

# **Verminderen van onbalanskosten op elektriciteitsverbruik**

Verbetering van het korte termijn voorspellingsmodel,  
klantgroepering en individuele klantkennis

B.T. Tromp

Afstudeerverslag



# Verminderen van onbalanskosten op elektriciteitsverbruik

---

*Verbetering van het korte termijn voorspellingsmodel, klantgroepering en individuele klantkennis*

B.T. Tromp

Afstudeerverslag  
Vrije Universiteit Amsterdam  
Faculteit der Exacte Wetenschappen  
Bedrijfskunde en Informatica  
De Boelelaan 1081a  
1081 HV Amsterdam

Opdrachtgever:  
Hot ITem BV  
Danzigerkade 19  
1013 AP Amsterdam

mei 2011



## Voorwoord

Vanaf het begin van mijn opleiding Bedrijfswiskunde en informatica aan de Vrije Universiteit Amsterdam heb ik naar dit verslag toe gewerkt. Gedurende de studie heb ik pieken en dalen gekend, maar vooral veel geleerd. Telkens als ik het even niet zag zitten, heeft het vooruitzicht naar dit verslag mij over het dode punt geholpen. Dit verslag is mijn afstudeerverslag en vormt de afsluiting van mijn fulltime opleidingscarrière. Na goedkeuring van dit verslag ben ik afgestudeerd en mag ik mij Master of Science noemen.

Het verslag behandelt de onbalanskosten op elektriciteitsverbruik en beschrijft drie manieren om deze te verminderen. De eerste manier is het huidige voorspellingsmodel vervangen door een model dat het verbruik voor de volgende dag beter voorspelt. De tweede manier is om alle klanten in groepen in te delen en om vervolgens elke groep apart te voorspellen. De derde manier analyseert klanten individueel op de bijdrage die het levert aan de onbalans.

Hot ITem is de opdrachtgever van dit onderzoek. De begeleiders van Hot ITem die mij hebben geholpen met dit onderzoek wil ik bedanken. Dit zijn de heren J. de Jonge en A. Scheepmaker. Zonder de uitleg, ervaring en enthousiasme van de heer De Jonge was het mij niet gelukt om dit verslag op deze manier tot stand te brengen. Mede door de ervaring en het overleg met de heer Scheepmaker is dit verslag geworden wat het is. Vanuit de VU heb ik een begeleider en tweede lezer tot mijn beschikking gehad. De heer S. Bhulai wil ik bedanken voor zijn ondersteuning, inzet en feedback als begeleider. De heer A. Roubos wil ik bedanken voor zijn inzet en feedback als tweede lezer.

*Brian Tromp*



## Samenvatting

Onbalanskosten worden veroorzaakt door het verschil tussen de voorspelling en het werkelijke verbruik van elektriciteit. In dit verslag wordt op drie manieren onderzocht of de onbalanskosten verminderd kunnen worden. De eerste manier is een onderzoek naar een beter voorspellingsmodel voor een dag vooruit. De tweede manier onderzoekt of de volgende dag beter te voorspellen is als klanten in groepen worden verdeeld. De derde manier is een onderzoek naar de mogelijkheden om klanten een individuele bijdrage van de onbalans toe te kennen.

[VERTROUWELIJK]

Geconcludeerd kan worden dat de grootste verbetering bereikt wordt door het huidige voorspellingsmodel te vervangen door het exponential smoothing model. Klanten in groepen indelen en de groepen apart voorspellen zorgt voor een verdere verbetering van de voorspelling. De beste voorspelling wordt verkregen door de som van de exponential smoothing groepsvoorspellingen zonder de groep van nieuwe en vertrekkende klanten mee te nemen. Verder is het mogelijk om klanten te onderscheiden op de bijdrage van onbalans. Sommige klanten zijn verantwoordelijk voor veel onbalans en sommige klanten zijn verantwoordelijk voor minder onbalans.

Aanbevolen wordt om vervolg onderzoek te doen naar het Multivariate Adaptive Regression Splines (MARS) model om de voorspelling nog een stap te verbeteren. Verwacht wordt dat dit model nog beter presteert dan het exponential smoothing model. Daarnaast zou verder onderzoek naar klantkennis ook kunnen bijdrage aan verdere verbetering van de voorspelling. Om een individuele onbalanspremie toe te passen in de praktijk is verder onderzoek nodig naar de manier waarop dit het beste mogelijk is.





# Inhoudsopgave

<b>1 INLEIDING .....</b>	<b>1</b>
1.1 ONBALANS .....	1
1.2 PROBLEEMSTELLING .....	2
1.3 DOELSTELLING .....	4
1.4 RELEVANTE INFORMATIE OVER HET ONDERZOEKSBEDRIJF .....	5
1.5 STRUCTUUR VAN HET VERSLAG .....	5
<b>2 DATAVERZAMELING.....</b>	<b>7</b>
<b>3 BESPREKING VAN DE LITERATUUR .....</b>	<b>9</b>
3.1 CRITERIA UIT DE LITERATUUR.....	10
3.2 DIEPERE UITWERKING VAN DE BESTE MODELLEN .....	13
3.2.1 <i>Exponential smoothing</i> .....	13
3.2.2 <i>ARIMA</i> .....	14
<b>4 ONDERZOEK ONTWERP.....</b>	<b>17</b>
4.1 VOORSPELLINGSMODEL .....	17
4.1.1 <i>Exponential smoothing</i> .....	17
4.1.2 <i>ARIMA</i> .....	18
4.2 KLANTGROEPERING .....	23
4.3 INDIVIDUELE KLANTKENNIS .....	25
<b>5 RESULTATEN .....</b>	<b>27</b>
5.1 VOORSPELLINGSMODEL .....	27
5.2 KLANTCLUSTERING .....	29
5.3 INDIVIDUELE KLANTKENNIS .....	31
<b>6 BESPREKING VAN DE PRAKTIJK .....</b>	<b>33</b>
<b>7 CONCLUSIE.....</b>	<b>35</b>
<b>8 VERVOLG ONDERZOEK.....</b>	<b>37</b>
<b>9 EVALUATIE VAN HET ONDERZOEK.....</b>	<b>39</b>
<b>10 REFERENTIES .....</b>	<b>41</b>
<b>11 BIJLAGE .....</b>	<b>43</b>
11.1 BIJLAGE A: SAMENVATTING VAN ARIMA REGELS .....	43
11.2 BIJLAGE B: MARS IMPACTANALYSE .....	45
11.3 BIJLAGE C: AFKORTINGEN .....	46



## 1 Inleiding

De prijs van elektriciteit is opgebouwd uit verschillende componenten. De eerste en belangrijkste component zijn de kosten die gepaard gaan met het opwekken van energie. Daarnaast zijn er risico's die vooraf ingecaluleerd moeten worden en is er altijd een vorm van overhead die bekostigd moet worden. Dit onderzoek richt zich op het risico van onbalans die in de vorm van een vaste premie per afgenomen megawattuur (MWh) aan de klant wordt doorberekend. De volgende paragraaf gaat dieper in op het thema onbalans en legt verder uit hoe hiermee wordt omgegaan. Elektriciteit en energie worden afwisselend gebruikt in dit verslag. Met beide termen wordt hetzelfde bedoeld.

### 1.1 Onbalans

Een energiemaatschappij moet elke dag een voorspelling doorgeven aan de netbeheerder. In deze voorspelling wordt aangegeven hoeveel elektriciteit de klanten de volgende dag gaan verbruiken en wordt aangegeven bij welke producenten ze dit gaan afnemen. Verbruik en afname moeten te allen tijde gelijk zijn wegens de kans op stroomuitval bij te veel onbalans. Onbalans ontstaat doordat de voorspelling niet overeen komt met het werkelijke verbruik. Dit is het geval bij zowel meer als minder verbruik dan voorspeld. Op het moment dat er onbalans dreigt, herstelt de netbeheerder dit door achter de hand gehouden capaciteit op of af te schakelen. De kosten die hier mee gemoeid zijn worden verhaald op de energiemaatschappij die hier verantwoordelijk voor is. Dit zijn de zogenoemde onbalanskosten.

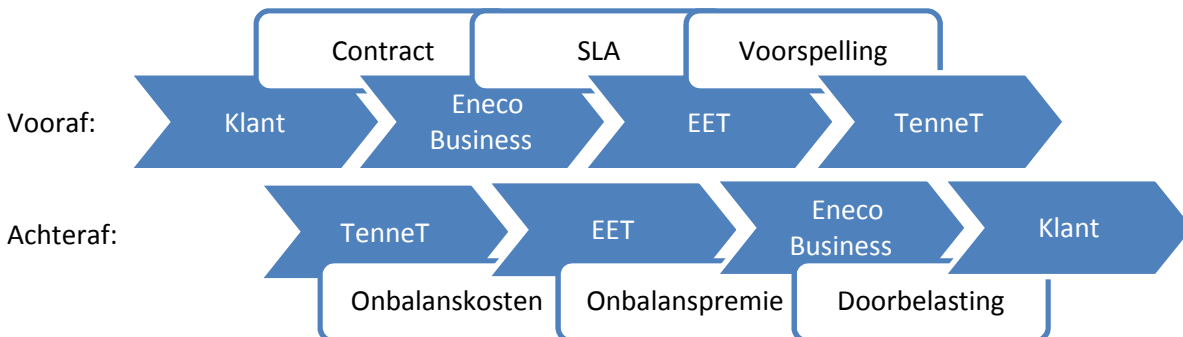
Eneco is de energiemaatschappij waar dit onderzoek zich op richt. Eneco heeft verschillende afdelingen waaronder Eneco Business en Eneco Energy Trade (EET). Eneco Business is de zogenoemde leverancier van elektriciteit voor de zakelijke klanten en EET de programmaverantwoordelijke. De leverancier heeft contracten afgesloten met klanten over de prijs, duur en hoeveelheid van het energieverbruik. De programmaverantwoordelijke voorspelt voor alle afdelingen het verbruik voor de volgende dag. Deze voorspelling moet voor 12 uur 's middags bij de netbeheerder binnen zijn en moet op kwartierbasis het totale verbruik weergeven.

TenneT is de netbeheerder in Nederland. Vooraf controleert TenneT of de voorspelling overeenkomt met de inkoop. Gedurende de dag herstelt zij de ontstane onbalans door de voorspelling en het werkelijk verbruik te vergelijken. Als het totaal van de voorspellingen een tekort of overschot oplevert, schakelt TenneT de reservecapaciteit respectievelijk op of af. Achteraf worden deze onbalanskosten doorberekend aan de programmaverantwoordelijken.

Door de rol van EET binnen Eneco, krijgt EET de onbalanskosten van TenneT gefactureerd als de voorspelling niet overeenkomt met het werkelijke verbruik. In de praktijk is het bijna onmogelijk een dag lang geen onbalans te veroorzaken. Om deze onbalanskosten te kunnen betalen heeft EET een service level agreement (SLA) met de afdelingen waarvoor ze voorspellen. Op dit moment hebben EET en Eneco Business een SLA dat voor iedere MWh dat verbruikt wordt een vaste onbalanspremie betaald moet worden. De hoogte van deze premie is vastgesteld aan de hand van een eenvoudig voorspellingsmodel. Dit model is de zogenoemde  $n - 7$  methode. Het model geeft als voorspelling voor dag  $n$  het verbruik van zeven dagen geleden. Wat bijvoorbeeld op maandag in het kwartier tussen 12 uur en kwart over 12 werd verbruikt is de voorspelling voor precies dat kwartier een week later.

Door de SLA met EET staat de onbalanspremie vast die Eneco Business moet betalen voor ieder MWh dat hun klanten afnemen. Doordat dit vast staat kan Eneco Business niet veel anders dan deze premie

direct overnemen in het contract met de klant. In Figuur 1 zijn de relaties nog een keer schematisch weergegeven.



Figuur 1: Relatiediagram van de partijen die een rol spelen in dit onderzoek

## 1.2 Probleemstelling

Klanten zijn kritischer en minder loyaal geworden aan hun leveranciers.<sup>1</sup> Door de komst van internet met hun vergelijkingssites kunnen klanten eenvoudig vergelijken en overstappen als de concurrent voordeliger of beter is. Wegens de eenzijdigheid van het product elektriciteit is de prijs een belangrijk onderdeel om onderscheidend te zijn ten opzichte van de concurrentie. Door het duurzame beleid dat Eneco voert, heeft het de naam een prijzige energiemaatschappij te zijn. Eneco Business zoekt naar mogelijkheden om de kosten te verlagen. Een kostenpost die geen waarde toevoegt is de doorbelasting van de onbalanspremie. Eneco Business wil de doorbelasting verlagen zodat de totaalprijs per MWh die ze kunnen aanbieden aan hun klanten concurrerender is. Een prijsverlaging voor de klant zonder er zelf op achteruit te gaan is alleen mogelijk als de SLA met EET gewijzigd wordt. Eneco Business ziet hier een oplossing voor. De SLA tussen Eneco Business en EET over de hoogte van de onbalanspremie is gebaseerd op het zojuist genoemde  $n - 7$  voorspellingsmodel. Door de onnauwkeurigheid van dit model komt er een hoge onbalanspremie uit voort. EET voorspelt zelf nauwkeuriger dan het  $n - 7$  model. Hierdoor betalen ze minder onbalanskosten aan TenneT dan ze binnen krijgen van Eneco Business via de onbalanspremie.

Als Eneco Business een nauwkeuriger model heeft dan het huidige model, kan er een nieuw SLA gemaakt worden over de hoogte van de onbalanspremie. Met een nauwkeuriger model komen de onbalanskosten lager uit waardoor de onbalanspremie ook verlaagd kan worden. Hierdoor kan de totaalprijs per MWh naar beneden. Een eis aan het nieuwe model is dat het eenvoudig is en niet meer werk moet opleveren: iedereen bij Eneco Business zou het nieuwe model moeten kunnen gebruiken en uitlegen tijdens de ontmoeting met EET over een nieuw SLA.

Op de hierboven beschreven manier levert de nieuwe SLA tussen Eneco Business en EET alleen iets op voor Eneco Business: Eneco Business betaalt namelijk minder onbalanspremie per MWh aan EET, maar EET moet evenveel onbalanskosten betalen aan TenneT. Dit komt doordat EET al een zeer nauwkeurige voorspelling heeft en het nieuwe model waarschijnlijk niet nauwkeuriger is dan het model dat EET reeds gebruikt. De reden hierachter is dat het nieuwe model eenvoudig moet blijven, maar EET zelf met een

<sup>1</sup> Davenport, T.H. en Harris, J.G. *Competing on analytics: The new science of winning*. Boston: Harvard Business School Press, 2007. Dit slaat niet alleen op de energiemarkt, maar is generiek toepasbaar.

uitgebreider en complexer model werkt. De ontmoeting tussen Eneco Business en EET zou makkelijker zijn als beide partijen er iets aan hebben. Om dit te bereiken wil Eneco Business EET helpen om nauwkeuriger te voorspellen. Aangezien voorspellen zelf de expertise is van EET en niet van Eneco Business, moet er gekeken worden naar wat Eneco Business wel heeft om de voorspelling mogelijk te verbeteren. De kennis die Eneco Business wel heeft en EET niet, is alles wat met de klant te maken heeft. EET voorspelt en ziet alleen het totaalverbruik van de klanten van Eneco Business. Eneco Business weet het verbruik van de individuele klant en verwacht hiermee over informatie te beschikken waarmee klanten apart, of in bepaalde groepen beter te voorspellen zijn. Het kunnen analyseren van aparte klanten en groepen zorgt ervoor dat het verwachte verbruikspatroon zuiverder wordt dan het geval is bij het totaalverbruik. Door de zuiverdere eigenschappen kan het verbruik beter voorspeld worden. Doordat EET niet differentieert op klantniveau, kan dit een verbetering betekenen voor de voorspelling van EET. Als dit inderdaad tot een verbetering leidt, kan het gesprek met EET een andere insteek krijgen en kunnen beide partijen een verbetering tegemoet zien.

