dagen. Het is ook logisch omdat een periode van vijf dagen precies een week voorstelt. Om de reeks stationair te maken, moet daarom de seasonal differencing worden toegepast met lag = 5. Het ARMA(p, q)(P, Q)₅ (of ARIMA(p, d, q)(P, D, Q)₅) model moet hiervoor een oplossing bieden.

In het softwarepakket R bestaat de functie `auto.arima()`. Op grond van de opgegeven selectiecriteria (meestal de minimale AIC), loopt de functie automatisch alle parameters langs om het beste model te vinden. Elk model wordt daarbij automatisch gevalideerd met de Dickey-Fuller-test, de Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test of de Phillips-Perron-test. Net als de Dickey-Fuller-test, toetst de KPSS-test en Phillips-Perron test de aanwezigheid van stationariteit in een reeks. Bij de uitvoeren horen ook de residuen en de schatting van de parameters. Meer informatie is te vinden in de documentatie van de “forecast”-package van R door R. J. Hyndman [12].

Het gemakkelijke model selectie en parameters schatting maakt het mogelijk om meerdere modellen met elkaar te vergelijken. Er is hier gekozen voor drie verschillende modellen.

ARMA-Model I

Dit is het ARIMA(2, 1, 2)(2, 0, 2)₅ model met drift en werd door de `auto.arima()`-functie gekozen als de beste oplossing (AIC als selectiecriteria en aanwezigheid van stationariteit wordt getoetst met de Dickey-Fuller-test). Het model heeft een AIC die gelijk is aan 2008.33. De Portmanteau-test heeft verder geen significante verschillen waargenomen op de residuen. Met andere woorden, de residuen is een witte ruis proces.

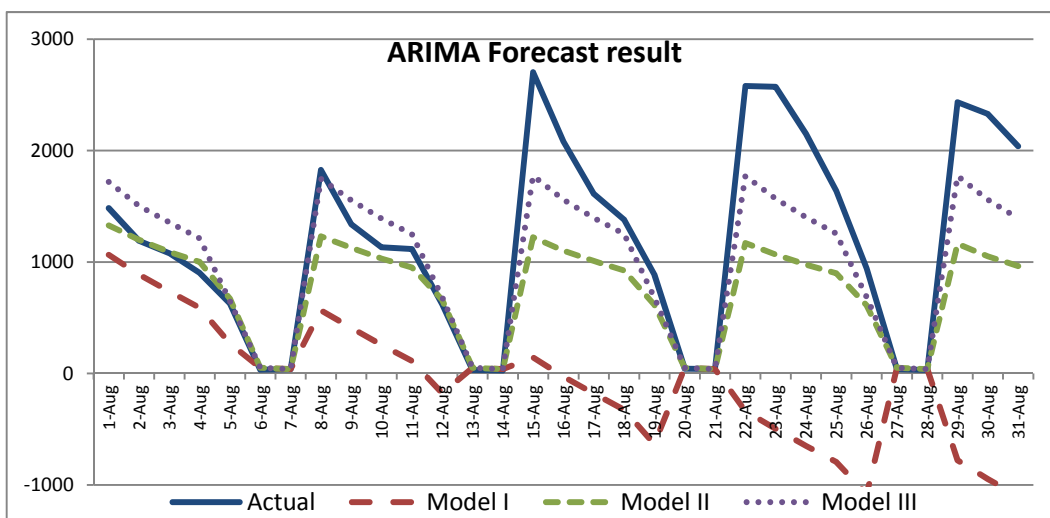
ARMA-Model II

Dit is het ARMA(3, 2)(2, 2)₅ model met een gemiddelde gelijk aan 0. (AIC als selectiecriteria en de default KPSS-test om te toetsen op aanwezigheid van stationariteit) Dit model heeft een AIC die gelijk is aan 2031.38. Het is theoretisch minder aantrekkelijk dan ARMA-Model I. De Portmanteau-test heeft verder geen significante verschillen waargenomen op de residuen.

ARMA-Model III

Dit is het ARMA(3, 2)(1, 2)₅ model met een gemiddelde dat ongelijk is aan 0. In dit model wordt vooraf handmatig de differencingmethode met lag = 5 toegepast. De reden hiervan is dat de data na de eerste keer differencing nog steeds significantie vertoont op lag = 5. De selectie van een model is gebaseerd op de AIC. Terwijl de aanwezigheid van stationariteit wordt getoetst met de Dickey-Fuller-test. Het model heeft verder een AIC die gelijk is aan 1968.41. De residuen zijn verder geslaagd voor de Portmanteau-test.

De voorspelde waarden van de drie modellen worden in *Figuur 6.1* uitgezet. Om het compleet te maken zijn ook de voorspelling in het weekeinde toegevoegd.



Figuur 6.1

De bijbehorende statistieken (gebaseerd op de data uit de evaluatie set) zijn in *Tabel 6.2* weergegeven.

Statistiek	ARMA-Model I	ARMA-Model II	ARMA-Model III
<i>ME</i>	1260.22	435.75	175.79
<i>MAE</i>	1265.38	454.15	299.01
<i>MSE</i>	2952764	482178	180370
<i>RMSE</i>	1718.36	694.39	424.70
<i>MPE</i>	68.38	13.57	-1.78
<i>MAPE</i>	84.24	31.13	24.85
<i>Theil's U</i>	1.04692	0.51089	0.29865
<i>AIC</i>	2008.33	2031.38	1968.41

Tabel 6.2

Het resultaat wordt verder in *Hoofdstuk 6.4* besproken.

6.3.2 HOLT-WINTERS MODEL

In *Hoofdstuk 6.1* was geconstateerd dat de trendeffecten en seizoenseffecten aanwezig zijn in de dataset. Daarom wordt het Holt-Winters model voor trend en seasonality gebruikt. Hier wordt gekozen voor het softwarepakket MS-Excel, omdat alle stappen precies te volgen zijn. Meer informatie over werken met de ES-modellen in MS-Excel is te vinden in het boek van Powell S. G. & Baker K. R. [18].

Initiële waarden

Bij de ES-modellen moeten de parameters α , β en γ een beginwaarde aannemen. Voor alle parameters geldt dat deze tussen 0 en 1 moeten zitten. Omdat deze anders ook waarden zoals oneindig kunnen aannemen. Voor het gemak wordt $\alpha = \beta = \gamma = 0,5$ genomen. Door de MSE of MAPE te minimaliseren (met bijvoorbeeld een Excel macro), kunnen de optimale waarden van de parameters worden gevonden.

Naast de parameters moeten ook startwaarden worden berekend. De startwaarden zijn nodig om het algoritme te initialiseren. Het niveau L_t wordt geïntialiseerd door het gemiddelde te nemen van het eerste seizoen. Merk op dat *Formule 6.11* precies de moving-average is met order s , waarmee de seizoenseffecten geëlimineerd kunnen worden.

$$L_s = \frac{1}{s}(Y_1 + Y_2 + \dots + Y_s) \quad (6.11)$$

Initiatie van de trend b_t gebeurt met het volgende:

$$\begin{aligned} b_s &= \frac{1}{s} \left(\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \dots + \frac{Y_{s+s} - Y_s}{s} \right) \\ &= ((Y_{s+1} + \dots + Y_{s+s}) - (Y_1 + \dots + Y_s)) \frac{1}{s^2} \end{aligned} \quad (6.12)$$

Seizoenseffecten S_t kan worden geïntialiseerd met het volgende:

$$S_1 = \frac{Y_1}{L_s}, S_2 = \frac{Y_2}{L_s}, \dots, S_s = \frac{Y_s}{L_s} \quad (6.13)$$

In *Figuur 9.3* uit de bijlage is een opzet van het model te zien. Vanaf $t = 6$, wordt *formule 5.9 – formule 5.12* uit *Hoofdstuk 5* toegepast.