Eneco Business ziet, als gezegd, groot verschil in soorten klanten. Zo kan er een onderscheid gemaakt worden tussen klanten die een stabiel verbruikspatroon hebben en klanten die juist een grillig verbruikspatroon hebben. Doordat er niet gedifferentieerd op klantniveau is het niet bekend welke klanten een stabiel of grillig patroon hebben. Hierdoor betaalt iedere klant dezelfde onbalanspremie. Dit is intuïtief niet rechtvaardig en Eneco Business wil onderzoeken of het mogelijk is om klanten te onderscheiden op de individuele bijdrage aan de onbalans.

Overigens komt het al voor dat klanten onderhandelen over de hoogte van de onbalanspremie en daarbij eisen dat die omlaag gaat. Wegens het willen behouden van sommige klanten, kan aan die eis voldaan worden. Als de ene klant minder onbalanspremie betaalt moet dat gecompenseerd worden door de andere klanten of moet Eneco Business dat uit eigen portefeuille betalen. Op dit moment is Eneco Business niet in staat een onbalanspremie op klantniveau toe te passen. Om te voorkomen dat Eneco Business meer uit eigen portefeuille moet betalen dan ze nu doet, wilt ze onderzoeken of het mogelijk is per klant vast te stellen in welke mate deze verantwoordelijk is voor de onbalans. Daarmee kan bekeken worden welke onbalanskosten dit met zich mee brengt en welke onbalanspremie dit oplevert. Deze onbalanspremie zou de klant dan moeten betalen en op die manier wordt ervoor gezorgd, dat Eneco Business geen onbalanskosten uit eigen portefeuille hoeft te betalen.

De hierboven beschreven probleemstellingen kunnen we vertalen in drie onderzoeksvragen. De eerste vraag is de hoofdvraag en beslaat het belangrijkste deel van dit onderzoek. Er is weinig tot geen voorkennis over de tweede en derde vraag. Deze dienen vooral als zoeklicht en kunnen een mogelijke opzet voor vervolgonderzoek vormen. De onderzoeksvragen zijn:

- *Is er een voorspellingsmodel dat beter is dan het huidige model?*

Eneco Business is niet tevreden over het huidige model. Het veroorzaakt een te hoge onbalanspremie. Ze verwacht dat er een eenvoudig model bestaat dat nauwkeuriger is dan het huidige model. Dit nieuwe model wilt ze gebruiken om de SLA over de onbalanspremie met EET te veranderen.

- *Kunnen klanten zo gegroepeerd worden dat de som van de groepsvoorspellingen beter is dan de voorspelling van het totaalverbruik?*

Eneco Business heeft niet de kunde om de voorspelling van EET over te nemen, maar kan EET waarschijnlijk wel helpen beter te voorspellen. Eneco Business heeft gegevens waardoor klanten gegroepeerd kunnen worden. Deze opdeling in groepen maakt een adequatere voorspelling wellicht mogelijk.

- *Kan er een onbalanspremie op klantniveau gemaakt worden?*

Klanten die goed voorspelbaar zijn, willen een lagere onbalanspremie. Om te voorkomen dat Eneco Business zelf voor de kosten opdraait, moeten klanten die minder goed te voorspellen zijn een hogere onbalanspremie betalen. De vraag is of klanten te onderscheiden zijn op de bijdrage van onbalans zodat een onbalanspremie op klantniveau te maken is.

### 1.3 Doelstelling

[VERTROUWELIJK]

Het doel van dit onderzoek is om te achterhalen of het mogelijk is om de onbalanskosten te verlagen. Als dit kan is de vervolgvraag met hoeveel het verlaagd kan worden. De eerste en belangrijkste doelstelling is om het voorspellingsmodel te verbeteren ten opzichte van het huidige model. Hiermee verlaagt Eneco Business hun onbalanskosten, maar voor EET levert het geen kostenbesparing op. Sterker nog, wat Eneco Business minder betaalt, komt minder bij EET binnen terwijl de kosten voor EET hetzelfde blijft. Hierdoor is het mogelijk dat de relatie tussen EET en Eneco Business op scherp gezet wordt.

Om ten opzichte van de concurrenten innovatief en onderscheidend te zijn, moeten Eneco Business en EET samenwerken om de voorspelling nauwkeuriger te maken. Een nieuwe manier om mogelijk de onbalanskosten te verlagen, is het vormen van klantgroepen. Deze groepen bevatten klanten met dezelfde eigenschappen waardoor de groep beter voorspelbaar is dan het totaalverbruik. De som van de groepsvoorspellingen zou beter moeten zijn dan de totaalvoorspelling. Doordat er weinig bekend is over dit onderdeel, is Eneco Business geïnteresseerd in de mogelijkheid de groepen samen te stellen en in de resultaten die daarmee behaald kunnen worden. Deze vorm van kosten verlagen geldt zowel voor EET als voor Eneco Business wat de SLA gesprekken kan versoepelen als Eneco Business de onbalanskosten wilt verminderen.

Naast Eneco Business zijn de klanten zelf ook op zoek naar kostenbesparingen. Door te onderhandelen over de onbalanspremie proberen ze een lagere prijs per MWh te krijgen. Wegens de importantie van sommige klanten, kan Eneco Business toegeven aan een lagere onbalanspremie, die ze dan zelf moeten betalen. Dit komt omdat ze nu nog geen manier hebben om onbalanskosten aan klanten afzonderlijk toe te kennen. Een mogelijk bijkomend voordeel is dat als Eneco Business wel individuele onbalanspremies kan berekenen het klantenbestand kan verschuiven. Goed voorspelbare klanten kunnen een lagere onbalanspremie krijgen waardoor Eneco Business interessant wordt voor dit soort klanten. Daar tegenover staat dat minder goed voorspelbare klanten een hogere onbalanspremie kunnen krijgen waardoor dit soort klanten kan worden afgestoten. De verwachting is dat een klantenbestand met goed voorspelbare klanten zelf ook beter voorspelbaar is. Hierdoor kunnen de onbalanskosten omlaag gaan en uiteindelijk ook de onbalanspremie.

Over het algemeen kan gezegd worden dat Eneco Business onbalanskosten wilt verlagen. Ze willen de kosten verlagen door een lagere onbalanspremie in de SLA met EET op te nemen en door klanten een individuele onbalanspremie door te berekenen. De onbalanspremie verlagen kan gerealiseerd worden

op twee manieren. Door een voorspellingsmodel dat nauwkeuriger is dan het huidige en door klanten te groeperen zodat ze beter voorspelbaar zijn. De eerste stap om klanten een individuele onbalanspremie te laten betalen is om te achterhalen of de klanten te onderscheiden zijn op hun bijdrage van de onbalans.

#### **1.4 Relevante informatie over het onderzoeksbedrijf**

Dit onderzoek wordt gedaan in opdracht van Hot ITem, een Amsterdams projectenbureau dat als doelstelling het blijvend verbeteren van de prestaties van bedrijven en overheidsinstanties heeft. Dit realiseert Hot ITem door middel van performance improvement projecten op het snijvlak van business en IT. Eneco is één van de klanten van Hot ITem uit de energiebranche. De heer J. de Jonge is een medewerker van Hot ITem en wordt beschouwd als een expert op het gebied van de energiemarkt. In dit onderzoek vertolkt hij zowel de rol van opdrachtgever vanuit Eneco, als van de begeleider vanuit Hot ITem.

De meeste energie en Eneco specifieke informatie is door middel van interviews met de heer De Jonge en de websites van TenneT en EET verkregen. Uitgebreidere informatie over de energiemarkt is verkregen door het boek van Burger (2007) over de energiemarkt en de verbonden risico's. Voor verdere uitleg en informatie over de energiemarkt, als dit verslag niet duidelijk of volledig genoeg is, wordt naar dit boek verwezen.

#### **1.5 Structuur van het verslag**

Het resterende deel van dit verslag is als volgt georganiseerd. In het volgende hoofdstuk wordt uitgelegd hoe de data eruit zien waar dit onderzoek mee gedaan is. Hoofdstuk drie geeft een beschrijving van de literatuur. Als eerste worden hier verschillende onderzoeken behandeld, waarna de twee beste modellen verder apart worden uitgewerkt. In hoofdstuk vier wordt beschreven hoe het onderzoek tot stand is gekomen. Het hoofdstuk is in drieën gesplitst met elk hun eigen onderzoeksvraag. Eerst wordt het voorspellingsmodel beschreven waarna de klantgroepering aan de beurt is. Als laatste onderdeel van dit hoofdstuk komt de individuele klantkennis aan bod. Elk deel van dit hoofdstuk beschrijft de gevolgde procedure en de beperkingen die dit onderzoek heeft. Hoofdstuk vijf laat de resultaten zien van de drie onderdelen van dit onderzoek. In hoofdstuk zes worden de behaalde resultaten vergeleken met de praktijk. Hoofdstuk zeven bevat de conclusies die getrokken kunnen worden aan de hand van de gevonden resultaten. In hoofdstuk acht worden mogelijke vervolg onderzoeken aangedragen die verder gaan op dit onderzoek. Tot slot beschrijft hoofdstuk negen een evaluatie over hoe het onderzoek is verlopen. Om de leesbaarheid te vergroten is in bijlage C is een lijst opgenomen met de afkortingen die in dit verslag aan bod komen.





## 2 Dataverzameling

[VERTROUWELIJK]

In de data kunnen er zowel positieve als negatieve waarden voor het energieverbruik optreden. Negatieve waarden betekenen dat er energie van het net wordt afgenomen en positieve waarden betekenen dat er energie aan het net wordt toegevoegd door de klant. Over het algemeen is het verbruik van de klanten die energie afnemen in absolute zin groter dan de waarden van de klanten die energie toevoegen.

Elke klant heeft zijn eigen afnamepatroon en is afzonderlijk te voorspellen. Groepen klanten zijn te voorspellen door het gesommeerde verbruik te voorspellen. Het patroon van het totaalverbruik heeft een duidelijk dag- en weekpatroon. Het patroon over het gehele jaar is in vergelijking met het dag- en weekpatroon nihil. Het dagpatroon is te herkennen aan de daluren gedurende de avond en nacht en de piek uren overdag. Het weekpatroon is te herkennen aan de doordeweekse dagen en het weekend. De zakelijke klanten van Eneco Business verbruiken over het algemeen meer energie door de week dan in het weekend. Maandag is de start van de werkweek, dinsdag, woensdag en donderdag hebben ongeveer een vergelijkbaar dagpatroon en vrijdag is de afsluiting van de werkweek. Zaterdag is het verbruik een stuk lager dan het gemiddelde doordeweekse verbruik en zondag is het verbruik nog een extra stukje lager. Figuur 2 laat het dag- en weekpatroon zien van het totaalverbruik gedurende twee weken.



Figuur 2: Dag- en weekpatroon van het totaalverbruik gedurende twee weken

Door de grote hoeveelheid datapunten, moet er een aantal stappen doorlopen worden om de data geschikt te maken voor het onderzoek. De afzonderlijke klantfiles worden ingelezen met SQL server 2008 en getransporteerd naar een SQL Server 2008 Analysis Services Server. Via een dataconnectie met PowerPivot op deze server is het verbruik bereikbaar voor Excel spreadsheets. Met draaitabellen in Excel kan het gewenste verbruik geladen worden om onderzoek op uit te voeren.

### 3 Bespreking van de literatuur

Voorspellingen over het verbruik van elektriciteit, zijn een veel besproken onderwerp in de wetenschappelijke literatuur. Wegens de grote hoeveelheid aan literatuur, kunnen we het ons veroorloven de lange termijn voorspelling buiten beschouwing te laten. We beperken ons in de bespreking van de literatuur tot de korte termijn voorspelling van het verbruik van elektriciteit.

In de literatuur worden meerdere modellen behandeld die van toepassing zijn op de eerste onderzoeksvraag van dit onderzoek. Deze modellen worden in dit onderzoek met elkaar vergeleken om zo het beste model te achterhalen om het verbruik te voorspellen. Hoewel het huidige model niet in de literatuur beschreven wordt, nemen we deze wel mee in de vergelijking.

Om tot een goede keuze te komen over welke modellen onderzocht moeten worden is er een lijst met kandidaat modellen opgesteld. Deze lijst is voortgekomen uit een eerste selectie op basis van beschikbaarheid en verwachte prestaties van de modellen uit de gelezen teksten. Zo zijn de meest eenvoudige modellen vanwege te lage verwachte prestaties weggelaten. Het huidige n-7 model is er als laatste bijgevoegd.

Lijst van kandidaat modellen:

- Exponential Smoothing
- ARIMA
- Neurale netwerken
- Meervoudige regressie analyse
- Principal Component Analysis
- n-7

Om een keuze te kunnen maken uit de kandidaat modellen, is er onderstaande lijst criteriapunten opgenomen om de verschillende modellen op te beoordelen. De modellen zijn met behulp van de literatuur te beoordelen op deze criteriapunten. De scores worden in de elke alinea apart weergegeven, maar aan het einde samengevat in Tabel 2. De totale waarde van een model geeft de uiteindelijke score aan. Op basis van deze score wordt een model gekozen om verder te onderzoeken. Hier wordt het vervolg van het onderzoek op gebaseerd. Hierbij moet worden opgemerkt dat er alleen gekeken wordt naar enkelvoudige modellen. Hoewel meervoudige of gecombineerde modellen over het algemeen beter presteren dan enkelvoudige modellen, vallen meervoudige modellen niet binnen dit onderzoek. De reden hierachter is dat Eneco Business een eenvoudig model zoekt.

Lijst van criteriapunten:

- Accuraatheid
- Eenvoud
- Robuustheid
- Inspanning
- Geschiktheid

Het belangrijkste criterium waar de modellen op beoordeeld worden is accuraatheid: ofwel de nauwkeurigheid waarmee het model kan voorspellen. Eenvoud is de wiskundige moeilijkheid die bij het ontwikkelen en gebruiken van het model hoort. Dit is een belangrijke eis van Eneco Business. Robuustheid wilt zeggen dat ondanks dat het model niet helemaal juist wordt gebruikt het toch goed moet blijven presteren. Daarnaast wilt het zeggen dat het model goed blijft presteren ondanks de niet

helemaal optimaal geschatte invoerparameters. Robuustheid houdt ook in dat het model goed moet blijven presteren als niet helemaal aan de aannames voldaan wordt. Inspanning behelst de moeite en tijd die het kost om met het model te werken. Geschiktheid wilt zeggen of het model voldoet aan de randvoorwaarden om een voorspelling te kunnen geven.

### 3.1 Criteria uit de literatuur

Naast de criteriapunten die worden beschreven, wordt er in dit stuk ook een kleine samenvatting gegeven van het onderzoek. Hierin wordt beschreven hoe het onderzoek is gedaan, met welke modellen, op basis van welke data, in welke periode en met welke voorspellingshorizon. Onder het laatste wordt verstaan hoe ver in de toekomst voorspeld is. Een voorspellingshorizon van 24 uur wilt bijvoorbeeld zeggen dat de volgende dag voorspeld moet worden met de gegevens die op dat moment beschikbaar zijn.