Nadat alle waarden voor F_t , L_t , b_t en S_t uit de trainingset zijn berekend, kan er gekeken worden naar de optimale waarde voor de parameters α , β en γ . Dit kan

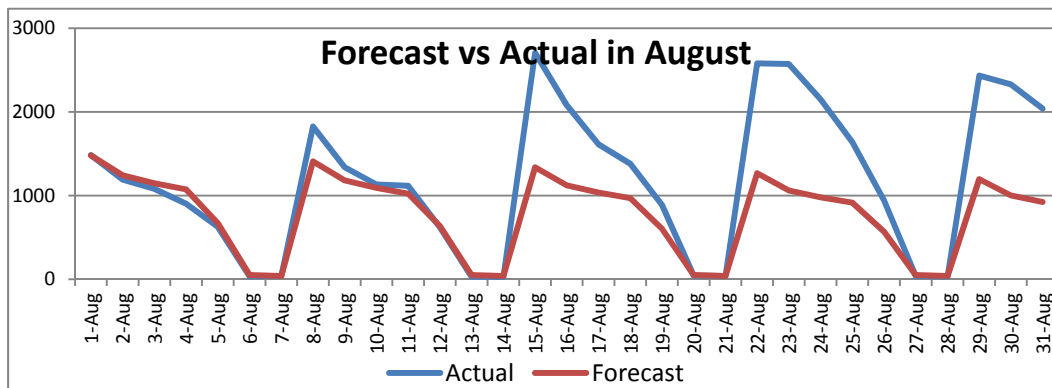
handmatig worden gedaan of met een macro in MS-Excel. Het idee is om de MSE te minimaliseren, rekening houdend met de voorwaarde: $0 \leq \alpha, \beta, \gamma \leq 1$.

En uiteindelijk, de parameters waarvoor geldt dat MSE geminimaliseerd is:

$$\begin{aligned}\alpha &= 0.312 \\ \beta &= 0.011 \\ \gamma &= 0.342\end{aligned}$$

In de praktijk kunnen data-analisten ook de parameters naar eigen inzicht kiezen. Bij het kiezen van de parameters moet er vooral gelet worden op overfitting. In het vervolg worden de parameters constant gehouden als hierboven.

De voorspelling (inclusief de weekeinden) die hieruit volgt is grafische weergegeven in *Figuur 6.2*. Het volledige resultaat is terug te vinden in de bijlage.



Figuur 6.2

De statistieken behorend bij de evaluatieset zijn als volgt:

Statistiek	HW-Model I
ME	408.05
MAE	435.77
MSE	457827
RMSE	676.63
MPE	12.16
MAPE	29.93
Theil's U	0.49

Tabel 6.3

Het resultaat wordt verder in *Hoofdstuk 6.4* besproken.

REG-Model B

In REG-Model B wordt eerst de logaritmen genomen over de data. In formule uitgedrukt:

$$\begin{aligned} \text{Logaritmen:} \quad \ln(X_t) = c + \alpha t + \gamma_1 \text{Maand} + \gamma_2 \text{Dag} & \quad (6.16) \\ & + \beta_1 \mathbb{I}_{(\text{Weekdag}=\text{Maandag})} + \dots \\ & + \beta_5 \mathbb{I}_{(\text{Weekdag}=\text{vrijdag})} \end{aligned}$$

REG-Model C

In REG-Model C wordt differencing met lag = 5 toegepast:

$$\begin{aligned} \text{Differencing:} \quad X_t - X_{t-5} = c + \alpha t + \gamma_1 \text{Maand} + \gamma_2 \text{Dag} & \quad (6.17) \\ & + \beta_1 \mathbb{I}_{(\text{Weekdag}=\text{Maandag})} + \dots \\ & + \beta_5 \mathbb{I}_{(\text{Weekdag}=\text{vrijdag})} \end{aligned}$$

De opzet van de data is te vinden in de bijlage (*Figuur 9.4*). Met de Regressiefunctie uit het Data-analyses tool-pack in MS-Excel kunnen de parameters worden geschat. Het resultaat is in *Tabel 6.4* weergegeven:

Coëfficiënten	REG-Model A	REG-Model B	REG-Model C
Intercept = c	7299.47	10.40	-1343.19
Maand = γ_1	-4685.33	-2.48	1326.48
Dag = γ_2	-162.18	-0.09	38.35
t = α	216.78	0.11	-61.71
Monday = β_1	0.00	0.00	0.00
Tuesday = β_2	-275.76	-0.14	-6.04
Wednesday = β_3	-520.76	-0.27	39.33
Thursday = β_4	-736.93	-0.39	75.26
Friday = β_5	-1384.10	-0.88	92.84

Tabel 6.4

Bij de uitvoer behoren ook de teststatistieken (Deze geven aan of een bepaalde variabele significant is. Variabelen die niet of weinig significant zijn, kunnen worden weggelaten. Zo worden alleen de belangrijkste variabelen behouden.

Een voorspelling met REG-Model C voor woensdag 3 augustus ($t = 153$) bijvoorbeeld, kan als volgt worden berekend:

$$\ln(F_{153}) = c + \alpha(153) + \gamma_1(8) + \gamma_2(3) + \beta_1 0 + \beta_2 0 + \beta_3 1 + \beta_4 0 + \beta_5 0$$

$$F_{153} = e^{c+153\alpha+8\gamma_1+3\gamma_2+\beta_3}$$

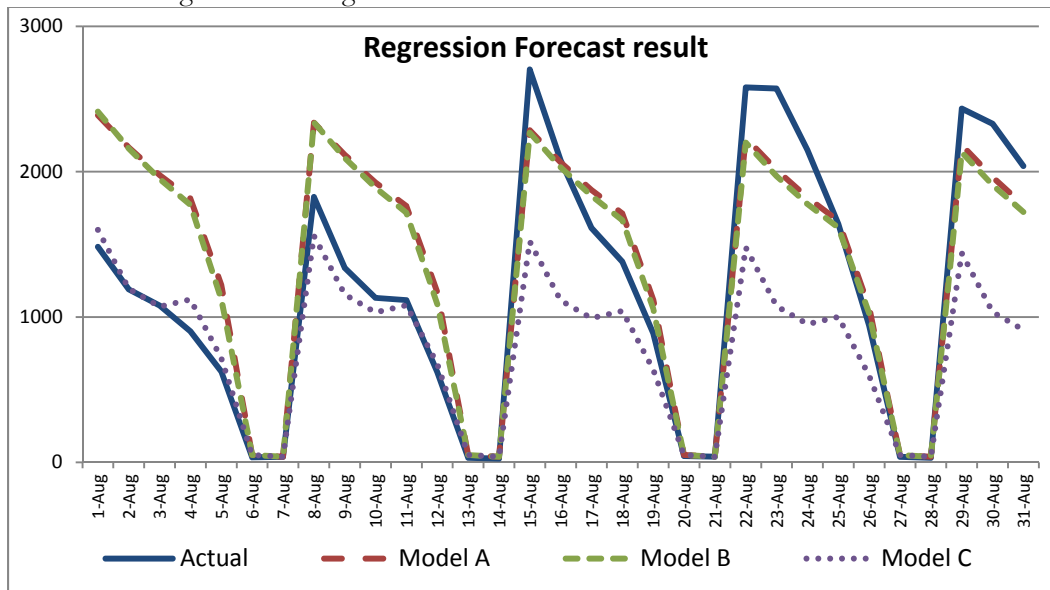
$$F_{153} = e^{10.40+153(0.11)-8(2.48)-3(0.09)-0.27} = 1949$$

Zodra de voorspellingen zijn gemaakt kunnen de statistieken van de drie modellen worden berekend (Zie Tabel 6.5):

Statistiek	Model A	Model B	Model C
ME	-195.78	-166.73	373.53
MAE	360.38	354.04	410.19
MSE	232782	223549	401224
RMSE	482.48	472.81	633.42
MPE	-30.53	-28.01	9.82
MAPE	37.36	35.84	29.30
Theil's U	0.23	0.24	0.40