Taylor (2003) vergelijkt een ARIMA (Autoregressive Integrated Moving Average) model met een exponential smoothing model. Het ARIMA model verwerkt net als het exponential smoothing model een dubbel seizoenspatroon. Dit zijn de dagelijkse en wekelijkse patronen van het verbruik. Het onderzoek is gedaan op basis van gegevens uit 2000 over het verbruik van energie per halfuur in Engeland en Wales. De voorspellingshorizon is van het volgende halfuur tot en met de volgende dag. Naast het standaard exponential smoothing model is er ook onderzoek gedaan naar een kleine toevoeging aan dit model. Hierin wordt een correctie op de voorspelfouten gemaakt met een autoregressie. Met deze toevoeging presteert dit nieuwe model beter dan het standaard exponential smoothing en het ARIMA model voor elke voorspellingshorizon. Het ARIMA model presteert tot bijna een dag van te voren beter dan het standaard exponential smoothing model, maar na 22,5 uur niet meer. Exponential smoothing is een redelijke kandidaat om de voorspeling uit te voeren wegens de robuustheid van dit model volgens Taylor (2003). Als aanbeveling wordt meegenomen dat hoewel het exponential smoothing model met autoregressie toevoeging het beste presteert, in de praktijk beter gebruik kan worden gemaakt van meerdere verschillende modellen. Mede door deze punten wordt exponential smoothing gewaardeerd met een vijf voor robuustheid en een vier voor accuraatheid. Het ARIMA model krijgt mede hierdoor een vier op accuraatheid en een twee voor robuustheid. Volgens Taylor (2003) is ARIMA een populair model en is exponential smoothing een veelbelovende nieuwkomer. Hierdoor scoren ARIMA en exponential smoothing een vier op geschiktheid.

In de tekst van Darbellay en Slama (2000) worden niet-lineaire modellen vergeleken met lineaire modellen. Voor niet-lineaire modellen gebruiken ze een neuraal netwerk als referentie en voor lineaire modellen gebruiken ze een ARIMA model. Het onderzoek is gedaan naar het energieverbruik per uur in de Tsjechische Republiek over de periode van 1994 tot en met 1995. Er wordt opgemerkt dat de gegevens een jaarlijks, wekelijks en een dagelijks patroon hebben. De voorspellingshorizon is van het volgende uur tot en met 36 uur in de toekomst. Naast het standaard neuraal netwerk en de ARIMA modellen, wordt er ook bekeken of temperatuur invloed heeft op de prestaties van beide modellen. Het blijkt dat beide modellen gelijkwaardig zijn, zowel met als zonder toevoeging van de temperatuur. De auteurs merken op dat lineaire modellen de voorkeur genieten boven niet-lineaire modellen vanwege de complexiteit van niet-lineaire modellen zonder verbeterd resultaat. Mede door deze punten worden neurale netwerken gewaardeerd met een één voor eenvoud en inspanning, maar toch met een drie voor accuraatheid. Het ARIMA model wordt gewaardeerd met een vier voor accuraatheid en een drie voor eenvoud en inspanning.

Ramanathan et al. (1997) hebben meervoudige regressieanalyse getoetst aan meerdere andere modellen. De analyse werd uitgevoerd op data van twee elektriciteitsmaatschappijen over twee

periodes van 1990 tot 1992. Dit onderzoek heeft een voorspellingshorizon van 16 uur tot en met 40 uur. In het gebruikte regressiemodel worden verschillende variabelen meegenomen: datum en tijd, temperatuur, huidige vraag en voorspellingsfouten. Ondanks het feit dat het model verkregen wordt via een kleinste kwadraten fout, is het een model geworden met vele variabelen. Daarnaast moet voor elk uur apart een model gemaakt worden om de uren te voorspellen. Verder wordt er ook onderscheid gemaakt tussen weekenddagen en doordeweekse dagen. Het komt er op neer dat een meervoudige regressieanalyse weliswaar een simpel en flexibel model is, maar dat de hoeveelheid parameters een probleem kan zijn. Hierdoor scoort het model een vier op eenvoud, maar een één op inspanning. Daarnaast scoort het model een drie voor robuustheid.

Soares en Medeiros (2008) vergelijken in hun onderzoek een meervoudig model met een ARIMA model en een neuraal netwerk. Ze voeren hun onderzoek uit op basis van gegevens uit het zuidoosten van Brazilië in de jaren 1990 tot en met 2000. Ze hebben een voorspellingshorizon van een dag tot en met een week. Het meervoudig model bestaat uit twee stappen. De eerste stap is een deterministische stap. De tweede stap is een lineaire autoregressie model. Voor ieder uur wordt er een apart voorspellingsmodel gemaakt. ARIMA krijgt de voorkeur boven het meervoudige model wegens de slechtere prestaties en de complexiteit van het meervoudige model. Ondanks dat ze gelijkwaardig presteren, zijn neurale netwerken ook te complex om de voorkeur te krijgen boven het ARIMA model. Hierdoor scoort een neuraal netwerk een drie op accuraatheid, maar krijgt ARIMA een vier. Daar staat tegenover dat een neuraal netwerk een één scoort op eenvoud en inspanning. ARIMA wordt met een drie gewaardeerd op deze punten.

Vier verschillende modellen worden vergeleken in de tekst van Taylor et al. (2006). De modellen zijn: Exponential smoothing, principal component analysis (PCA), ARIMA en neurale netwerken. In dit onderzoek worden voorspellingen gedaan op basis van gegevens uit zowel Engeland en Wales als Brazilië. De gegevens uit Engeland en Wales zijn per half uur en komen uit het jaar 2000. De gegevens uit Brazilië zijn per uur en komen uit 1996. Het onderzoek heeft een voorspellingshorizon van een half uur tot en met een dag. Van de vier modellen presteren exponential smoothing en PCA even goed. Het ARIMA model doet het iets minder goed en het neurale netwerk presteert het minst goed van de vier modellen. Mede hierdoor krijgen exponential smoothing, PCA en ARIMA een vier voor accuraatheid, maar krijgt neurale netwerken een drie. Over het exponential smoothing model wordt gezegd: *“In addition to its forecasting performance, it is important to note that, of the four sophisticated methods that we considered, this method is comfortably the simplest and quickest to implement”* Taylor et al. (2006, , blz. 14-15). Hierdoor scoort exponential smoothing een vier op eenvoud en inspanning. Neurale netwerken hebben twee vaak voorkomende valkuilen. Namelijk overfitting en een complexe architectuur. Daarnaast presteert een neuraal netwerk waarschijnlijk pas goed als er veel data beschikbaar is en de data in verschillende datasets wordt opgesplitst. Dit verklaart mede de één op eenvoud en inspanning. Ook geven Taylor et al. (2006) aan dat hun data lineair zijn en dat een niet-lineair model, zoals een neurale netwerk, niet goed geschikt is voor dit soort onderzoek. Hierdoor scoort het een twee voor geschiktheid. Wegens het niet verder beschrijven van het PCA model wordt deze gewaardeerd met een twee op geschiktheid en gelijk aan het ARIMA model voor eenvoud en inspanning. Er wordt niks gezegd over de robuustheid van de modellen waardoor PCA en neurale netwerken dezelfde score krijgen als het ARIMA model, namelijk een twee.

Als er externe variabelen worden meegenomen in de voorspelling naar het verbruik van elektriciteit, is temperatuur het meest gebruikt volgens Peirson en Henley (1994). Zelf onderzoeken ze of temperatuur geschikt is als externe variabele in een meervoudige regressieanalyse. Ze komen tot de conclusie dat de impact van temperatuur niet goed te gebruiken is voor de voorspelling. Temperatuur is te sterk

gecorrleerd aan de maand van het jaar en het uur van de dag om nog extra informatie te bevatten. Datum en tijd zijn sterkere voorspellers dan temperatuur. Hierdoor krijgt meervoudige regressieanalyse een twee op geschiktheid en accuraatheid.

Voor het huidige n-7 model is er geen literatuur beschikbaar. Dit model wordt op basis van ervaring door dit onderzoek beoordeeld. Het huidige model krijgt een één voor accuraatheid wegens de onnauwkeurigheid van het model. Het model scoort een 5 voor eenvoud en een vier voor inspanning. Dit komt doordat er geen wiskundig aspect in zit en weinig tijd en moeite kost om ermee te werken. Op robuustheid scoort het model een twee wegens het gebrek aan lerend vermogen. Mede door de voorspellingshorizon van een week en de beschikbaarheid van recentere data scoort dit model een twee op geschiktheid.

Om een overzicht te geven van de prestaties van de hierboven beschreven modellen, kunnen de mean absolute percentage error (MAPE) van de verschillende modellen met elkaar vergeleken worden. Om de accuraatheid van een model te waarderen moet ook de voorspellingshorizon meegenomen worden. Wegens de ongelijke voorspellingshorizon krijgen de modellen met de kleinste MAPE uit Tabel 1 niet direct de hoogste beoordeling voor accuraatheid in Tabel 2. In dit onderzoek moet de voorspelling voor de volgende dag worden verkregen aan de hand van het verbruik van de vorige dag. Effectief komt het er op neer dat dit onderzoek een voorspellingshorizon heeft van 24 uur tot en met 48 uur. Hierdoor weegt de MAPE van voorspellingen die in de buurt liggen van deze horizon zwaarder mee dan de MAPE van voorspellingen die verder uit de buurt liggen.

**Tabel 1: Prestatie per model uit de literatuur**

Model	MAPE	Horizon
Exponential Smoothing	1,3% - 3,0%	24 uur
ARIMA	1,7% - 3,0%	24 uur
Neurale Netwerken	2,1% - 4,2%	24 - 40 uur
Meervoudige Regressie Analyse	2,5% - 3,2%	16 - 40 uur
Principal Component Analysis	1,5% - 3,0%	24 uur
n-7	2,6% - 3,5%	168 uur

Om een eenvoudig overzicht te hebben van de verschillende modellen beoordeeld op verschillende criteria, zijn de scores in Tabel 2 weergegeven. De scores zijn afgeleid uit de teksten en lopen van 1 tot en met 5. De score 1 staat voor zeer matig of zwak. Een 5 houdt een sterke of uitstekende score in. Naast de scores van de modellen op de criteria, is er ook een totaal score aangegeven. Om tot deze totaalscore te komen zijn de verschillende scores op de criteria bij elkaar opgeteld. Hoe hoger de score, hoe beter het model. In Tabel 2 staan de modellen gerangschikt op totaalscore. De beste modellen boven en de minder goed onder in de tabel.

Tabel 2: Modellen met totaalscore en selectiecriteria

Model	Totaal	Accuraat	Eenvoud	Robuust	Inspanning	Geschikt
Exponential Smoothing	21	4	4	5	4	4
ARIMA	16	4	3	2	3	4
Principal Component Analysis	14	4	3	2	3	2
n-7	14	1	5	2	4	2
Meervoudige Regressie Analyse	12	2	4	3	1	2
Neurale Netwerken	9	3	1	2	1	2

## 3.2 Diepere uitwerking van de beste modellen

Met de uitkomst van Tabel 2 komen exponential smoothing en het ARIMA model als beste uit de bus. Beide modellen zullen nader worden onderzocht.

### 3.2.1 Exponential smoothing

Voor de basisprincipes van exponential smoothing wordt er verwezen naar Makridakis (1998). Hierin wordt uitgelegd wat exponential smoothing inhoudt en hoe het is opgebouwd. De teksten van Taylor (2003) en Taylor et al. (2006) gaan dieper in op het exponential smoothing model. De grootste verdieping is de mogelijkheid om dubbele seizoenspatronen mee te nemen. Dit model wordt het double seasonal exponential smoothing model genoemd. Taylor et al. (2006) hebben op dit model nog een kleine aanvulling gegeven door een extra  $\phi$  term toe te voegen. In de tekst van Taylor (2003) wordt er een uitgebreide beschrijving gegeven om de start waardes te initialiseren. De voorspelling met het exponential smoothing model wordt gegeven door de volgende formules:

$$S_t = \alpha(y_t / (D_{t-s_1} W_{t-s_2})) + (1 - \alpha)(S_{t-1} + T_{t-1})$$

Vergelijking 1

$$T_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)T_{t-1}$$

Vergelijking 2

$$D_t = \delta(y_t / S_t W_{t-s_2}) + (1 - \delta)D_{t-s_1}$$

Vergelijking 3

$$W_t = \omega(y_t / S_t D_{t-s_1}) + (1 - \omega)W_{t-s_2}$$

Vergelijking 4

$$\hat{y}_t(k) = (S_t + kT_t)D_{t-s_1+k}W_{t-s_2+k} + \phi^k(y_t - ((S_{t-1} + T_{t-1})D_{t-s_1}W_{t-s_2}))$$

Vergelijking 5

$S_t$  geeft het basisniveau aan en  $T_t$  de trend;  $D_t$  en  $W_t$  zijn de seizoensperiode indicatoren voor respectievelijk dag- en weekpatronen;  $s_1$  en  $s_2$  zijn de lengtes van de periodes van respectievelijk het dag- en weekpatroon;  $\alpha$ ,  $\gamma$ ,  $\delta$  en  $\omega$  zijn de smoothing parameters;  $y_t$  is het werkelijke verbruik op tijdstip  $t$ ; en  $\hat{y}_t(k)$  is de voorspelling op tijdstip  $t$  met horizon  $k$ . De toevoeging bij de schatter met de  $\phi$  parameter is de uitbreiding van Taylor et al. (2006) ten opzichte van de Taylor (2003) formules en is om de eerste autocorrelatie mee te nemen.

De initiële trend,  $T_0$ , wordt gekozen als het gemiddelde van: (1)  $1/s_2$  van het verschil tussen het gemiddelde van de eerste en tweede week en (2) het gemiddelde van de eerste differentie van de eerste week. Het initiële basisniveau,  $S_0$ , wordt gekozen als het gemiddelde van de eerste twee weken minus  $(s_2 + 0.5) \times T_0$ . De initiële waarden van het dagpatroon worden gezet op het gemiddelde van de eerste zeven dagen van de ratio van de actuele waarde met een  $s_1$ -punts gecentreerde moving average (MA). De initiële waarden van het weekpatroon worden gezet op het gemiddelde van de eerste twee weken van de overeenkomstige periodes van dezelfde dag van de ratio tussen de actuele waarden en de  $s_2$ -punts gecentreerde MA, gedeeld door de overeenkomstige initiële waarden van het dagpatroon,  $D_t$ .

Door eenmalig de som van de kwadratische error te minimaliseren, worden de parameters geschat en wordt het uiteindelijke model voltooid.

### 3.2.2 ARIMA

De teksten van Taylor et al. (2006) en Taylor (2003) gebruiken vergelijkbare methodes om aan een ARIMA model te komen. Het verschil tussen beide modellen komt voort uit het feit dat het op basis van andere data zijn uit gevoerd. Uit beide teksten kunnen we een multiplicative double seasonal ARIMA model destilleren. Dit model wordt als volgt beschreven.

$$\phi_p(L)\Phi_{P_1}(L^{s_1})\Omega_{P_2}(L^{s_2})\nabla^d\nabla_{s_1}^{D_1}\nabla_{s_2}^{D_2}(y_t - c) = \theta_q(L)\Theta_{Q_1}(L^{s_1})\Psi_{Q_2}(L^{s_2})\varepsilon_t$$

Vergelijking 6

Hier is  $y_t$  de vraag in periode  $t$ ;  $c$  is een contante term;  $s_1$  en  $s_2$  zijn de lengtes van de periodes van respectievelijk dag- en weekpatroon;  $L$  is de lag operator;  $\nabla$  is de difference operator;  $\nabla_{s_1}$  en  $\nabla_{s_2}$  zijn de seizoensdifference operators;  $d$ ,  $D_1$  en  $D_2$  zijn de orders van differencing;  $\varepsilon_t$  is een witte ruis term; en  $\phi_p$ ,  $\Phi_{P_1}$ ,  $\Omega_{P_2}$ ,  $\theta_q$ ,  $\Theta_{Q_1}$  en  $\Psi_{Q_2}$  zijn de polynomiale functies van respectievelijk order  $p$ ,  $P_1$ ,  $P_2$ ,  $q$ ,  $Q_1$ ,  $Q_2$ . Dit model kan uitgedrukt worden als  $ARIMA(p, d, q) \times (P_1, D_1, Q_1)_{s_1} \times (P_2, D_2, Q_2)_{s_2}$ . Taylor (2003) komt voor de vraag per half uur in Engeland en Wales uit op een  $ARIMA(2,0,0) \times (2,0,1)_{48} \times (2,0,2)_{336}$ . Om een voorbeeld te geven is Vergelijking 7 de uitgewerkte versie van Vergelijking 6 van het zojuist beschreven model met  $s_1 = 48$  en  $s_2 = 336$ .

$$(1 - 1,02L + 0,08L^2)(1 - 0,32L^{48} - 0,31L^{96}) \times (1 - 0,16L^{336} - 0,44L^{672})(y_t - 21625) \\ = (1 - 0,16L^{96})(1 - 0,15L^{336} - 0,19L^{672})\varepsilon_t$$

Vergelijking 7

Uit het onderzoek van Taylor et al. (2006) komt voor de vraag per uur in zuid-oost Brazilië een  $ARIMA(3,0,3) \times (3,0,3)_{24} \times (3,0,3)_{168}$ . In hetzelfde onderzoek, maar dan voor de vraag per half uur in Engeland en Wales, komen ze uit op een  $ARIMA(2,0,1) \times (2,0,1)_{48} \times (1,0,2)_{336}$ . In beide teksten en voor alle modellen wordt niet precies beschreven hoe en waarom ze op deze modellen komen. Er wordt wel beschreven dat via de Box-Jenkins methodologie de meest bruikbare modellen verkregen zijn en dat de autocorrelatie functie (ACF) en partiële autocorrelatie functie (PACF) gebruikt zijn om de orders van het model te bepalen. Het model met de laagste Schwartz Bayesian information Criterion (SBC) met toegestane voorspelfouten is dit het uiteindelijke model geworden. Logaritmische en gedifferentieerde transformaties leverden geen verbetering op in de SBC.

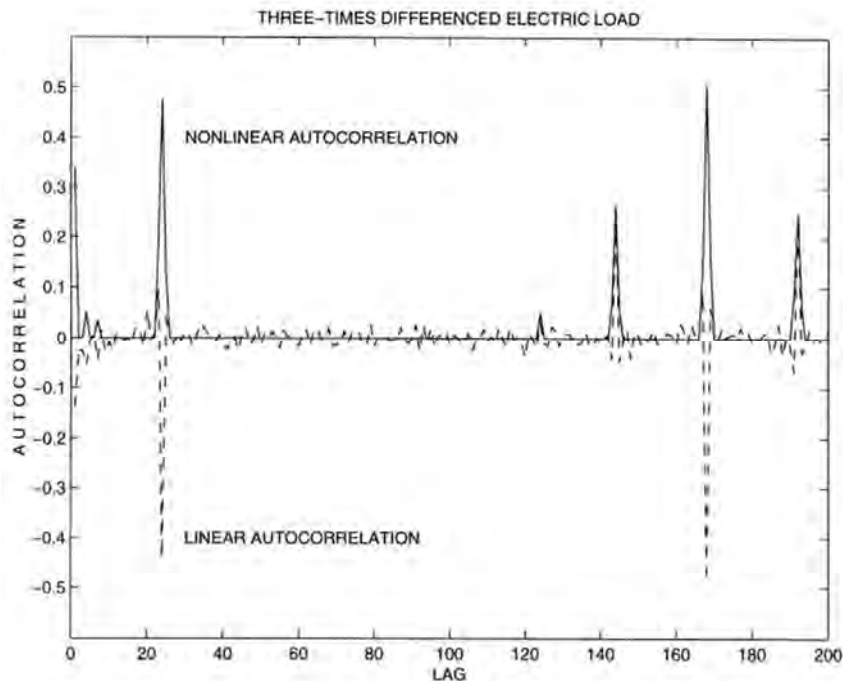
In de tekst van Darbellay en Slama (2000) wordt op een andere manier een ARIMA model verkregen. Hier worden de stappen die genomen worden om tot het model te komen uitgebreid uitgelegd waardoor het geheel betrouwbaarder lijkt. Wegens de eigenschap van niet stationaire data en de



seizoenperiodes is de data drie keer gedifferentieerd. De data worden als volgt gedifferentieerd waarbij de originele data zijn:

$$\begin{aligned}u_t &= y_t - y_{t-1} \\v_t &= u_t - u_{t-24} \\w_t &= v_t - v_{t-168}\end{aligned}$$

Na de eerste keer gedifferentieerd te hebben, blijven alleen de dagelijkse en wekelijkse autocorrelaties over. Na de tweede en derde differentiatie om respectievelijk de dag- en weekpatronen eruit te halen, blijft de autocorrelatie functie over uit Figuur 3.



Figuur 3: Autocorrelatie functie na drie keer gedifferentieerd van Darbellay en Slama (2000)

In Figuur 3 zien we vijf pieken. Deze pieken liggen op 1 week (lag = 168), 1 dag, 1 uur, 6 en 8 dagen. Verder dan lag = 200 blijkt er geen significante autocorrelatie meer te zijn en veronderstellen ze dat het verder een stationaire tijdreeks is. De lineaire autocorrelatie functie laat twee negatieve pieken zien. Het ARIMA model waarmee ze komen is een  $ARIMA(1,1,1) \times (1,1,1)_{24} \times (1,1,1)_{168}$ . De parameters worden geschat via de gegeneraliseerde kleinste kwadraten methode.

De hierboven beschreven teksten zijn relevant voor dit onderzoek omdat ze hetzelfde onderwerp hebben als dit onderzoek. De teksten zijn echter niet volledig over hun werkwijze, aangezien ze niet beschrijven hoe ze tot een voorspelling komen met een verdere horizon dan een stap vooruit. Evenmin geven ze aan op welke grond ze de keuze van hun modellen hebben gemaakt. Kortom, om toch een ARIMA model te kunnen gebruiken voor dit onderzoek is verdere studie naar ARIMA modellen noodzakelijk.

Om tot een ARIMA model te komen, moeten de data stationair zijn en moeten de voorspelfouten aan de eis van witte ruis voldoen volgens De Gunst (2009). De data zijn stationair als de huidige verwachtingswaarde en de verwachte waarde van een punt in de toekomst of in het verleden niet

afhankelijk zijn van de tijd. Dit moet gelden voor elk datapunt. Dat de voorspelfouten aan de eis van witte ruis moeten voldoen houdt in dat de foutterm van de voorspelling onafhankelijk en identiek verdeeld moet zijn volgens een normaalverdeling met gemiddelde nul en standaard deviatie sigma. Om een keuze te kunnen maken in de MA en AR order, wordt er gekeken naar de ACF en de PACF. Als de ACF een duidelijk aantal significante termen heeft, is dat de order van de MA. Als de PACF een duidelijk aantal significante termen heeft, is dat de order van de AR. Indien beide niet duidelijk zijn, worden beide orders bepaald aan de hand van de Akaike Information Criterion (AIC).

Als de data niet stationair zijn, kan stationaire data verkregen worden door middel van drie methodes. Via de small trend methode, filtering of differencing worden de trend en het seizoenspatroon uit de data gefilterd om zo stationaire data over te houden.

Nau (2005) heeft een duidelijk stappenplan om het juiste ARIMA model te identificeren. Hiervoor zijn dertien regels opgesteld om de data stationair te maken, de orders te bepalen voor de AR en MA term en om de seizoenspatronen mee te nemen. Een samenvatting van de regels is opgenomen in Bijlage A.

Met deze dertien regels is de informatie bijna volledig die nodig is voor dit onderzoek om een ARIMA model op te stellen. Tot nu toe is er informatie om één stap vooruit te voorspellen, terwijl dit onderzoek 24 uur tot 48 uur vooruit moeten voorspellen. Het probleem met ARIMA modellen is dat het de voorgaande realisatie en de voorspellingsfouten gebruikt. Echter, als er verder dan één stap voorspeld moet worden, is er een gat in de formule doordat de realisatie nog niet bekend is. Logischerwijs is de voorspelfout ook niet bekend. Hier hebben Cizek et al. (2005) een oplossing voor. Dit kan verholpen worden door de reeds voorspelde waardes te gebruiken als de realisatie nog niet bekend is en de voorspellingsfout gelijk te stellen aan nul. Vergelijking 8, Vergelijking 9 en Vergelijking 10 geven de wiskundige notatie hiervan.

$$\begin{aligned}\hat{y}_t(k) &= E[y_{T+k}|Y_T] \\ &= \phi_1 E[y_{T+k-1}|Y_T] + \dots + \phi_{p+d} E[y_{T+k-p-d}|Y_T] + E[\varepsilon_{T+k}|Y_T] \\ &\quad + \theta_1 E[\varepsilon_{T+k-1}|Y_T] + \dots + \theta_q E[\varepsilon_{T+k-q}|Y_T]\end{aligned}$$

Vergelijking 8

$$E[y_{T+j}|Y_T] = \begin{cases} y_{T+j} & j \leq 0 \\ y_T(j) & j > 0 \end{cases}$$

Vergelijking 9

$$E[\varepsilon_{T+j}|Y_T] = \begin{cases} \varepsilon_{T+j} & j \leq 0 \\ 0 & j > 0 \end{cases}$$

Vergelijking 10

Met deze informatie en de vergelijkingen om een verdere horizon te voorspellen wordt nu alleen nog maar de initiële waardes gemist. Om tot een volledige voorspelling te komen, wordt er gebruik gemaakt van achterwaarts voorspellen. Dit is nodig omdat de voorspelling gebruik maakt van eerdere voorspellingen. Echter, de eerste voorspelling zal er een zijn die zonder voorgaande voorspellingen voorspeld wordt. Gelukkig helpt hier de eigenschap van stationaire data. Stationaire data hebben hetzelfde verloop in de toekomst als in het verleden. Hierdoor kan met hetzelfde model het verleden voorspeld worden en kan er via iteraties uiteindelijk de juiste waardes gekregen worden.

## 4 Onderzoek ontwerp

In dit hoofdstuk wordt beschreven hoe dit onderzoek is verlopen en wat er is gedaan om tot een resultaat te komen. Het hoofdstuk is in drieën gedeeld met elk deel zijn eigen onderzoeksvraag. Als eerst wordt het voorspellingsmodel beschreven. Dit onderdeel is nog verder in een exponential smoothing en ARIMA paragraaf gesplitst. Als tweede wordt er een beschrijving gegeven over het onderzoek ontwerp over de onderzoeksvraag van de klantgroepering. Als derde onderdeel wordt er uitgelegd hoe de resultaten van de individuele klantkennis tot stand zijn gekomen. Elk deel of paragraaf bestaat uit een onderzoeksprocedure en een beschrijving van de beperking van het onderzoek.

### 4.1 Voorspellingsmodel

Aan de hand van Tabel 2 komen exponential smoothing en ARIMA als beste modellen naar voren. Beide modellen worden gebruikt voor dit onderzoek en worden vergeleken met het huidige  $n - 7$  model.

De data die gebruikt zijn voor dit deel van het onderzoek, is het verbruik van energie van alle klanten van Eneco Business over een periode van twaalf weken, lopend van 6 juni 2007 tot en met 2 september 2007. De eerste week is gebruikt voor de initialisatie fase. Daarna zijn zeven weken gebruikt om de parameters te schatten. De laatste 4 weken zijn gebruikt om het model te evalueren op robuustheid.

#### 4.1.1 Exponential smoothing

##### Onderzoeksprocedure

De volgende onderzoekstappen zijn doorlopen om tot de gevonden resultaten te komen. Deze stappen zijn voornamelijk gebaseerd op de tekst van Taylor (2003) en van Taylor et al. (2006).

De voorspelling is opgebouwd uit vier componenten. Deze componenten beschrijven het basisniveau, de trend en een dag- en weekindex. Deze worden gegeven door Vergelijking 1 tot en met Vergelijking 4 uit hoofdstuk 3. Vergelijking 5 is de voorspelling zelf voor het exponential smoothing model.

De initiële waardes worden verkregen aan de hand van de tekst van Taylor (2003) zoals in het vorige hoofdstuk uitvoerig is beschreven. Om dit te realiseren zijn de datapunten van één week nodig. Om de optimale initiële parameters van het model te verkrijgen wordt de som van de kwadratische fout geminimaliseerd. Deze som wordt verkregen door het verschil tussen de voorspelling en het werkelijke verbruik te kwadrateren over de eerste vier weken. Voor elke volgende dag worden de parameters opnieuw geschat om de voorspelling te krijgen door de voorspelfouten tot en met de twee dagen voor de voorspellende dag te minimaliseren. Dit gebeurt tot en met de achtste week. Een schatting van de robuustheid van het model wordt verkregen door de overige vier weken te voorspellen zonder de gevonden parameters te veranderen. Na afloop van de voorspelling worden de fouten van de voorspelling nog nader onderzocht.

Om het exponential smoothing model te kunnen vergelijken met het huidige model worden de MAPE van beide modellen berekend. Dit is een veel gebruikte eenheid om een uitspraak te doen over de kwaliteit van een voorspelling. Dit is ook de eenheid die in alle gelezen literatuur gebruikt wordt. De MAPE wordt gegeven door Vergelijking 11. Door de MAPE over de tweede tot en met achtste week te nemen wordt een indicatie verkregen over hoe accuraat het model is. Door de MAPE over de laatste vier weken te nemen, wordt een indicatie over de robuustheid van het model verkregen.

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

Vergelijking 11

Om de onbalanskosten te schatten die het exponential smoothing model zou opleveren, wordt er een paar aannames gedaan. Er wordt aangenomen dat de onbalansprijs vast is over elke periode. Hierdoor wordt er geen rekening gehouden met dure of goedkope tijden van elektriciteit of met voorspellingsoverschot of tekort. Verder wordt aangenomen dat de onbalanskosten zich evenredig verhouden met de MAPE. Dit houdt in dat een lage MAPE lage onbalanskosten veroorzaakt.

### Beperkingen van het onderzoek

Voor optimaal gebruik vereist het model regelmatig een update van de parameters. Hoe vaak deze update plaats moet vinden is niet bekend en hangt af van de snelheid van verandering in verbruik en de robuustheid van het model. Verder is de optimale periode waarover de parameters worden geschat niet bekend. Hoe meer data meegenomen worden, hoe geëffender en robuuster het model, maar ook hoe ouder de data. Te oude data hebben geen toegevoegde waarde en kan het model zelfs verslechteren door de grote veranderingen die ondertussen hebben plaatsgevonden. Te weinig data meenemen maakt het model volatiel, waardoor grote fouten kunnen ontstaan door toevalligheden.

Door de aannames die gedaan worden in dit onderzoek, kan de praktijksituatie anders zijn dan de resultaten die in dit onderzoek gehaald worden. Desondanks worden de aannames gebruikt om het onderzoek te vereenvoudigen en overzichtelijk te houden.

### 4.1.2 ARIMA

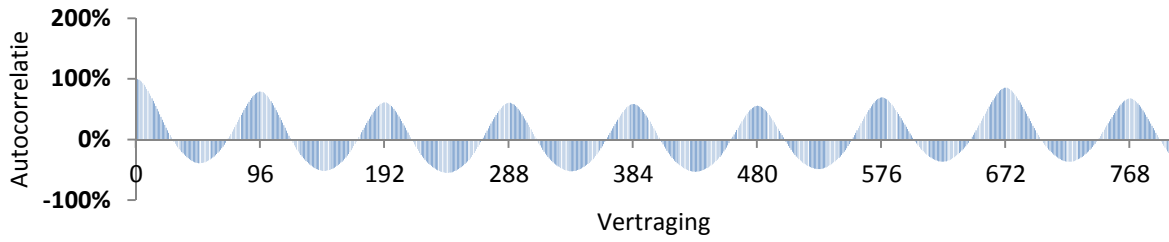
#### Onderzoeksprocedure

De volgende onderzoekstappen zijn doorlopen om tot de gevonden resultaten te komen. Deze stappen zijn voornamelijk gebaseerd op de dertien regels van Nau (2005) en de voorspellingshorizon techniek van Cizek et al. (2005). Daarnaast is de methode van Darbellay en Slama (2000) als basismethode gebruikt.