Tabel 6.5

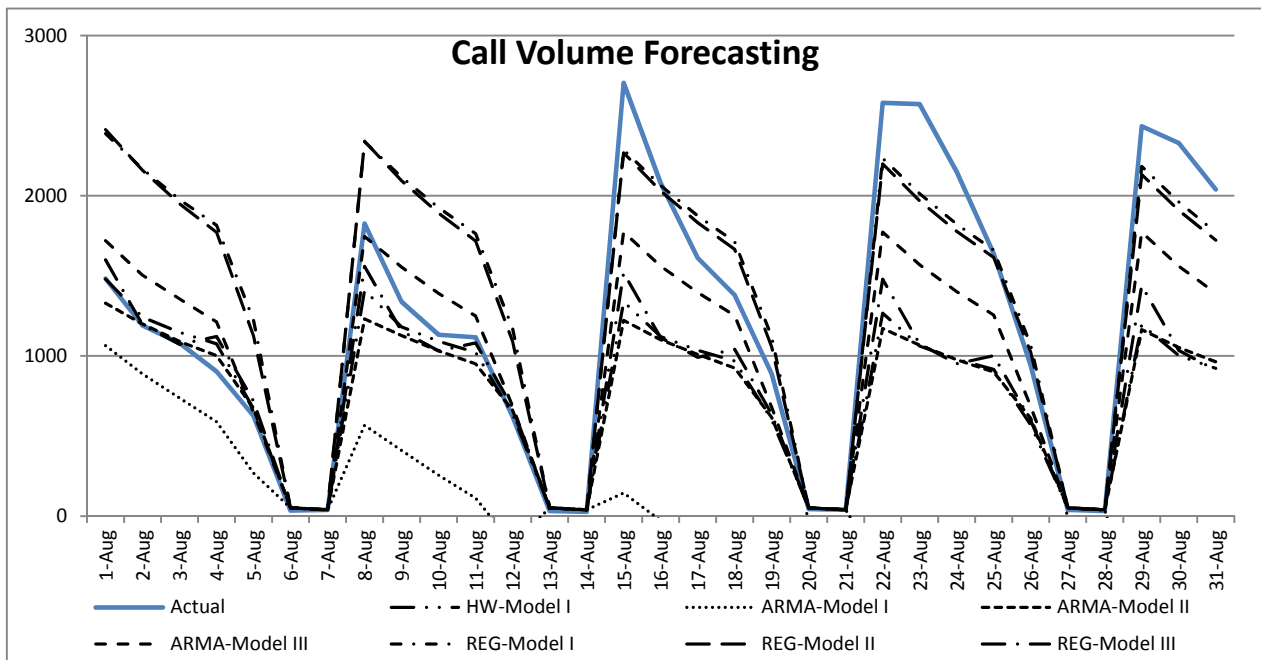
En de grafische weergave:



Figuur 6.3

6.4 RESULTAAT

Alle voorspellingen die gemaakt zijn kunnen worden teruggevonden in *Tabel 9.4* in de bijlage. De voorspellingen zijn ook in *Figuur 6.4* weergegeven. In de figuur zijn verder ook de werkelijke waarden en voorspellingen in het weekeinde opgenomen. De statistieken ter beoordeling van de prestatie zijn samengevat in *Tabel 6.6*. De resultaten zijn wederom gebaseerd op de evaluatie set.



Figuur 6.4

Actual Aug = 36939		ARMA			Regressie		
Statistiek	HW	Model I	Model II	Model III	Model A	Model B	Model C
Voorspelling Augustus	24293	-2128	23431	31489	43008	42108	25360
%Afwijking tov werkelijk	-34.2%	-105.8%	-36.6%	-14.8%	16.4%	14.0%	-31.3%
ME	408.05	1260.22	435.75	175.79	-195.78	-166.73	373.53
MAE	435.77	1265.38	454.15	299.01	360.38	354.04	410.19
MSE	457827	2952764	482178	180370	232782	223549	401224
RMSE	676.63	1718.36	694.39	424.7	482.48	472.81	633.42
MPE	12.16	68.38	13.57	-1.78	-30.53	-28.01	9.82
MAPE	29.93	84.24	31.13	24.85	37.36	35.84	29.3
Theil's U	0.49	1.05	0.51	0.3	0.23	0.24	0.4

Tabel 6.6

Uit *Tabel 6.6* is meteen te zien dat het $\text{ARMA}(3, 2)(1, 2)_5$ model (ARMA-Model III) het beste presteert. De afwijking van de voorspelling is 15% van de werkelijkheid. Het model heeft zeer gunstige MSE en MAPE tegenover andere modellen. De werkelijke answer rate van augustus was 66.7%. Het aantal potentiële bellers kan daarom veel lager zijn vanwege de repeated calls. De afwijking kan daardoor lager zijn. Maar kunnen we concluderen dat dit model het “beste” uit de reeks is? In *Figuur 6.4* valt meteen op dat de voorspelling van week 33 ruim 35% onder het werkelijk ontvangen calls zit. Hetzelfde komt voor bij de overige modellen. Is dit model dan alsnog het “beste”? Of presteren alle modellen in principe even goed?

Figuur 6.4 laat ook zien dat het $\text{ARIMA}(2, 1, 2)(2, 0, 2)_5$ model (ARMA-Model I) niet een geschikt model is. Naarmate er verder in de toekomst wordt gekeken, wijken de voorspellingen steeds verder af en werden zelfs negatief. Dit is echter niet mogelijk in deze context.

Het ARMA-Model II heeft een hogere AIC dan ARMA-Model I. En zal theoretisch slechter presteren. Het resultaat laat echter het tegenovergestelde zien. Een mogelijke verklaring hiervoor is dat AIC op de trainingset gebaseerd is, terwijl MSE en MAPE dat op de evaluatieset zijn. Dit is duidelijk een vorm van overfitting, waardoor het beeld ontstaat dat ARMA-Model I beter is. Dit is ook te zien in *Figuur 9.5* uit de bijlage.

De resultaten van regressiemodellen A en B verschillen zo weinig van elkaar dat men zich afvraagt of de transformatie met logaritme inderdaad nodig was. Hoewel de statistieken anders aangeven heeft dit model geen overeenkomstige punten met de werkelijke datapunten. Maar toch wijkt REG-Model B maar 14% van de werkelijkheid af. Oftewel: een model dat misschien geschikt is voor de lange termijn.

Tot nu toe werd alleen gekeken naar het resultaat op de lange termijn. Als er alleen gekeken wordt naar de eerste twee weken van augustus, dan valt het op dat er twee modellen zijn die zeer goed presteren, namelijk het Holt-Winters-model en het aangepaste regressiemodel met differencing (REG-Model C). De voorspellingen van deze modellen voor week 31 en 32 wijken minder dan 5% af van de werkelijkheid (zie ook *Tabel 6.7*). De modellen maken accurate voorspellingen voor de korte termijn. Juist door de onverwachte volumestijging in week 33 worden deze modellen slecht gewaardeerd.

De voorspelling van ARMA-Model II wijkt ook minder dan 10% af van de werkelijkheid. Het is daarom ook een geschikt model voor kortetermijnvoorspellingen.

<i>Actual = 11442</i>		ARMA			Regressie		
Statistiek	HW	Model I	Model II	Model III	Model A	Model B	Model C
Voorspelling Week 31 en 32	11135	4895	10473	13243	19060	18727	11389
%Afwijking tov werkelijk	97.32%	42.78%	91.53%	115.74%	166.58%	163.67%	99.53%
<i>ME</i>	21.93	467.65	69.24	-128.61	-544.14	-520.38	3.81
<i>MAE</i>	79.66	474.94	105.82	140.09	544.14	520.38	80.85
<i>MSE</i>	17804	390545	33844	33552	427435	396934	13590
<i>RMSE</i>	133.43	624.94	183.97	183.17	653.78	630.03	116.58
<i>MPE</i>	-11.42	31.71	-7.94	-23.55	-63.76	-60.63	-13.03
<i>MAPE</i>	18.10	55.59	19.72	24.18	63.76	60.63	18.88
<i>Theil's U</i>	0.17	0.56	0.26	0.11	0.47	0.48	0.12

Tabel 6.7

Tabel 6.7 bevestigt dat ARMA-Model III niet het beste model is. Het model presteert goed op de lange termijn, maar op de korte termijn zijn er veel betere modellen beschikbaar.