De eerste stap is om de ACF van de data te onderzoeken. De ACF van de originele data worden weergegeven in Figuur 4. Na één keer gedifferentieerd te hebben, wordt de ACF van Figuur 5 verkregen. Na de seizoendifferentiatie is Figuur 6 de ACF. Hierdoor wordt het dagpatroon verwijderd. Volgens regel 12 mag er maximaal 2 keer gedifferentieerd worden, maar uit de tekst van Darbeallay en Slama (2000) blijkt dat er nog een keer gedifferentieerd wordt. Na een derde keer differentiëren wordt de ACF uit Figuur 7 gevonden. Hierdoor wordt het weekpatroon verwijderd. Met een dagpatroon bestaande uit 96 kwartieren en een weekpatroon bestaande uit 672 kwartieren, wordt de originele data ( $y$ ) als volgt gedifferentieerd.

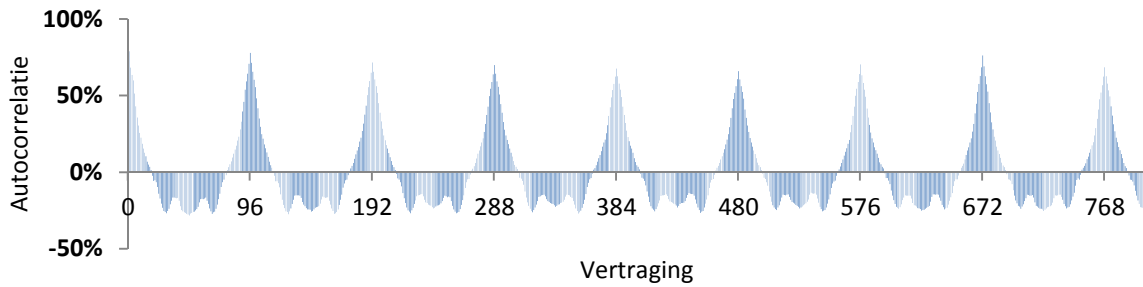
1.  $u_t = y_t - y_{t-1}$
2.  $v_t = u_t - u_{t-96}$
3.  $w_t = v_t - v_{t-672}$

### ACF op originele data



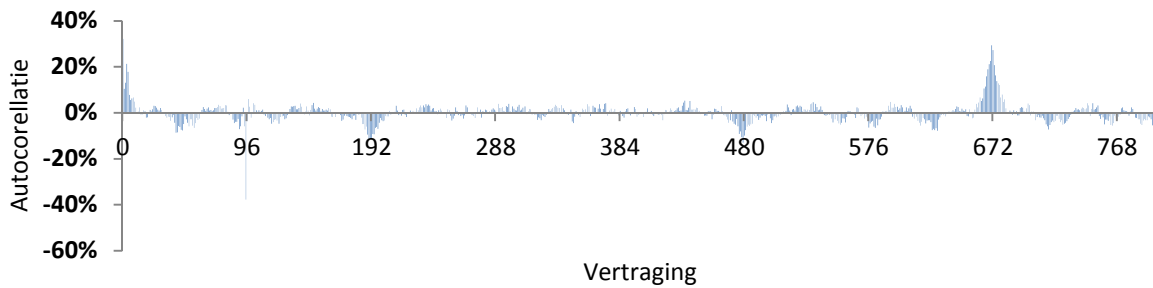
Figuur 4: Autocorrelatie functie over de originele data

### ACF na 1x differentiatie



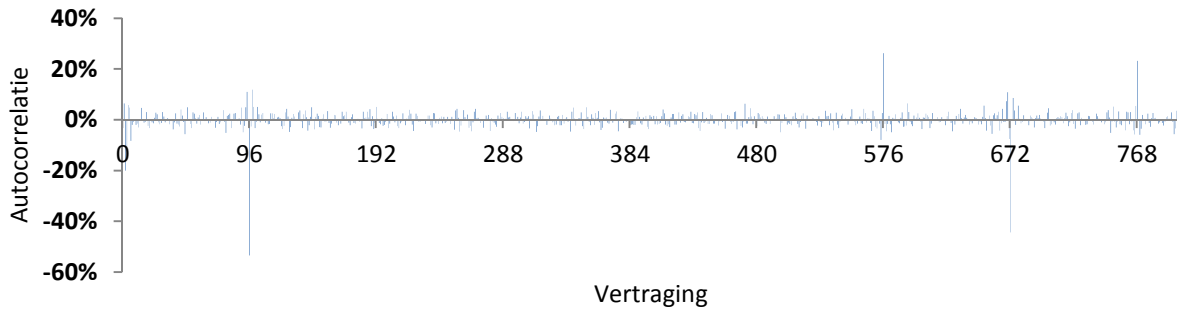
Figuur 5: Autocorrelatie functie over de data na 1x differentiëren

### ACF na 2x differentiatie



Figuur 6: Autocorrelatie functie over de data na 2x differentiëren

## ACF na 3x differentiatie



Figuur 7: Autocorrelatie functie over de data na 3x differentiëren

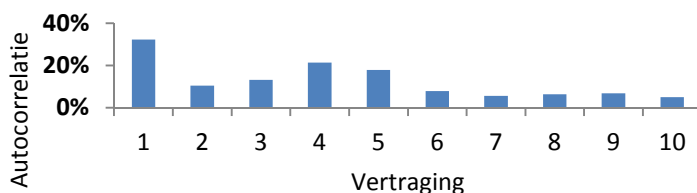
Volgens regel drie is de optimale orde van differentieën het aantal waarvan de standaard deviatie het laagst is. In Tabel 3 zijn de standaard deviaties opgenomen en blijkt dat twee keer differentieën optimaal is. Om de methode van Darbellay en Slama (2000) niet te passeren, wordt verder gewerkt met een differentie orde van twee en drie.

Tabel 3: Standaard deviaties van de verschillende datasets

Data	Standaard deviatie
Originele data	105.560
Na 1x differentieën	9.119
Na 2x differentieën	5.914
Na 3x differentieën	6.883

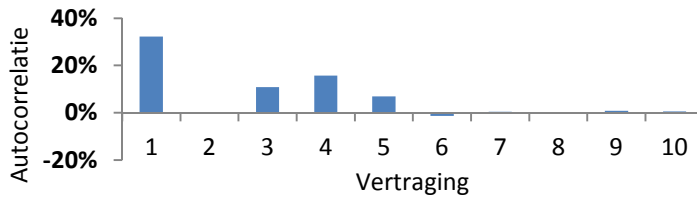
Figuur 8 laat de ACF zien en Figuur 9 laat de PACF zien voor de eerste 10 vertragingen nadat de data twee keer gedifferentieerd zijn. Figuur 10 laat de ACF zien en Figuur 11 laat de PACF zien na drie keer gedifferentieerd te zijn voor de eerste 10 vertragingen. Alle figuren laten zowel bij de PACF als de ACF geen scherpe afkapping zien, waardoor er volgens De Gunst (2009) zowel AR als MA termen gebruikt moeten worden en op basis van de AIC het beste model verkregen kan worden. Hierbij moet wel rekening gehouden worden met een unit root in zowel de AR als de MA term volgens regels 9, 10 en 11.

## ACF na 2x differentiatie



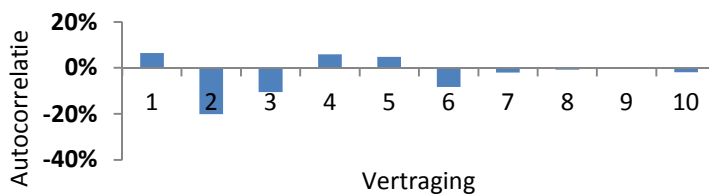
Figuur 8: Autocorrelatie functie over de eerste 10 vertragingen na 2x differentieën

## PACF na 2x differentiatie



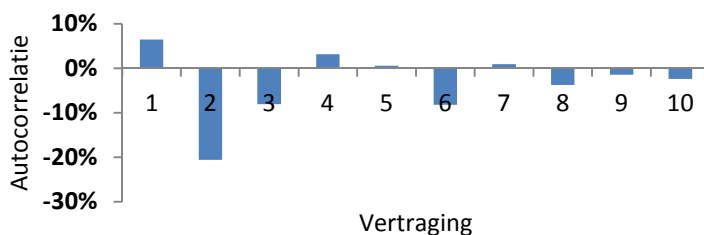
Figuur 9: Partiële autocorrelatie functie over de eerste 10 vertragingen na 2x differentieren

## ACF na 3x differentiatie



Figuur 10: Autocorrelatie functie over de eerste 10 vertragingen na 3x differentieren

## PACF na 3x differentiatie



Figuur 11: Partiële autocorrelatie functie over de eerste 10 vertragingen na 3x differentieren

Volgens regel dertien moet er een AR term komen voor de seizoensperiode als de ACF positief uitslaat. We hebben te maken met een MA term als de seizoensperiode van de ACF negatief uitslaat. Aan de hand hiervan is er na twee keer differentieren een MA term voor het dagpatroon en na drie keer differentieren een MA term voor de dag- en weekpatroon. Meer dan één orde van de seizoenterm moet voorkomen worden volgens de regels van Nau (2005). In de tekst van Darbellay en Slama (2000) gebruiken ze wel meerdere seizoenstermen en zelfs AR en MA termen door elkaar. Om bij voorbaat geen modellen uit te sluiten, wordt er naar een breed spectrum van mogelijke modellen gekeken, maar gaat de voorkeur uit naar een model volgens de regels van Nau (2005). Het voorkeursmodel is een  $ARIMA(2,1,2) \times (0,1,1)_{96}$ .

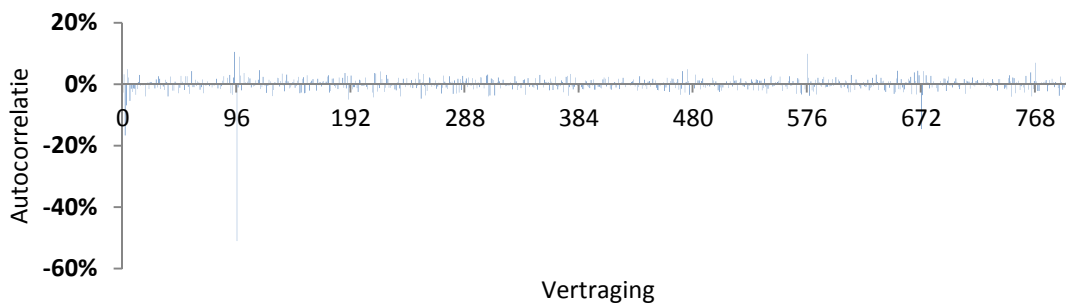
Via iteratief achterwaarts voorspellen worden de juiste waardes verkregen voor de eerste voorspellingen.

### Beperkingen van het onderzoek

Volgens De Gunst (2009) is het meest belangrijke van een ARIMA model de controle of de voorspelfouten voldoen aan de eis van witte ruis. Taylor et al. (2006) hebben het ook over toegestane modellen met betrekking tot voorspelfouten. In het geval dat de voorspelfout niet voldoet aan deze eis, is het model mogelijk niet geschikt. Juist deze eis is een probleem voor dit onderzoek. Er moet te ver vooruit voorspeld worden, zodat de hoogst gecorrleerde periode niet meegenomen kan worden. Het model moet 24 tot 48 uur vooruit voorspellen, terwijl de waarde van 24 uur het belangrijkste is om te voorspellen. Het dagpatroon kan niet meegenomen worden in de voorspelling. Hierdoor zijn de voorspelfouten teveel met elkaar gecorrleerd en voldoet het niet aan de eis van witte ruis.

Als voorbeeld is het  $ARIMA(2,1,0) \times (1,1,1)_{96} \times (1,1,1)_{672}$  model gegeven met de ACF van de voorspelfouten, zie Figuur 12. Hierin is te zien dat de voorspelfouten niet voldoen aan de eis van witte ruis door de grote correlatie op 24 uur. Verder zien we ook nog een correlatie op een week en zes en acht dagen, maar deze zijn minder groot en daardoor minder belangrijk. Dit geldt ook voor de correlatie van het eerste kwartier. Met een significantie niveau van 10% zouden de correlaties van het eerste kwartier, zes en acht dagen misschien nog kunnen, maar de correlatie van 24 uur is nauwelijks afgenomen en nog steeds ruim 50%. Dit is te hoog om te negeren.

### ACF van de voorspelfouten



Figuur 12: Autocorrelatie functie van de voorspelfouten van het  $ARIMA(2,1,0) \times (1,1,1)_{96} \times (1,1,1)_{672}$  model

Door verder naar de voorspelfouten te kijken, kan er misschien achterhaald worden of er meer onregelmatigheden in zitten. Uit Tabel 4 blijkt dat het gemiddelde -2 is. Met een significantieniveau van 5% is dit niet significant anders dan nul. De standaard deviatie is kleiner geworden dan de data na drie keer gedifferentieerd te zijn. Verder is de skew -0,07 en de excess-kurtosis 9,49. Deze zijn wel significant anders dan nul. De kwantielen zien er goed uit en komen overeen met de hierboven beschreven situatie. Dit betekent dat het minimum in absolute zin groter is dan het maximum en dat het eerste en derde kwantiel dicht bij het gemiddelde en de mediaan liggen. Naast dat Tabel 4 aangeeft dat er inderdaad geen sprake is van witte ruis, is er zelfs ook sprake van dat de voorspelfouten niet normaal verdeeld zijn. Wel is er een Autoregressive Conditional Heteroskedasticity (ARCH) effect. Dit houdt in dat er periodes zijn met relatief grote fouten, gevolgd door periodes met relatief minder grote fouten.



Tabel 4: Voorspelfouten analyse van het ARIMA(2,1,0) x (1,1,1)<sub>96</sub> x (1,1,1)<sub>672</sub> model

Descriptive Statistics		Significance Test			5,00%	Test	p-vlaue	Result?
		Target	p-vlaue	Different?		White-noise	0,00%	FALSE
<b>AVERAGE:</b>	-2	0,000	48,69%	FALSE		Normal Distributed?	0,00%	FALSE
<b>STD DEV:</b>	5.790					ARCH Effect?	0,00%	TRUE
<b>SKEW:</b>	-0,07	0,000	0,47%	TRUE				
<b>EXCESS-KURTOSIS:</b>	9,49	0,000	0,00%	TRUE				
<b>MEDIAN:</b>	-27							
<b>MIN:</b>	-59.801							
<b>MAX:</b>	50.058							
<b>Q 1:</b>	-2.749							
<b>Q 3:</b>	2.738							

Dit probleem geldt niet alleen voor het ARIMA model in het voorbeeld, maar voor alle ARIMA modellen die toegepast zijn op dit onderzoek. Voor elk ARIMA model is het dagpatroon alleen in het model op te nemen als voorspelling. In de voorspelling zit de aanname dat de voorspelfout nul is, terwijl dit in de praktijk bijna nooit het geval zal zijn.

De voorspelfout voldoet niet aan de eigenschap van witte ruis. Zonder deze eigenschap is het model niet goed toepasbaar en het blijkt dat het model met het iteratief achterwaarts voorspellen niet convergeert naar een optimale voorspelling. Dit komt waarschijnlijk wegens de grote correlatie met de dag en de voorspellingshorizon van een dag en langer is het niet mogelijk om dit op te lossen binnen de tijd die hiervoor staat. Daarnaast is het wel mogelijk om met exponential smoothing een resultaat te behalen dat volgens de literatuur gelijkwaardige verbetering biedt. Het exponential smoothing model heeft niet de eis dat de voorspelfout aan witte ruis moet voldoen. Hierdoor wordt het onderzoek naar het ARIMA model gestopt.

## 4.2 Klantgroepering

### Onderzoeksprocedure

Als eerst worden de data behandeld om het mogelijk te maken om de klanten te groeperen. Vervolgens worden de klanten gegroepeerd aan de hand van een Excel add-in. De groepen worden apart voorspeld en de resultaten worden vergeleken met voorspellingen over het totaalverbruik.

Voor dit deel van het onderzoek is gebruik gemaakt van de data van Eneco Business over het verbruik van alle klanten van 11 juni 2007 tot en met 3 juli 2007.

Om tot klantgroepen te komen moeten er een aantal stappen worden doorlopen. Deze stappen zijn gebaseerd op de data voorbereidingsstappen van Fischer (2004). De eerste stap om klanten te groeperen is om de uitschieterende waardes eruit te filteren. Nieuwe klant of juist vertrekkende klanten worden voor dit onderzoek aangemerkt als uitschieters. Klanten worden gezien als nieuwe dan wel vertrekkende klanten als er op de eerste dan wel laatste dag van de periode geen waardes worden gemeten. Een klant die zowel de eerste dag als de laatste dag van de periode geen waarde heeft, wordt als vanzelfsprekend ook niet meegenomen. Van de bijna 14.000 klanten zijn er ruim 2.400 klanten die de eerste en/of laatste dag geen waardes hebben. Deze klanten worden niet weggegooid maar komen in een aparte

groep van nieuwe en vertrekkende klanten. Zo kunnen ze alsnog meegenomen worden in voorspellingen zonder dat ze de verdere groepering verstoren.

Om de overige klanten met elkaar te kunnen vergelijken, moeten ze veralgemeniseerd worden. Per klant worden de volgende stappen doorlopen. Als eerst wordt de lineaire trend verwijderd. Met behulp van het data analysis pakket van Excel wordt via de multiple regression analysis de helling bepaald en van de data afgehaald. Ten tweede wordt van deze verkregen data het gemiddelde af getrokken om het vervolgens te delen door de standaard deviatie. Er is bewust gekozen om de dag- en weekpatroon niet van de data af te halen om de klanten met dezelfde patronen bij elkaar in de groep te zetten. Op deze manier maakt het niet uit hoe groot het gemiddelde verbruik is of hoe groot de patronen zijn. De klanten zijn naar verhouding bijgesteld waardoor ze met elkaar te vergelijken zijn qua verbruik.

Om via deze veralgemeniseerde data de klanten aan groepen toe te kennen is gebruik gemaakt van de data mining add-in van Excel. Deze add-in heeft een detect categories functie die klanten aan de hand van input variabele in verschillende groepen verdeeld. Er zijn vijf dagen gekozen om als input variabele de groepen mogelijk te maken. Dit is een maandag, woensdag, vrijdag, zaterdag en zondag. Er is gekozen om geen dinsdag en donderdag mee te nemen omdat het afname patroon van die dagen het meeste lijken op een al meegenomen dag, namelijk woensdag. De detect categories functie heeft de klanten verdeeld in elf verschillende groepen. De keuze voor elf groepen heeft de add-in zelf bepaald door die optie aan te vinken. Met behulp van PowerPivot is er een relatie aangemaakt om elke klant aan een groep toe te kennen.

Door deze toekenning kan er vanuit Excel met de draaitabel het verbruik van elke groep afzonderlijk opgevraagd worden. Elke groep kan afzonderlijk voorspeld worden via het exponential smoothing model. Door de voorspellingen van alle groepen bij elkaar op te tellen, is er een voorspelling over het totaalverbruik verkregen. De voorspelling over het totaalverbruik wordt ook rechtstreeks berekend via het exponential smoothing model. Als beide voorspellingen met elkaar vergeleken worden kan er bepaald worden of groepsvoorspellingen een verbetering opleveren ten opzichte van de totaalvoorspelling. De vergelijking wordt gedaan aan de hand van de eerder gebruikte MAPE.

Wegens de onvoorspelbaarheid van de groep nieuwe en verlatende klanten wordt er apart gekeken of de voorspellingen zonder deze groep beter zijn. De stappen van hiervoor worden doorlopen zonder de groep van nieuwe en verlatende klanten mee te nemen in de groepsvoorspelling en het totaalverbruik.

### **Beperkingen van het onderzoek**

De gevolgde procedure is niet de enige manier om tot groeeringen te komen. Ook staan de voorbereidende stappen niet vast. Het is niet bekend of eerder geprobeerd om te weten of deze stappen het gewenste resultaat opleveren.

Er is gekozen om groepen te maken aan de hand van het verbruikspatroon. Hierdoor zijn er dagen als input variabele gekozen. Wegens het werkbaar houden van het model is er gekozen om niet met meer dan vijf dagen als inputvariabele te werken. Elke dag heeft 96 kwartieren wat betekent dat er  $5 \times 96 = 480$  input variabelen zijn meegenomen om te groeperen. Een andere manier van groeperen kan betere resultaten opleveren, maar wegens het gebrek aan tijd is dit niet uitgebreider onderzocht.

Verdere analyse op de klanten is niet mogelijk wegens de geanonimiseerde data. Klanten hebben een nummer gekregen waardoor er niet bekend is welke klanten het zijn en of er enig verband in zit.

### 4.3 Individuele klantkennis

#### Onderzoeksprocedure

Aan de hand van een kleine subset wordt gekeken hoe de totaalvoorspelling van de subset opgebouwd is ten opzichte van de afzonderlijke klantvoorspellingen. In dit geval bestaat de subset uit vier willekeurig gekozen klanten. Per tijdseenheid wordt gekeken naar het verschil tussen de voorspelling en het werkelijke verbruik. Dit is het deel dat verantwoordelijk is voor de onbalans. Er wordt per klant onderzocht of de klantvoorspelling te hoog is geweest als de totaalvoorspelling groter is dan het werkelijke verbruik. Andersom wordt er ook gekeken of de klantvoorspelling te laag is geweest als de totaalvoorspelling kleiner is dan het werkelijke verbruik. Met dit verschil wordt het bijdragepercentage berekend over de klanten die aan de hier bovengenoemde situatie voldoen. Klanten die niet aan bovengenoemde situatie voldoen krijgen een bijdragepercentage van nul. Met het bijdragepercentage en het totale verschil is het mogelijk om een gewogen bijdrage per klant te bepalen per tijdseenheid. Door de bijdrage per klant te sommeren kan het totale gewogen bijdragepercentage berekend worden.

Een tweede manier is om het gemiddelde van de bijdragepercentages te nemen. De bijdragepercentages worden verkregen op dezelfde manier als zojuist is beschreven. Aan de hand van het gemiddelde kan het totale bijdragepercentage berekend worden. Het verschil tussen de eerste en deze manier is dat op de eerste manier het bijdragepercentage een gewogen gemiddelde is en de tweede manier een standaard gemiddelde is. Het gewogen gemiddelde wordt bepaald aan de hand van de werkelijke onbalans.

Een derde manier is om de MAPE per klant te bepalen door de klanten afzonderlijk te voorspellen. De MAPE per klant wordt verkregen door Vergelijking 11 toe te passen. Deze manier geeft een indicatie van de voorspelbaarheid voor iedere klant. Aan de hand van deze voorspelbaarheid is het mogelijk om een totale bijdragepercentage per klant te berekenen.

#### Beperkingen van het onderzoek

Er moet rekening gehouden worden met de schaal van het onderzoek. In de subset zijn maar vier klanten mee genomen, maar in werkelijkheid zijn het bijna 14.000 klanten. Daarnaast is voor dit gedeelte van het onderzoek zes uur meegenomen.

De eerste en tweede manier berekenen de bijdrage door de historie te bekijken. Deze manieren hebben echter geen voorspellende waarde wegens teveel afhankelijkheden. De bijdrage waarvoor de klanten verantwoordelijk waren in het verleden wil niet zeggen dat ze dezelfde bijdrage in de toekomst zullen hebben. De derde manier geeft een indicatie van de voorspelbaarheid. Dit betekent niet dat het dezelfde indicatie geeft voor de bijdrage aan onbalans uit het verleden, maar het geeft wel een indicatie voor de bijdrage van de onbalans in de toekomst.

De derde manier om de voorspelbaarheid te bepalen is gebaseerd op de MAPE. Een nadeel om de MAPE te vergelijken is, dat de MAPE gevoelig is voor de grootte waarvoor het de MAPE moet bepalen. Bijvoorbeeld, een verschil van één op een verbruik van 100 is volgens de MAPE minder erg dan een verschil van één op een verbruik van 50. Voor de onbalans maakt dit niet uit, waardoor de MAPE een vertekend beeld kan geven. Hier moet rekening mee gehouden worden als deze manier wordt toegepast.

Er zijn drie manieren onderzocht. Een vierde manier is om de prijs van de onbalans te kijken. Zo zijn er dure en goedkopere tijden. Dit heeft invloed op de prijsbijdrage van de onbalans.



## 5 Resultaten

Als eerste worden de resultaten van de onderzoeksvraag over het voorspellingsmodel gegeven. Vervolgens over de verbetering door middel van de groeperingen van klanten en als laatste onderdeel van dit hoofdstuk worden de resultaten van de individuele klantkennis analyse getoond. Wegens de hoeveelheid van de afzonderlijke voorspellingen en andere resultaten, en de dynamiek van de methode zijn de concrete onderzoeksresultaten niet opgenomen in de bijlage. Dit is wel op te vragen bij ondergetekende en na goedkeuring door Hot ITem.

### 5.1 Voorspellingsmodel

In deze paragraaf wordt antwoord op de eerste onderzoeksvraag gegeven. De vraag is: “*Is er een voorspellingsmodel dat beter is dan het huidige model?*”. Er is geprobeerd op twee manieren antwoord gegeven. Het exponential smoothing model en het ARIMA model kwamen als beste uit de bus en zijn nader onderzocht.

Er is gebleken dat niet aan alle eisen van een ARIMA model voldaan kan worden waardoor er is besloten om met dit model te stoppen. De voorspelfouten voldeden niet aan de eis van witte ruis waardoor het model niet gebruikt kan worden. Er was geen convergerende oplossing om de voorspelfouten te minimaliseren.

Via exponential smoothing is er wel een positief antwoord op de vraag te geven. Exponential smoothing is beter dan het huidige  $n - 7$  model. Het exponential smoothing model is op twee manieren getest. De eerste manier is de gebruikelijke manier en heeft optimale parameters. De tweede manier is de test op robuustheid en gebruikt de gevonden parameters om verder te voorspellen. In Tabel 5 is het resultaat te zien van de voorspellingen. Om de verbetering vast te stellen is er ook het  $n - 7$  model meegenomen in de tabel. De modellen zijn op basis van hun MAPE te beoordelen over hun accuraatheid.

Tabel 5: Prestaties van de verschillende modellen

Model	MAPE
Exponential Smoothing	2,20%
Test op robuustheid	2,53%
$n-7$	3,25%

Tabel 5 laat zien dat het exponential smoothing model een duidelijke verbetering is ten opzichte van het  $n - 7$  model. De test op robuustheid is minder goed dan het exponential smoothing model met optimale parameters, maar is alsnog een verbetering ten opzichte van het huidige model. Dit geeft aan dat het model redelijk robuust is. In Tabel 6 staan de optimale parameters waarmee het exponential smoothing model voorspeld heeft.

Tabel 6: Optimale parameters van het exponential smoothing model

Basisniveau $\alpha$	Trend $\gamma$	Dagpatroon $\delta$	Weekpatroon $\omega$	AR $\phi$
0,001	0,002	0,571	0,222	0,000

Om te onderzoeken hoe de voorspelfouten zich verhouden wordt in Tabel 7 de analyse weergegeven. Hierin valt op dat zowel het gemiddelde als de mediaan ver van nul liggen. Deze blijkt met een significantieniveau van 5% ook inderdaad anders te zijn dan nul. De standaard deviatie is 17.120. De skew is 0,36 en de excess-kurtosis is 1,67 ook deze zijn significant anders dan nul. Dit betekent dat de verdeling right-skewed is met een grotere rechter staart in vergelijking met de linker staart, maar dat de staarten dun zijn en de piek van de verdeling hoog is. Dit zien we maar gedeeltelijk terug in de kwantielen. Het eerste en derde kwantiel zijn in absolute zin kleiner dan de standaard deviatie, wat overeenkomt met de excess-kurtosis, maar het minimum is in absolute zin groter dan het maximum. Dit is niet te verwachten met de verdeling die skewed to the right is. Verder valt op dat de voorspelfouten geen witte ruis blijken te zijn en dat ze niet eens normaal verdeeld zijn. Wel is er sprake van een ARCH effect.

Tabel 7: Voorspelfouten analyse van het exponential smoothing model

Descriptive Statistics		Significance Test 5,00%			Test	p-vlaue	Result?
		Target	p-vlaue	Different?			
<b>AVERAGE:</b>	-1.322	0,000	0,00%	TRUE	<b>White-noise</b>	0,00%	FALSE
<b>STD DEV:</b>	17.120				<b>Normal Distributed?</b>	0,00%	FALSE
<b>SKEW:</b>	0,36	0,000	0,00%	TRUE	<b>ARCH Effect?</b>	0,00%	TRUE
<b>EXCESS-KURTOSIS:</b>	1,67	0,000	0,00%	TRUE			
<b>MEDIAN:</b>	-1.856						
<b>MIN:</b>	-88.878						
<b>MAX:</b>	70.595						
<b>Q 1:</b>	-11.911						
<b>Q 3:</b>	7.814						

Om de voorspelfouten van het exponential smoothing model te vergelijken met de voorspelfouten van het n-7 model, worden ook de voorspelfouten van het n-7 model geanalyseerd. Dit wordt in Tabel 8 weergegeven. Hierin zien we dat het gemiddelde, de standaard deviatie, de skew en de excess-kurtosis groter zijn dan bij het exponential smoothing model. Dit geeft de indicatie dat het exponential smoothing model betere voorspelfouten heeft waardoor de voorspelling zelf ook beter zijn. Verder zien we in Tabel 8 dat, met een significantieniveau van 5%, het gemiddelde, de skew en de excess-kurtosis significant anders zijn dan nul. Dit komt terug in de kwantielen. Daarnaast gedragen de voorspelfouten zich niet als witte ruis en zijn ze niet normaal verdeeld. Wel is er sprake van een ARCH effect.