7. CONCLUSIE EN AANBEVELINGEN

Het doel van het werkstuk was om te laten zien hoe een callcenter data analyst een call volume forecast kan maken met het ARMA-model, het Hol-Winters-model en het aangepaste regressiemodel. Dit werd gedaan aan de hand van een real-life dataset.

De dataset werd grondig geanalyseerd en alle oneffen effecten werden eruit gefilterd. Vervolgens werd de dataset in twee parten gedeeld, de trainingset en de evaluatieset. Met de trainingset werden de parameters van de modellen geschat en hiermee werd voorspellingen gemaakt. De nauwkeurigheid van de voorspellingen, werden uiteindelijk geëvalueerd met de evaluatieset. Daarna werd de nauwkeurigheid van de voorspellingen gemeten aan de hand van een aantal statistische indicatoren. Uitgaande van de statistieken zal een bijpassend model worden gekozen.

De evaluatie van het resultaat heeft echter geen eenduidigheid opgeleverd over de keuze van het model. Dit heeft een aantal oorzaken:

- De statistieken laten geen eenduidige keuze over. Een model met hogere waardering, is nog niet per definitie het beste model, omdat er van het gemiddelde wordt uitgegaan. Als een model gemiddeld beter scoort, betekent dat nog niet dat het model per datapunt gezien ook de juiste voorspelling kan maken.
- Een voorspellingshorizon van 31 dagen geeft grote onzekerheid. De onzekerheid verslechtert de positie van een model. Vooral modellen die aan het begin nauwkeurig konden voorspellen maar minder nauwkeurig aan het eind, kregen hierdoor een slechte waardering.
- In week 33 is het aantal calls ten opzichte van de week daarvoor met bijna 50% gestegen. Door de stijging ontstond meer onzekerheid waardoor de positie van de modellen verslechterden. Uiteindelijk bestond er ook geen eenduidigheid meer over de keuze.

De genoemde problemen kunnen worden voorkomen als de voorspellingshorizon korter is, of als de voorspelling wekelijks aangescherpt mag worden. Hiermee zullen de voorspellingen beter aansluiten op de werkelijkheid. In de praktijk betekent dat dat het werkschema van telefonisten wekelijks kan worden aangepast.

De keuze van het model hangt af van de voorspellingshorizon. Indien de keuze op de korte termijn valt, verdient het Holt-Winters model de voorkeur, omdat deze allereerst vrij nauwkeurige voorspellingen doet op de korte termijn, maar ook omdat het model gemakkelijk te implementeren is. Als de voorspelling op de lange termijn is, sluit het ARMA-model dan weer beter aan op de werkelijkheid. Kortom, de keuze hangt af van de situatie en de beschikbare data, maar ook van de voorkeur van de data-analist.

De oorzaak van de stijging in week 33 kan een jaarlijks fluctuatie zijn (bijvoorbeeld omdat vakantiegangers massaal terugkeren na de vakantieperiodes) of het gevolg van een evenement dat heeft plaatsgevonden. Door onderzoek naar de oorzaak ervan, is dit in de toekomst te voorkomen.

Verder is heel goed naar voren gekomen dat de voorspellingen in de weekeinden niet gemodelleerd hoeven te worden. Het is al voldoende door slechts het gemiddelde te nemen. Dezelfde techniek kan natuurlijk ook worden toegepast op feestdagen.

Het is ook naar voren gekomen dat de keuze van een model niet volledig afhangt van een aantal statistieken, maar vooral van verstand en ervaring. Naarmate er steeds meer gegevens beschikbaar zijn, moet er worden gekeken of er verbeteringen en/of andere nieuwe mogelijkheden zijn. Hiermee wordt de nauwkeurigheid van voorspellingen constant op niveau gebracht.

De stappen die nodig zijn om een voorspelling te maken met het AR(I)MA(p , d , q) model kunnen kort worden samengevat als:

1. Onderzoek de aanwezigheid van trend en seizoenseffecten. Gebruik bij constatering de small trend methode of differencing om ze eruit te filteren.
2. Onderzoek de stationariteit van de tijdreeks met behulp van de Dickey-Fuller-test.
3. Gebruik de ACF/PACF plot om de waarde van p , d en q te herkennen. Kies daarbij een aantal modellen dat verder onderzocht moeten worden.
4. Bereken de Akaike Information Criterion (AIC) van elk van de gekozen modellen en schat de parameters daarvan.
5. Onderzoek de residuen van het model met de laagste AIC. Indien de residuen een witte ruis proces is, dan kan het model worden gebruikt om voorspellingen te maken. Zo niet, kies dan een ander model.
6. Maak de voorspelling en bekijk de prestatie van model. Beoordeel de geschiktheid voor verder gebruik.

Het gebruik van ARMA-modellen heeft een aantal voordelen. Zo is het model geschikt voor langetermijnvoorspellingen. En indien het juiste model wordt gekozen, is het model ook geschikt voor kortetermijnvoorspellingen. Het model is daarnaast gemakkelijk te implementeren, mede door het gekozen softwarepakket. Het nadeel is dat het model vrij ingewikkeld is. De datareeksen moeten aan veel eisen voldoen. Het vereist heel wat onderzoek, inzicht en kennis om een goed voorspelling te kunnen maken. Blindelings vertrouwen op wat het software pakket voorschotel, levert niet altijd bevredigend resultaat op.

Bij het maken van voorspellingen met het Exponential Smoothing model, moet de aanwezigheid van trend en seizoenseffecten ook worden onderzocht. Van de Single Exponential Smoothing methode wordt gebruikt gemaakt, wanneer er geen trend- en seizoenseffecten aanwezig zijn. De Holt's Lineairmethode wordt gebruikt als er alleen een trend aanwezig is. En de Holt-Winters methode wordt gebruikt als de trend- en de seizoenseffecten allebei aanwezig zijn (wat meestal het geval is). Kies vervolgens de beginwaarde van de parameters en initieer het algoritme met de startwaarden. Bereken de overige waarden voor alle tijdstippen. Optimaliseer of kies daarna de parameters waarbij MSE zo laag mogelijk is. Maak vervolgens de voorspelling en beoordeel de geschiktheid van het model voor verder gebruik.

Het model is aantrekkelijk voor het gebruik in de praktijk, mede dankzij de eenvoudige implementatie, maar ook door de uitstekende prestaties op de korte termijn. Verder is het model vrij gemakkelijk te begrijpen voor beginners. Een nadeel van het model dat niet direct uit het resultaat volgt, is dat het model slecht reageert op de uitschieters (zie *Figuur 9.6*). Zodra er een uitschieter is geweest, reflecteert deze meteen op de voorspelling van de komende periode. De data analist moet hiermee oppassen.

Het aangepaste regressiemodel maakt voorspellingen door eerst de (dummy-) variabelen te kiezen. Met de regressieanalyse worden de parameters geschat. Bij de uitvoer van de regressieanalyse behoren ook de teststatistieken van de variabelen. Hiermee kunnen variabelen worden beoordeeld op hun significantie. Indien alle variabelen significant zijn, kunnen voorspellingen worden gemaakt. En uiteindelijk moet worden beoordeeld of het model geschikt is voor verder gebruik.

In het onderzoek dient het regressiemodel als een vergelijking op de ARMA-modellen en de ES-modellen, maar eveneens om een idee te geven dat data analisten ook creatief moeten zijn om correcte voorspellingen te kunnen doen.

Naast de genoemde ARMA-modellen en Exponential Smoothing modellen zijn er nog tal van andere modellen die ook efficiënte voorspellingen kunnen doen. De keuze van het model hangt niet alleen af van de statistische indicatoren, zoals de MSE of de

MAPE, maar ook van de menselijke beoordeling. Vaak presteert een model in de praktijk veel beter dan de statistieken suggereren. Degene die de voorspelling doet, moet daarom goed bedacht zijn op de informatie waarover hij/zij beschikt om een juiste voorspelling te doen.

Als laatste nog dit: een correcte voorspelling doen is slechts het begin van de prestatieverbetering van een callcenter. Want naast de voorspelling van het aantal, moet bijvoorbeeld ook de gespreksduur worden onderzocht, het aantal telefonisten worden berekend die nodig zijn, en uiteindelijk ook hoe de call agents werkelijke worden ingepland. Dit alles vereist een hoeveelheid kennis en inzicht. Pas als alle onderdelen de nodige aandacht hebben gekregen, kunnen goede prestaties worden verwacht. En dit is niet enkel het werk van een callcenter data analist, maar van het hele managementteam.

8. LITERATUURLIJST

- [1] Andrews, B. H. & Cunningham, S. M. (1995). L.L. Bean improves call-center forecasting. *Interfaces*, 25: 6, 1–13.
- [2] Avramidis, A. N., Deslauriers, A. & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50, 896–908.
- [3] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and Control*, 3rd edition. Englewood Cliffs; Prentice-Hall.
- [4] Box, G. E. P. & Pierce, D. A. (1994). Distribution of the residual autocorrelations in autoregressive-integrated moving-average time series models, *Journal of the American Statistical Association*, 65, 1509-1526.
- [5] Brockwell, P. J. & Davis, R. A. (1996). *Introduction to Time Series and Forecasting*, 2nd edition. p36. New York; Springer-Verlag.
- [6] Brown, R. G. (1956). Exponential Smoothing for predicting demand. Presented at the tenth national meeting of the Operations Research Society of America. San Francisco, 16 november 1956.
- [7] Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc*, 100, 36–55.
- [8] Chatfield, C. (2000). *Time-Series Forecasting*. Chapman and Hall / CRC Press.
- [9] Dickey, D. A. & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, 74, p427-431.
- [10] Gunst, de M. C. M. (2009). *Statistical Models Reader*. Amsterdam; Vrije Universiteit.
- [11] Holt, C. C. (1957). Forecasting seasonals and trend by exponentially weighted moving averages. Pittsburgh, PA; Technical report, Carnegie Institute of Technology.
- [12] Hyndman, R. J. (2011). Forecasting functions in time series in R. Verkregen op 24 December, 2011 van <http://cran.r-project.org/web/packages/forecast/forecast.pdf>

- [13] Jongbloed, G. & Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Appl. Stoch. Models Bus. Ind.*, 17, 307–318.
- [14] Ljung, G. M. & Box, G. E. P. (1978), On a measure of lack of fit in time series models, *Biometrika*, 65, 297-303.
- [15] Magee, J. F. (1967). *Production Planning and Inventory Control*. New York; McGraw-Hill.
- [16] Makridakis, S. G. (1990). *Forecasting, planning, and strategy for the 21st century*. New York; The Free Press.
- [17] Makridakis, S. G., Wheelwright, S.C. & Hyndman R.J. (1998). *Forecasting: Methods and Applications*, 3rd edition. Hoboken; John Wiley and Sons, Inc.
- [18] Powell, S. G. & Baker, K. R. (2007). *Management Science: The art of modeling with spreadsheets*, 2nd edition. Hoboken; John Wiley and Sons, Inc.
- [19] Tanir, O. & Booth, R. J. (1999). Call center simulation in Bell Canada. In *Proceedings of the 1999 Winter Simulation Conference* (P. A. Farrington, N. H. B. Nembhard, D. T. Sturrock and G. W. Evans, eds.) 1640–1647. IEEE Press, Piscataway, NJ.
- [20] Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, 54, 253–265.
- [21] Winters, P. R. (1960). Forecasting sales by exponentially weighed moving averages. *Management Science*, 6, 324–342.

[22] Locatie van de R Project software pakketten:

<http://cran.r-project.org/web/packages/quadprog/index.html>

<http://cran.r-project.org/web/packages/tseries/index.html>

<http://cran.r-project.org/web/packages/zoo/index.html>

<http://cran.r-project.org/web/packages/forecast/index.html>

<http://cran.r-project.org/web/packages/fracdiff/index.html>

<http://cran.r-project.org/web/packages/Rcpp/index.html>

<http://cran.r-project.org/web/packages/RcppArmadillo/index.html>

9. BIJLAGE

Tabel 9.1 Dagelijkse ontvangen calls

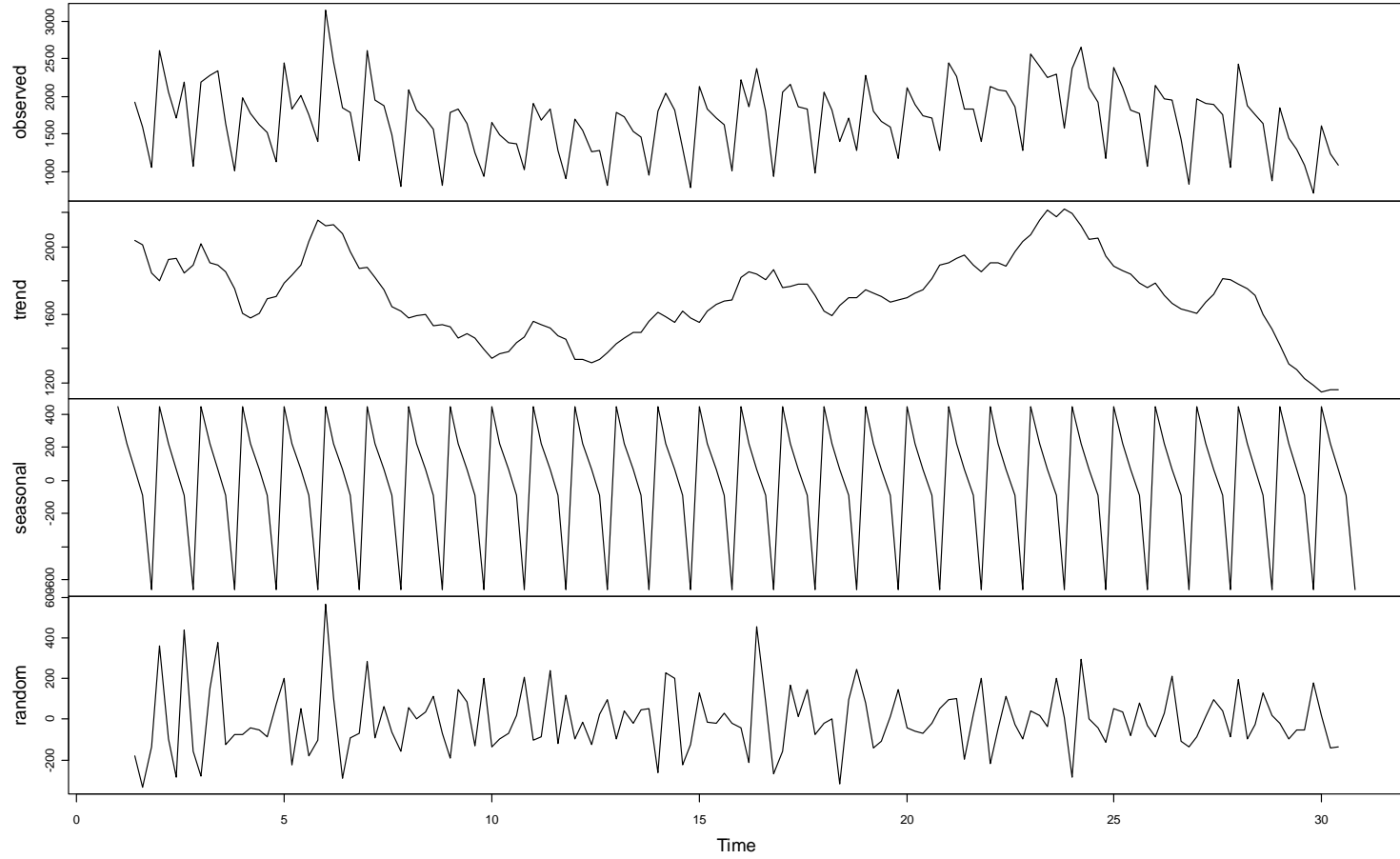
Week		Maandag	Dinsdag	Woensdag	Donderdag	Vrijdag	Zaterdag	Zondag
1	Training set	2754	2871	1928	1588	1050	96	52
2		2611	2049	1710	2192	1075	45	53
3		2183	2284	2341	1642	1015	68	33
4		1981	1766	1622	1523	1125	52	31
5		2437	1838	2012	1760	1394	38	25
6		3137	2462	1853	1793	1140	110	31
7		2606	1952	1883	1496	803	36	24
8		2091	1823	1702	1560	812	34	22
9		1785	1836	1643	1244	938	54	34
10		1653	1494	1383	1365	1021	67	47
11		1903	1681	1829	1273	914	47	17
12		1691	1552	1262	1274	813	85	94
13		1781	1731	1539	1455	954	29	34
14		1799	2039	1823	1310	795	67	34
15		2136	1830	1709	1624	1011	43	21
16		2222	1861	2363	1799	797	34	28
17		45	2161	1858	1837	980	19	23
18		2050	1824	1407	103	396	25	22
19		2275	1809	1663	1596	1174	41	27
20		2108	1889	1743	1706	1284	68	40
21		2447	2259	1826	1828	1400	76	76
22		2129	2087	1572	92	365	55	63
23		2556	2395	2249	2290	1578	73	58
24		143	2644	2117	1921	1172	37	26
25		2383	2117	1823	1779	1074	54	22
26		2146	1973	1947	1434	827	51	40
27		1970	1912	1886	1764	1061	49	23
28		2421	1879	1759	1642	877	31	32
29		1849	1438	1288	1086	707	27	35
30		1611	1236	1091	1128	730	41	34
31	Evaluatie set	1483	1190	1079	902	626	34	38
32		1826	1338	1132	1116	621	31	26
33		2704	2080	1612	1380	888	43	39
34		2579	2571	2149	1639	942	38	31
35		2434	2329	2039				