Tabel 8: Voorspelfouten analyse van het n-7 model

Descriptive Statistics		Significance Test			5,00%	Test	p-vlaue	Result?
		Target	p-vlaue	Different?				
<b>AVERAGE:</b>	4.431	0,000	0,00%	TRUE		<b>White-noise</b>	0,00%	FALSE
<b>STD DEV:</b>	27.001					<b>Normal Distributed?</b>	0,00%	FALSE
<b>SKEW:</b>	0,81	0,000	0,00%	TRUE		<b>ARCH Effect?</b>	0,00%	TRUE
<b>EXCESS-KURTOSIS:</b>	2,73	0,000	0,00%	TRUE				
<b>MEDIAN:</b>	3.578							
<b>MIN:</b>	-68.701							
<b>MAX:</b>	137.391							
<b>Q 1:</b>	-11.975							
<b>Q 3:</b>	18.153							

De voorspelfouten van het exponential smoothing model uit Tabel 7 en de voorspelfouten van het n-7 model uit Tabel 8 zijn niet te vergelijken met de voorspelfouten van het ARIMA model uit Tabel 4. Dit komt omdat

### Wat kan het Eneco Business opleveren?

[VERTROUWELIJK]

Tabel 9: Onbalanskosten met de verschillende modellen

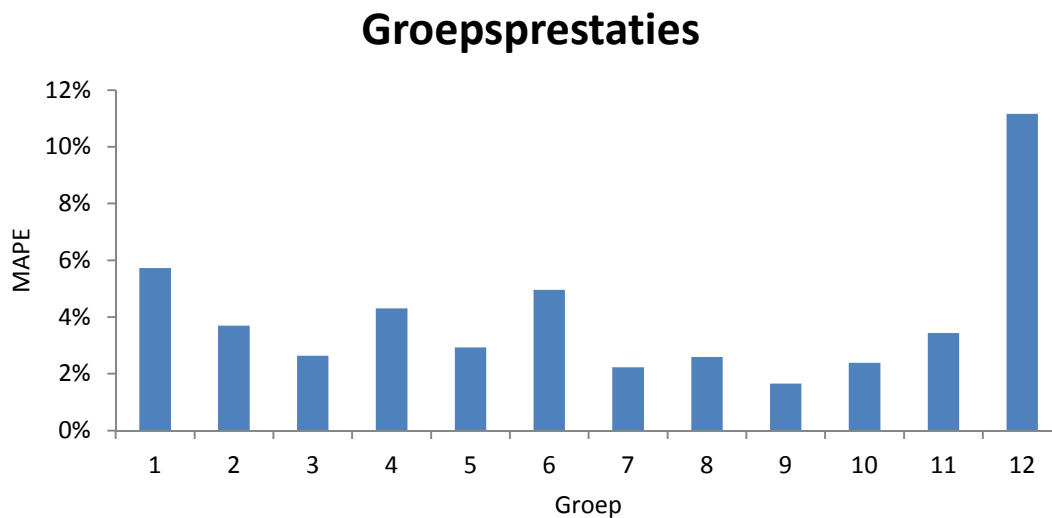
## 5.2 Klantclustering

De tweede onderzoeksvraag wordt in deze paragraaf behandeld. De vraag is: "Kunnen klanten zo gegroepeerd worden dat de som van de groepsvoorspellingen beter is dan de voorspelling van het totaalverbruik?".

[VERTROUWELIJK]

Tabel 10: Prestaties van de groepsvoorspelling

Zoals gezegd is de groepsvoorspelling uit Tabel 10 opgebouwd door de som van de verschillende groepsvoorspellingen te nemen. Figuur 13 laat voor iedere groep de verkregen MAPE zien van de voorspellingen. Hierin zien we dat groep van nieuwe en vertrekkende klanten (groep twaalf) een hoge MAPE heeft. In de praktijk is deze groep eenvoudig te voorspellen aan de hand van start en einddatum van het contract, maar het model kan hier niet goed mee overweg. De overige groepen hebben een gelijkwaardige MAPE. Groep negen is echter een positieve uitschieter en heeft de laagste MAPE van 1,65%. Wegens de geanonimiseerde data is het niet bekend welke klanten dit zijn waardoor er geen verdere analyse op gemaakt kan worden.



Figuur 13: MAPE van de verschillende groepen

[VERTROUWELIJK]

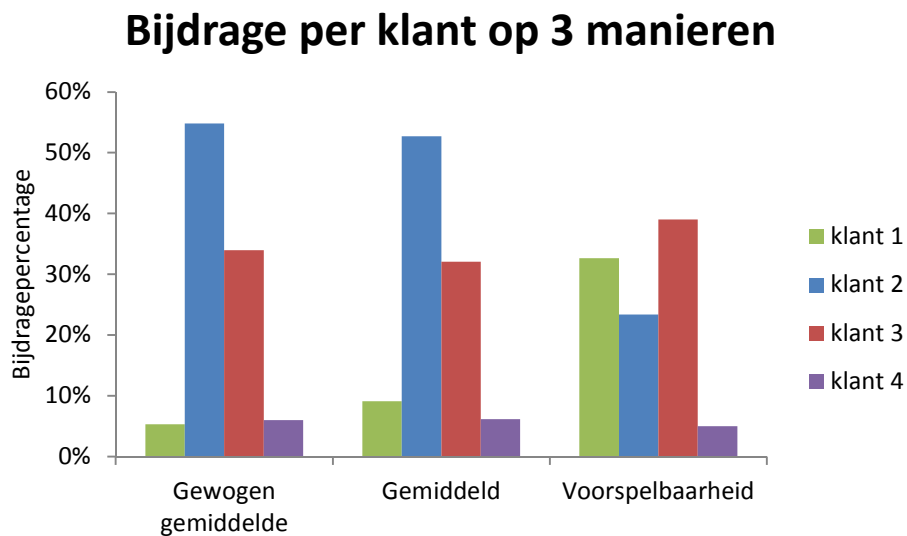


Tabel 11: Prestaties van de groepsvoorspelling zonder de groep van nieuwe en vertrekkende klanten

### 5.3 Individuele klantkennis

In deze paragraaf wordt de derde onderzoeksvraag beantwoord. De vraag is: “*Kan er een onbalanspremie op klantniveau gemaakt worden?*”. Door de individuele bijdrage op onbalans en voorspelbaarheid te analyseren wordt geprobeerd een antwoord te geven.

Aan de hand van een subset zijn op drie manieren de bijdragepercentages op de onbalans berekend. Figuur 14 laat het resultaat zien van de bijdrage per klant op de drie manieren.



Figuur 14: Onbalansbijdrage per klant op drie verschillende manieren

Manier 1 en 2 komen redelijk met elkaar overeen. Het grootste verschil is dat volgens manier 2 klant 1 iets meer bijdraagt waardoor klant 2 en 3 iets minder bijdragen. Voor Klant 4 is de bijdrage op beide manieren hetzelfde. Manier 3 komt niet goed overeen met manieren 1 en 2. Klant 4 is nog redelijk

constant, maar vooral klant 1 en 2 hebben een ander bijdragepercentage. Klant 1 is van minder dan 10%, volgens manier 1, naar meer dan 30% volgens manier 3 gegaan. Klant 2 laat een omgekeerd percentage zien. Van meer dan 50%, volgens manier 1 en 2, naar minder dan 25% volgens manier 3. Het percentage van klant 3 is iets omhoog gegaan, maar blijft tussen de 30 en 40%.

Manieren 1 en 2 onderscheiden de klanten op de bijdrage van de onbalans. Manier 1 geeft de gewogen gemiddelde bijdrage en manier 2 geeft het standaard gemiddelde. Manier 3 geeft een algemene indicatie van voorspelbaarheid aan door de MAPE om te zetten in bijdragepercentages. Voor klant 3 en 4 geven de MAPE waardes een redelijke indicatie voor de bijdrage van de werkelijke onbalans. Voor klant 1 en 2 geeft de MAPE echter een verkeerd beeld. Voor klant 1 is de MAPE hoog, maar de bijdrage laag. Voor klant 2 is het andersom, klant 2 heeft een redelijk lage MAPE maar een hoog bijdragepercentage.

Het blijkt dat klant 1 een kleine klant is ten opzichte van de andere klanten. Hierdoor geeft de voorspelbaarheid aan de hand van de MAPE een vertekend beeld. Dit zou de oorzaak kunnen zijn van het verschil tussen de voorspelbaarheid en gemiddelde bijdrage.

Hieruit volgt dat de MAPE niet altijd het juiste bijdragepercentage weergeeft. Als klanten goed voorspelbaar zijn, wil dat nog niet zeggen dat ze automatisch een lage bijdrage leveren aan de onbalans. Dit geldt voornamelijk als er grote verschillen bestaan tussen het verbruik van klanten.

Om terug te komen op het antwoord op de onderzoeksvraag: Ja, er kan onbalanspremie op klantniveau gemaakt worden. Klanten zijn te onderscheiden op de bijdrage aan de hand van de historische bijdrage. Dit betekent echter niet dat de klanten dezelfde bijdrage zullen leveren in de toekomst. Aan de hand van de MAPE is de indicatie voor toekomstige bijdrage van de onbalans betrouwbaarder door de meer voorspellende waarde van de methode, mits de klanten een vergelijkbare orde van grote in het verbruik hebben. Daarbij komt dat de voorspelbaarheid nog geen garantie geeft dat het ook daadwerkelijk goed voorspeld wordt. Het hangt van de praktijk situatie af welke manier het beste gebruikt kan worden. Als achteraf de onbalans verrekend kan worden met de klant, kan de eerste of tweede manier gebruikt worden. Er zou dan bijvoorbeeld een extra toeslag of teruggave kunnen komen op het voorschot van wat de klant al heeft betaald. Als de onbalanspremie vooraf vastgesteld moet worden kan de derde manier gebruikt worden. Hierin wordt de voorspelbaarheid van de klant meegenomen waardoor het risico op onbalans in de toekomst wordt bepaald. Met deze manier moet wel rekening gehouden worden met de grote van de klant.

## 6 Bespreking van de praktijk

Tot nu toe zijn alleen de literatuur en het eigen onderzoek besproken. Hoewel dit de belangrijkste onderdelen zijn, is het daarnaast ook interessant te kijken wat er in de praktijk gebeurt. Om dit te onderzoeken, is er geprobeerd contact te leggen met verschillende partijen. Deze partijen zijn actief op het gebied van korte termijn voorspellingen van elektriciteit. De volgende vragen staan centraal bij de communicatie met de verschillende partijen:

- Welk model wordt er gebruikt voor de kort termijn voorspelling?
- Hoe presteert dit model?
- Wordt er klantkennis meegenomen om te voorspellen?

Van de verschillende benaderde bedrijven waren Eneco, Essent, RWE en Mikan in staat en bereid om te helpen bij dit praktijkonderzoek. Door middel van mailwisseling is er herhaaldelijk contact geweest en werd er antwoord gegeven op de gestelde vragen. Ondanks de bereidheid antwoord te geven, gaf ieder bedrijf vooraf aan dat ze waarschijnlijk niet te diep op de stof konden ingaan wegens concurrentiegevoelige informatie. Mede hierdoor is er besloten om niet in te gaan op de onderzoeksvraag of er voor klanten individuele onbalansprijzen worden berekend. Dit om een bijna zeker antwoord dat het te gevoelig ligt te ontlopen.

Voor Eneco is er gesproken met een medewerker van EET. Als antwoord op de eerste vraag gaf hij aan dat er gebruik wordt gemaakt van meerdere modellen. Het software pakket dat gebruikt wordt, maakt gebruik van een base-load leerset, ARIMA model, lineaire regressie en een neuraal netwerk. De uitkomst wordt nagekeken door een expert die vaak de laatste aanpassingen doet om de uiteindelijke voorspelling te krijgen. De MAPE die Eneco bereikt met deze methode is ongeveer 1,7%. Klantkennis wordt sporadisch meegenomen en uitsluitend als het grote invloed heeft op de voorspelling.

Een medewerker van Essent Energy Trading wilde niet veel informatie kwijt over de door hen gebruikte modellen. Wel werd aangegeven dat weersinvloeden voor sommige type klanten niet van grote invloed zijn. De MAPE die ze bereiken met hun voorspellingen ligt tussen de 1,5% en 2,5%. Klantinformatie wordt niet meegenomen omdat klanten elkaar kunnen opheffen waardoor de totaalvoorspelling beter is dan de afzonderlijke voorspellingen bij elkaar. Wel gaf hij aan het interessant te vinden om te kijken wat de resultaten zouden zijn van een onderzoek naar de voorspelling van het totaalverbruik, de grootste klanten en een combinatie hiervan.

Een medewerker van RWE gaf aan dat de voorspelling grotendeels gebaseerd is op de similar day methode. Dit is een methode die een voorspelling levert aan de hand van het verbruik van overeenkomstige dagen. Aan de hand van weersinvloeden en een ARIMA model krijgt de voorspelling zijn uiteindelijke vorm. Op deze manier is een MAPE tussen de 1,5% en 2,5% gebruikelijk. Een antwoord op de vraag over klantkennis werd omzeild waardoor de suggestie werd gewekt dat er niet veel mee gedaan wordt.

De oprichter van Mikan gaf aan dat zij de Multivariate Adaptive Regression Splines (MARS) methode gebruiken om te voorspellen. Volgens hem presteert deze methode veel beter dan de exponential smoothing, ARIMA en neurale netwerken en is zij daarbij relatief eenvoudig en zeer snel. Naar aanleiding van deze opmerking en een snelle scan op het internet hierover werd er in een laat stadium van het onderzoek overwogen deze methode mee te nemen. Na overleg met de opdrachtgever is er een impactanalyse gemaakt voor deze change. Hierin heeft de opdrachtgever de keuze gemaakt om door te

gaan met klantgroepen en individuele klantkennis in plaats van de MARS methode uit te werken. De impactanalyse is opgenomen in Bijlage B. De oprichter van Mikan gaf aan dat de MAPE van de MARS methode rond de 1,5% ligt. Hierbij werd vermeld dat er af en toe negatieve uitschieters bij zitten van rond de 5%. Over het meenemen van klantkennis in de voorspelling kon hij nog weinig kwijt omdat hij er zelf nog volop mee bezig was. Wel kon hij patroonherkenning aanbevelen om mogelijk iets over de verschillende klanten te kunnen zeggen.

**Tabel 12: Prestaties van de verschillende partijen**

	MAPE
Mikan	1,5% - 5%
Eneco	1,7%
Essent	1,5% - 2,5%
RWE	1,5% - 2,5%

In Tabel 12 zien we nog een keer de prestaties van de verschillende partijen samengevat terug. In de tabel valt op dat de beste waarden van de verschillende partijen ongeveer gelijkwaardig zijn, maar dat EET geen bovengrens aangeeft. De overige partijen geven een bereik aan waartussen de MAPE meestal ligt. Dit bereik komt redelijk overeen met dit onderzoek en met de onderzoeken uit de literatuur.

## **7 Conclusie**

[VERTROUWELIJK]



## 8 Vervolg onderzoek

Dit onderzoek is opgedeeld in drieën met elk een onderzoeksvraag. Op alle onderzoeksvragen zijn positieve antwoorden gegeven, maar de verwachting is dat elke vraag nog beter beantwoord kan worden.

Als verbeterd model is het exponential smoothing model gevonden. Dit model is eenvoudig en snel te gebruiken. Een ander model dat eenvoudig en snel te gebruiken is, is het MARS model. Van dit model wordt verwacht dat de prestaties beter zijn dan van het exponential smoothing model. Als vervolg onderzoek kan het MARS model onderzocht worden.

In het vervolg onderzoek over klantgroeperingen zou onderzocht kunnen worden welke klantkennis nog meer aan de verbetering kan bijdragen. Ook kan dit gedeelte van het onderzoek verder geoptimaliseerd worden. Zo kunnen er andere data pre-process stappen onderzocht worden, net als een andere manier van groeperen.

Het vervolg onderzoek naar onbalanspremies op klantniveau zou zich moeten concentreren op de premieberekening zelf. Welke berekening zou Eneco Business het beste kunnen toepassen? In dit onderzoek zou het nuttig zijn om de praktijk situatie mee te nemen. Het onderzoek zou zich moeten richten op de haalbaarheid van de verschillende manieren.