= Geeft de datapunten aan die zijn aangepast

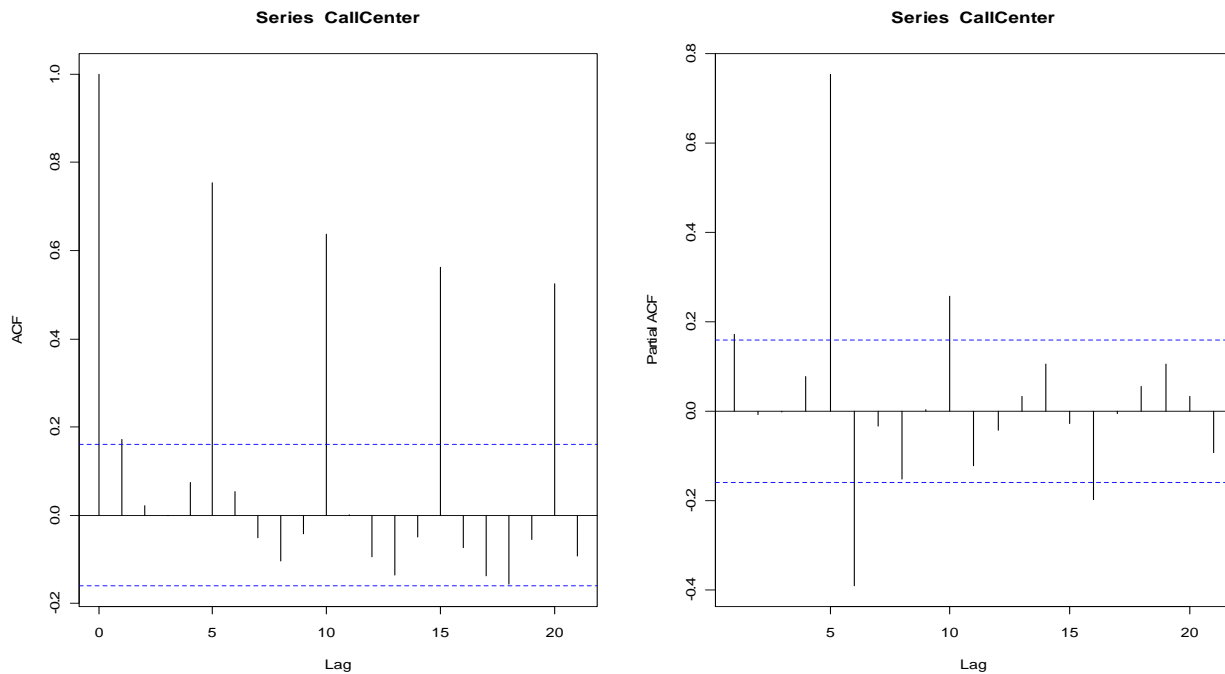
Tabel 9.2 Aangepaste Data punten

Weeknummer	Weekdag	Oorspronkelijk	Aanpassing	Reden
16	Vrijdag	797	935	Goede vrijdag
17	Maandag	45	2052	2 ^e paasdag
17	Zaterdag	19	43	Koninginnedag
18	Donderdag	103	1710	Bevrijdingsdag
18	Vrijdag	396	1286	Na-effect Bevrijdingsdag
22	Woensdag	1572	2063	Voor-effect hemelvaartsdag
22	Donderdag	92	1856	hemelvaartsdag
22	Vrijdag	365	1275	Na-effect hemelvaartsdag
24	Maandag	143	2362	2e pinksterdag

Decomposition of additive time series



Figuur 9.1 Decompositie van de data



Figuur 9.2 ACF/PACF plot van de data

Dickey Fuller Test op oorspronkelijke data:

```
> adf.test(CallCenter)
```

Augmented Dickey-Fuller Test

```
data: CallCenter
Dickey-Fuller = -1.7963, Lag order = 5, p-value = 0.6613
alternative hypothesis: stationary
```

R Codes voor ARMA Model

```
# Lees data in
CallCenter <- scan("CallCenterz.txt")
```

```
# Laad functiepakketten
library(quadprog)
library(tseries)
library(stats)
library(forecast)
```

```

# Plot van data
ts_CallCenter <- ts(CallCenter, freq=5)
fil_CallCenter <- decompose(ts_CallCenter)
plot(fil_CallCenter)
acf(CallCenter)
pacf(CallCenter)

# Dickey Fuller Test
adf.test(CallCenter)

Model1.fit <- auto.arima(ts_CallCenter, trace=TRUE, test=c("adf"))
# Portmanteau test
Box.test(Model1.fit$residuals, type = c("Box-Pierce"),fitdf = 4, lag=5)
Box.test(Model1.fit$residuals, type = c("Ljung-Box"),fitdf = 4, lag=5)
# Plot residuals
plot(resid(Model1.fit))
acf(resid(Model1.fit))
pacf(resid(Model1.fit))
# Maak voorspelling
forecast(Model1.fit,h=23)

Model2.fit <- auto.arima(ts_CallCenter, trace=TRUE)
# Portmanteau test
Zelfde als Model 1
# Plot residuals
Zelfde als Model 1
# Maak voorspelling
Zelfde als Model 1

Model3.fit <- auto.arima(ts(diff(CallCenter, lag=5, differences=1), freq=5), trace=TRUE)
# Portmanteau test
Zelfde als Model 1
# Plot residuals
Zelfde als Model 1
# Maak voorspelling
predict <- forecast(Model3.fit,h=23)
values <- c(as.vector(t(CallCenter[146:150])),predict$mean)
for(i in 1:23) {
  values[i + 5] <- values[i + 5] + values[i]
}
Voorspelling <- values[6:28]