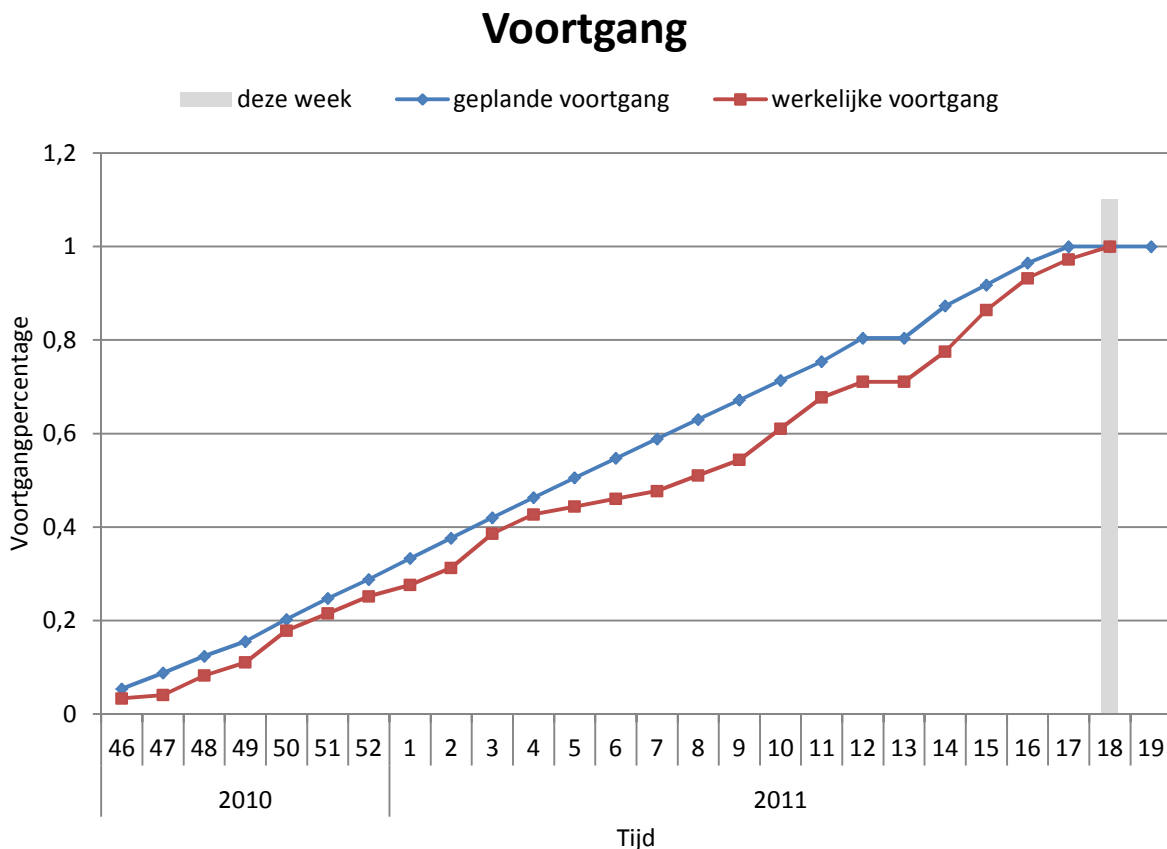


## 9 Evaluatie van het onderzoek

Deze evaluatie is gebaseerd het boek Projectmatig creëren 2.0 van Bos en Harting (2006). Naast deze evaluatie is het ook gebruikt als leidraad voor het projectmanagementonderdeel van dit onderzoek.

Aan dit onderzoek is een opdrachtformulering vooraf gegaan. Hierin werd ik ondergedompeld in de wereld van elektriciteit en kwam boven zwemmen met een goedgekeurd stagevoorstel. Aan het begin van deze onderzoek is de opdracht verder gespecificeerd en heb ik een plan van aanpak geschreven. Achteraf is gebleken dat doelstellingen toen nog niet helemaal helder waren. Gedurende het onderzoek werd door regelmatig overleg met de opdrachtgever en begeleiders de richting steeds duidelijker. Door de onbekendheid over het verloop van dit onderzoek in het begin is het lastig het eindresultaat te toetsen aan het plan van aanpak.

De planning aan het begin van dit onderzoek is redelijk gehaald. Uit Figuur 15 zien we dat de werkelijke voortgang structureel achter liep op de planning, maar het project blijft binnen de marge van twee weken uitloop. We zien in Figuur 15 een terugkerende golf in de werkelijke voortgang. Het blijkt dat elke golf een fase is in het project. Aan het begin van een fase loopt de geplande voortgang uit op de werkelijke voortgang, maar gedurende de fase loopt de werkelijke voortgang in op de geplande voortgang. Toch was het niet mogelijk om de planning helemaal bij te halen.



Figuur 15: Voortgang van het onderzoek

De planning is mede redelijk gehaald door twee situaties. Het eerste voorval is de voorspellingstool waar het plan van aanpak het over heeft. Deze tool is uiteindelijk afgeslankt in de vorm van een Excel sheet met een Powerpivot connectie. Hiermee kunnen de juiste klanten ingeladen worden waarna het verbruik voorspeld kan worden. Het tweede voorval is het besluit geweest om de MARS methode niet mee te nemen. Het onderzoek dreigde verder achter op de planning te raken als ook de MARS methode onderzocht zou worden.

Gedurende het onderzoek zijn er twee beslismomenten geweest. Het eerste moment was de goedkeuring van het plan van aanpak. Hierdoor kon het onderzoek inhoudelijk van start gaan. Het tweede beslismoment was de tussentijdse presentatie in het bijzijn van de opdrachtgever. Ondanks de soms stevige kritiek werd het resultaat goedgekeurd en kon de uitwerking beginnen. Op het moment van schrijven moet het derde beslismoment nog volgen en zal plaats vinden na de eindpresentatie. De begeleiders van Hot ITem en van de *vrije* Universiteit zullen beslissen of het onderzoek, verslag en presentatie voldoen aan de gestelde eisen.

Tijdens het onderzoek is er niet echt sprake geweest van samenwerking, maar is er meer sprake geweest van overleg en verantwoording. In het begin van het onderzoek is er veel contact geweest met de opdrachtgever om het onderzoek te doorgronden. Dit is voornamelijk tijdens het schrijven van het plan van aanpak geweest. Gedurende het onderzoek werd er één uur per week gereserveerd voor de wekelijkse update met één van de Hot ITem begeleiders. Hierin werd de voortgang besproken en knelpunten gesignaleerd. Het onderzoek werd bij Hot ITem uitgevoerd, hierdoor waren de lijntjes met de begeleiders kort. De begeleiders waren enthousiast en inspirerend waardoor het mijn enthousiasme en inzet versterkte. De wekelijkse update mailtjes naar de VU begeleider zijn er een beetje bij in geschoten. Er is maar af en toe mailcontact geweest tussen mij en de VU begeleider. Hierdoor was de betrokkenheid niet groot.

Door dit onderzoek te doen heb ik twee lessen geleerd die buiten de inhoud van dit onderzoek vallen. De eerste les is het zorgvuldiger omgaan met de Deming-cyclus. De plan, do, check, act stappen heb ik als zeer belangrijk ervaren, maar zijn puntjes voor verbetering. In de eerste fase van het onderzoek had ik vooral moeite met het plannen. In latere fases waren het voornamelijk de check en act stappen die nog verbeterd konden worden. De tweede les is de documentatie gedurende het onderzoek. Het blijft voor mij lastig om alles helder en zorgvuldig te documenteren zodat ik het ook nog terug kan lezen of kan gebruiken zonder het opnieuw te schrijven. Dit is een valkuil voor mij en daar zal ik verder aan moeten werken.

Uiteindelijk is dit onderzoek beter gegaan dan ik me vooraf had voorgesteld. Dit onderzoek zou het hoogtepunt van mijn opleiding moeten betekenen en dat is het ook voor mijn gevoel. Ik kon verschillende vakken van de studie herkennen en toepassen in dit onderzoek, waardoor ik een voldaan gevoel kreeg over heel mijn opleiding. Daarnaast heeft dit onderzoek bereikt wat het wilde bereiken, maar kostte het minder moeite dan ik me vooraf had bedacht. Dat wil overigens niet zeggen dat het makkelijk is geweest, maar het onderzoek gaf me meer energie dan het energie kostte.

## 10 Referenties

- Bos, J. en Harting, E. *Projectmatig creëren 2.0*. Schiedam: Scriptum, 2006.
- Burger, M., Greaber, B. en Schindlmayr, B. *Managing Energy Risk: An Integrated View on Power and Other Energy Markets*. New York: Wiley, 2007.
- Cizek, P., Härdle, W. en Weron, R. *Forecasting with ARIMA models*. 3 maart 2005. <http://fedc.wiwi.hu-berlin.de/xplore/tutorials/xegbohtmlnode38.html> (geopend maart 11, 2011).
- Darbellay, G.A. en Slama, M. „Forecasting the short-term demand for electricity: Do neural networks stand a better chance?” *International Journal of Forecasting* 16 (2000): 71-83.
- Fischer, M. „Forecasting Energy Demand.” *Mikan Solutions!: Energy*. 2004. <http://www.mikan.nc/app8.htm> (geopend november 19, 2010).
- Gunst, de M.C.M. *Statistical Models*. Colledictaat, Amsterdam: Vrije Universiteit, 2009.
- Makridakis, S., Wheelwright, S.C. en Hyndman, R.J. „Exponential Smoothing Methods.” In *Forecasting: Methods and Applications*, 135-184. New York: Wiley, 1998.
- Nau, B. *Forecasting: Decision 411*. 16 mei 2005. <http://www.duke.edu/~rnau/411home.htm> (geopend januari 28, 2011).
- Peirson, J. en Henley, A. „Electricity load and temperature.” *Energy Economics* 16 (1994): 235-243.
- Ramanathan, R., Engel, R., Granger, C.W.J., Vahid-Araghi, F. en Brace, C. „Short-run forecasts of electricity loads and peaks.” *International Journal of Forecasting* 13 (1997): 161-174.
- Soares, L.J. en Medeiros, M.C. „Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazil data.” *International Journal of Forecasting* 24 (2008): 630-644.
- Taylor, J.W. „Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing.” *The Journal of the Operational Research Society* 54, nr. 8 (2003): 799-805.
- Taylor, J.W., Menezes, de L.M. en McSharry, P.E. „A comparison of univariate methods for forecasting electricity demand up to a day ahead.” *International Journal of Forecasting* 22 (2006): 1-16.



## 11 Bijlage

### 11.1 Bijlage A: Samenvatting van ARIMA regels

#### Identifying the order of differencing and the constant:

- Rule 1: If the series has positive autocorrelations out to a high number of lags, then it probably needs a higher order of differencing.
- Rule 2: If the lag-1 autocorrelation is zero or negative, or the autocorrelations are all small and patternless, then the series does *not* need a higher order of differencing. If the lag-1 autocorrelation is -0.5 or more negative, the series may be overdifferenced. **BEWARE OF OVERDIFFERENCING!!**
- Rule 3: The optimal order of differencing is often the order of differencing at which the standard deviation is lowest.
- Rule 4: A model with no orders of differencing assumes that the original series is stationary (among other things, mean-reverting). A model with one order of differencing assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth). A model with two orders of total differencing assumes that the original series has a time-varying trend (e.g. a random trend or LES-type model).
- Rule 5: A model with no orders of differencing normally includes a constant term (which represents the mean of the series). A model with two orders of total differencing normally does not include a constant term. In a model with one order of total differencing, a constant term should be included if the series has a non-zero average trend.

#### Identifying the numbers of AR and MA terms:

- Rule 6: If the partial autocorrelation function (PACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is positive--i.e., if the series appears slightly "underdifferenced"--then consider adding one or more AR terms to the model. The lag beyond which the PACF cuts off is the indicated number of AR terms.
- Rule 7: If the autocorrelation function (ACF) of the differenced series displays a sharp cutoff and/or the lag-1 autocorrelation is negative--i.e., if the series appears slightly "overdifferenced"--then consider adding an MA term to the model. The lag beyond which the ACF cuts off is the indicated number of MA terms.
- Rule 8: It is possible for an AR term and an MA term to cancel each other's effects, so if a mixed AR-MA model seems to fit the data, also try a model with one fewer AR term and one fewer MA term--particularly if the parameter estimates in the original model require more than 10 iterations to converge.
- Rule 9: If there is a unit root in the AR part of the model--i.e., if the sum of the AR coefficients is almost exactly 1--you should reduce the number of AR terms by one and increase the order of differencing by one.
- Rule 10: If there is a unit root in the MA part of the model--i.e., if the sum of the MA coefficients is almost exactly 1--you should reduce the number of MA terms by one and reduce the order of differencing by one.
- Rule 11: If the long-term forecasts appear erratic or unstable, there may be a unit root in the AR or MA coefficients.

**Identifying the seasonal part of the model:**

- Rule 12: If the series has a strong and consistent seasonal pattern, then you should use an order of seasonal differencing--but never use more than one order of seasonal differencing or more than 2 orders of total differencing (seasonal+nonseasonal).
- Rule 13: If the autocorrelation at the seasonal period is positive, consider adding an SAR term to the model. If the autocorrelation at the seasonal period is negative, consider adding an SMA term to the model. Do not mix SAR and SMA terms in the same model, and avoid using more than one of either kind.

## 11.2 Bijlage B: MARS Impactanalyse

### Inleiding:

Er is een voorspellingsmethode die mogelijk beter voorspelt dan de methodes die nu in het project worden meegenomen. Het past niet in de huidige planning om deze nieuwe methode mee te nemen. Om toch binnen de planning te blijven is het mogelijk om de overige analyses te schrappen of deels te schrappen. Wilt u aangeven of de nieuwe voorspellingsmethode meegenomen moet worden?

### Probleem:

Een specialist op dit gebied heeft mij op een andere methode gewezen die beter lijkt te zijn dan de ARIMA, exponential smoothing en lineaire regressie methodes die nu meegenomen worden. De nieuwe methode is de Multivariate Adaptive Regression Splines (MARS) methode. Deze methode beheers ik niet en kost extra tijd om onder de knie te krijgen. In de huidige planning heb ik hier geen tijd voor en zal de klantkennisanalyse vervallen en gaat het voor een groot gedeelte ten koste van de klantgroepering. Hieronder is de huidige en aangepaste planning te zien:

Huidige planning:				Aangepaste planning:			
Activiteit	week 9	week 10	week 11	Activiteit	week 9	week 10	week 11
Klantkennisanalyse				Klantkennisanalyse			
Klantgroepering				Klantgroepering			
voorspellen				voorspellen			
				MARS methode			

### Baten:

De MARS methode heeft een MAPE van 1,5% tot 5% volgens de specialist. De methodes die nu gebruikt worden hebben een geschatte MAPE tussen de 2,75% en 3%, terwijl de N-7 een geschatte MAPE heeft van 3,5%. Dit betekent dat de MARS methode niet beter hoeft te zijn, maar dat er een kans is dat het beter voorspelt dan de andere. Waar de andere methodes maar een half procent verbetert, kan de MARS misschien twee keer zo goed voorspellen dan de benchmark.

### Kosten:

Klantkennisanalyse wordt niet meer meegenomen en klantgroepering vervalt voor een groot gedeelte. Het gevolg daarvan is dat er in dit onderzoek geen duidelijkheid wordt gegeven van de gevolgen van klantkennis en groepering op de onbalans.

### Voorstel:

Aangezien klantkennis en groeperingen in het domein van Eneco Business liggen en voorspellingstechnieken meer het domein van Eneco Energy Trade is, raad ik aan om af te zien van deze verandering. Mocht er tijd over zijn kan het verder onderzocht worden en als advies meegeven worden.

*Kunt u zo snel mogelijk aangeven of u mijn voorstel wil volgen?*

### 11.3 Bijlage C: Afkortingen

ACF	AutoCorrelatie Functie
AIC	Akaike Information Criterion
AR	AutoRegressive
ARCH	AutoRegressive Conditional Hetroskedasticity
ARIMA	AutoRegressive Integrated Moving Average
EET	Eneco Energy Trade
kWh	Kilowattuur
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Splines
MWh	Megawattuur
PACF	Partiële AutoCorrelatie Functie
PCA	Principal Component Analysis
PV	ProgrammaVerantwoordelijke
SBC	Schwartz Bayesian information Criterion
SLA	Service Level Agreement
TWh	Terawattuur