```

		Alpha	0.5	Beta	0.5	Gamma	0.5		
		Period							
		t	X(t)	L(t)	b(t)	S(t)	Forecast		
Initial Values	1		2754			1.35		X1/L5	
	2		2871			1.41			
	3		1928			0.95		Gemiddelde(X1-X5)	
	4		1588			0.78			
	5		1050	2038.2	-22.2	0.52		(Som(X6:X10) - Som(X1-X5))/25	
Forecast	6		2611	1974.2	-43.1	1.34	2724	Formule 5.9	
	7		2049	1692.9	-162.2	1.31	2720	Formule 5.11	
	8		1710	1669.2	-92.9	0.99	1448	Formule 5.12	
	9		2192	2194.9	216.4	0.89	1228	Formule 5.10	
	10		1075	2249.0	135.2	0.50	1242		
	⋮		⋮	⋮	⋮	⋮	⋮		
	146		1611	1249.4	0.1	1.24	1359		
	147		1236	1219.2	-15.1	1.03	1299		
	148		1091	1148.4	-43.0	0.97	1202		
	149		1128	1175.8	-7.8	0.93	1001		
	150		730	1231.2	23.8	0.58	659		

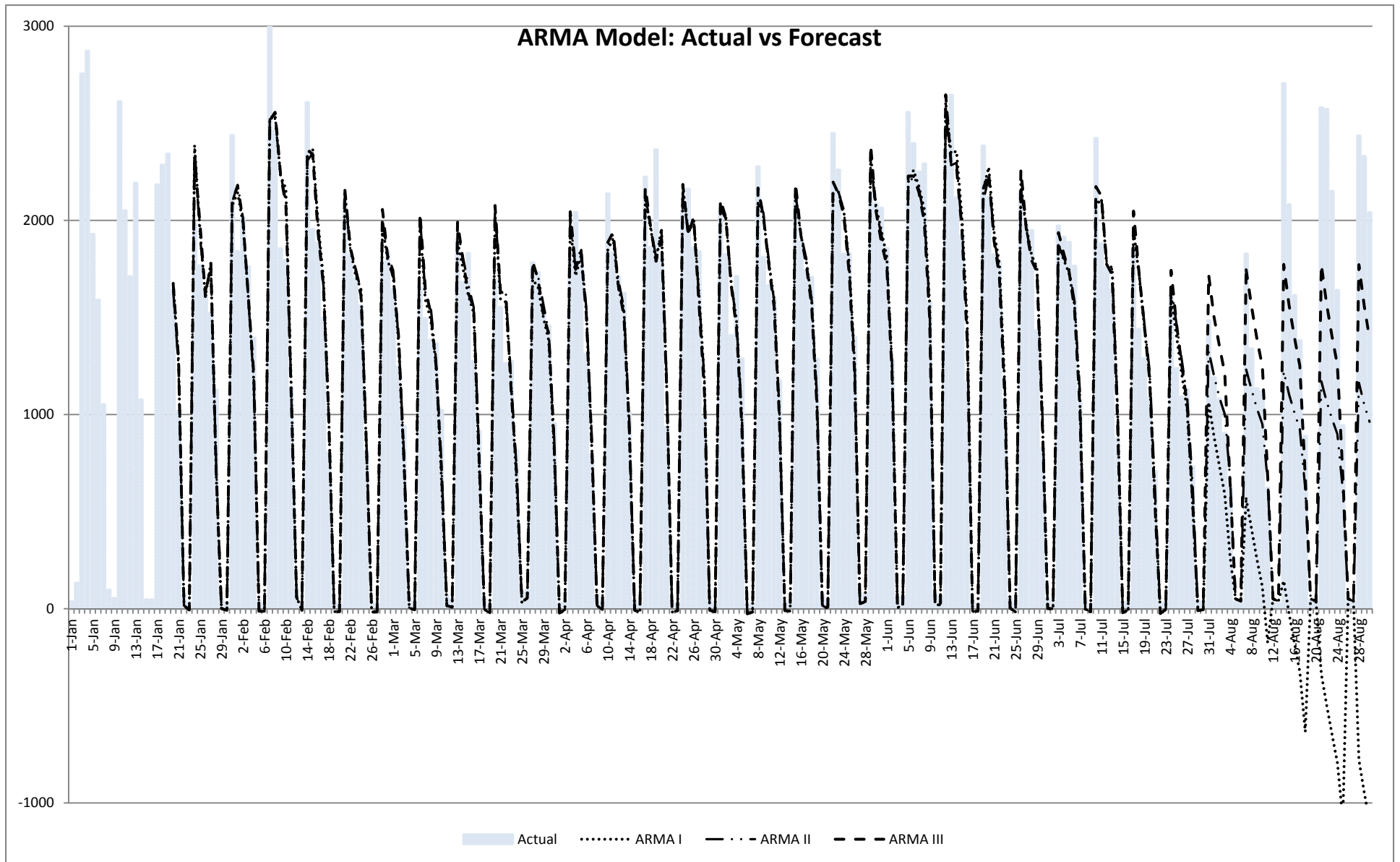
Figuur 9.3 Opzet Holt-Winters Methode in Excel

Datum	X _t			Dummy Variabelen Z _t			Dummy Variabelen β _i				
	Observed	LN(t)	Differ(t)	t	Maand	Dag	Monday	Tuesday	Wednesday	Thursday	Friday
3-1-2011	2754	7.9208	-30.310	1	1	3	1	0	0	0	0
4-1-2011	2871	7.9624	-30.310	2	1	4	0	1	0	0	0
5-1-2011	1928	7.5642	-30.310	3	1	5	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
27-7-2011	1091	6.9948	-197.000	148	7	27	0	0	1	0	0
28-7-2011	1128	7.0282	42.000	149	7	28	0	0	0	1	0
29-7-2011	730	6.5930	23.000	150	7	29	0	0	0	0	1

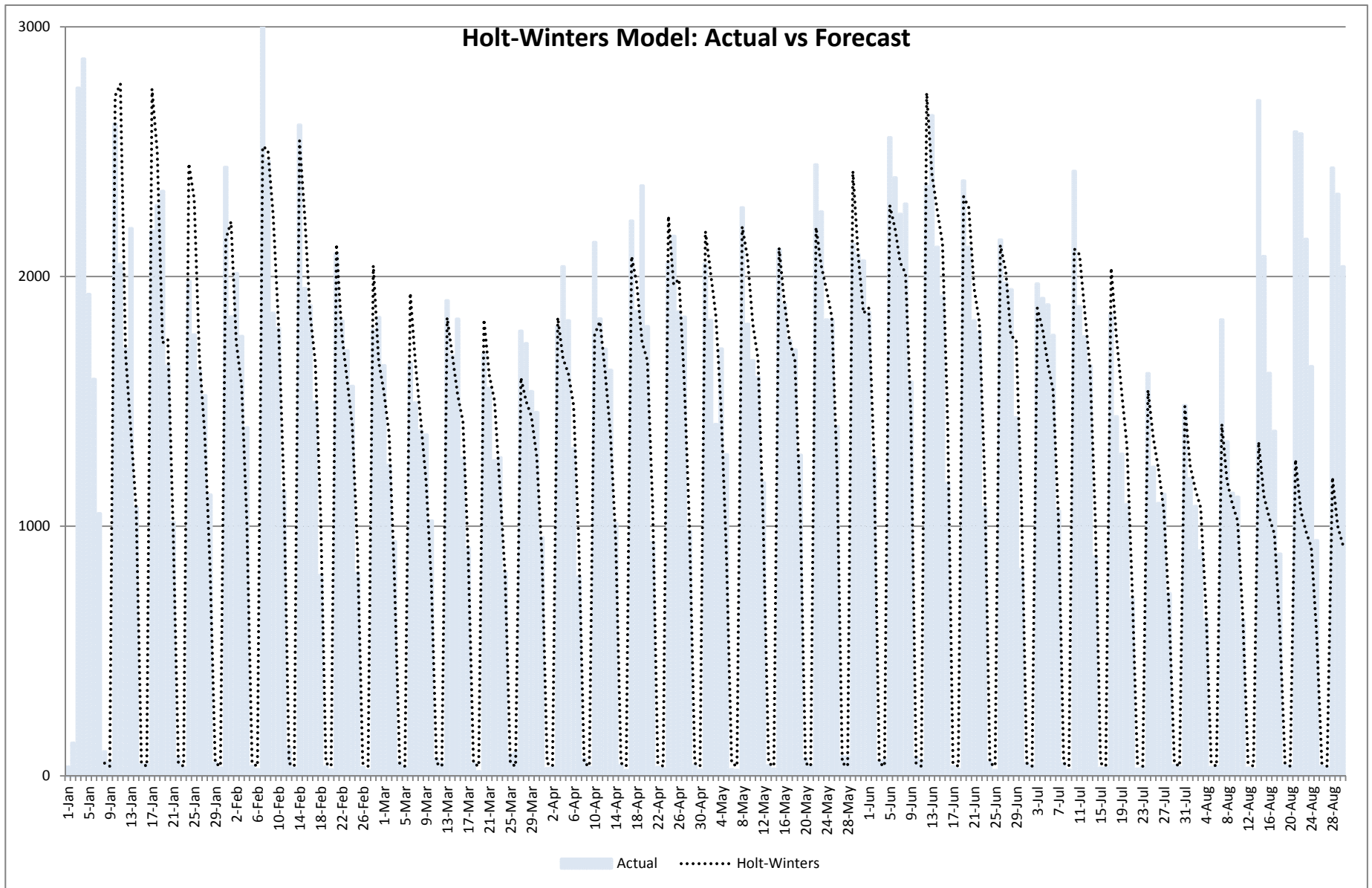
Figuur 9.4 Opzet Regressie Methode in Excel

Tabel 9.3 Resultaat van alle modellen

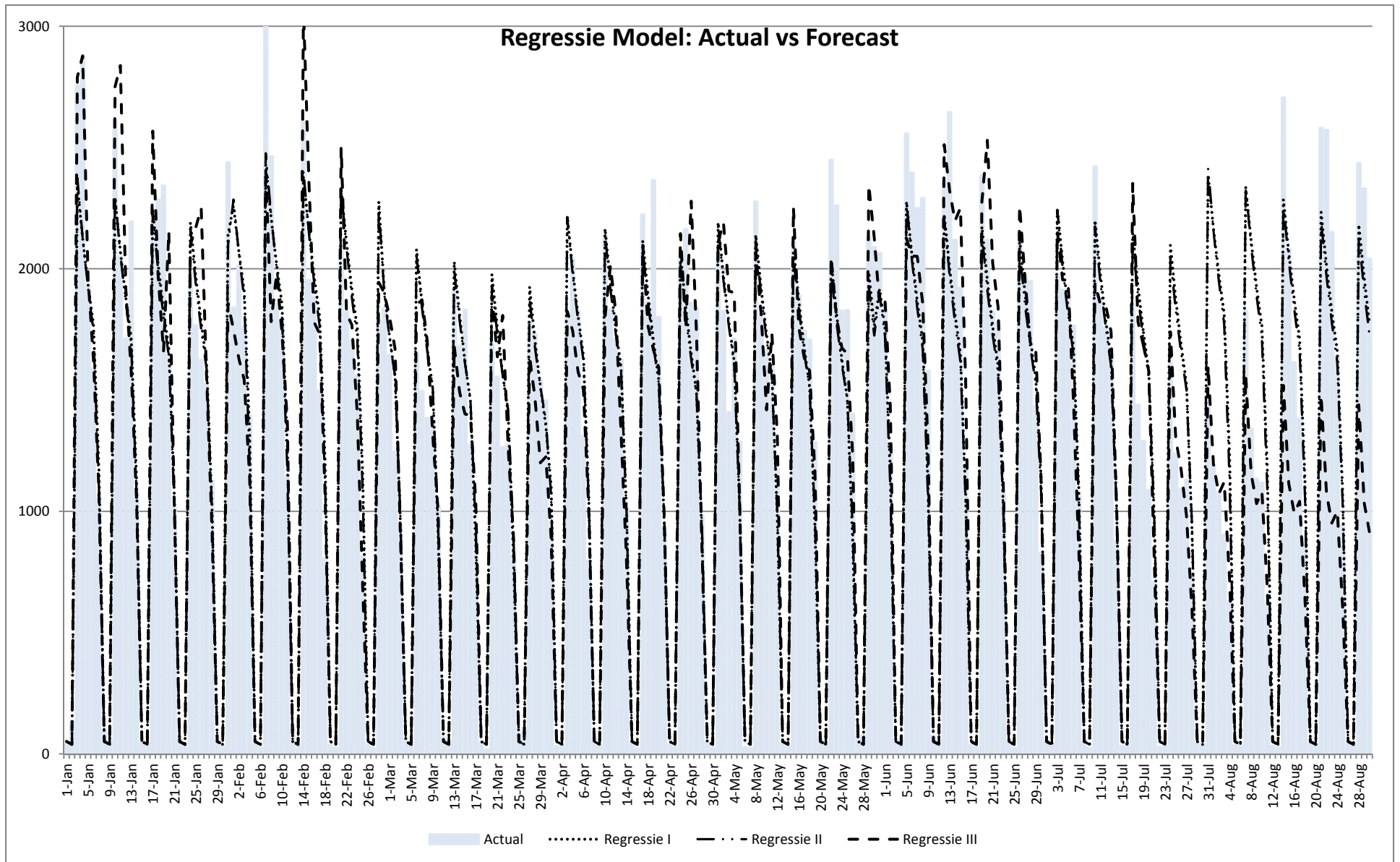
Datum	Actual	Forecast						
		HoltWinters	ARMA			Regressie		
			Model I	Model II	Model III	Model A	Model B	Model C
8/1/2011	1483	1480	1064	1329	1719	2388	2413	1599
8/2/2011	1190	1242	887	1199	1503	2167	2159	1195
8/3/2011	1079	1148	737	1091	1356	1976	1949	1072
8/4/2011	902	1075	589	1002	1212	1815	1772	1121
8/5/2011	626	670	268	671	648	1222	1124	718
8/6/2011	34	51	51	51	51	51	51	51
8/7/2011	38	39	39	39	39	39	39	39
8/8/2011	1826	1409	566	1231	1746	2337	2339	1559
8/9/2011	1338	1181	410	1128	1554	2115	2093	1155
8/10/2011	1132	1091	256	1031	1390	1925	1890	1032
8/11/2011	1116	1022	113	951	1249	1764	1718	1081
8/12/2011	621	637	-175	660	685	1171	1090	677
8/13/2011	31	51	51	51	51	51	51	51
8/14/2011	26	39	39	39	39	39	39	39
8/15/2011	2704	1338	144	1222	1773	2285	2268	1519
8/16/2011	2080	1121	-28	1101	1560	2064	2030	1115
8/17/2011	1612	1035	-178	1009	1396	1874	1833	992
8/18/2011	1380	969	-325	924	1252	1712	1666	1041
8/19/2011	888	603	-634	617	681	1120	1057	637
8/20/2011	43	51	51	51	51	51	51	51
8/21/2011	39	39	39	39	39	39	39	39
8/22/2011	2579	1267	-336	1171	1772	2234	2199	1479
8/23/2011	2571	1061	-496	1068	1568	2013	1968	1074
8/24/2011	2149	979	-649	977	1401	1822	1777	951
8/25/2011	1639	916	-793	899	1255	1661	1615	1001
8/26/2011	942	570	-1085	610	686	1068	1025	597
8/27/2011	38	51	51	51	51	51	51	51
8/28/2011	31	39	39	39	39	39	39	39
8/29/2011	2434	1196	-778	1164	1772	2183	2132	1439
8/30/2011	2329	1001	-946	1052	1560	1961	1908	1034
8/31/2011	2039	923	-1097	963	1392	1771	1723	911



Figuur 9.5 Voorspelling vs Werkelijk: ARMA Modellen



Figuur 9.6 Voorspelling vs Werkelijk: Holt-Winters Model



Figuur 9.7 Voorspelling vs Werkelijk: Regressie Model

