Tijdreeks Voorspellingen

met neurale netwerken,
nearest-neighbour-methoden
en trend-fitting

Erwin de Winter
(erwindewinter@yahoo.com)
Sentient Machine Research

7 februari 2002

afstudeerscriptie Kunstmatige Intelligentie
richting Autonome Systemen
Faculteit Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

Inhoudsopgave

1. Introductie

Deze scriptie is het resultaat van mijn onderzoek naar het voorspellen van tijdreeksen. Dit onderzoek is verricht in een stage bij Sentient Machine Research (SMR) te Amsterdam. Dit bedrijf is in 1990 opgericht door de huidige directeur Marten den Uyl met als doelstelling adaptieve technieken praktisch toepasbaar te maken bij het oplossen van bedrijfsproblemen. Deze technieken worden toegepast in datamining-producten voor het opsporen van patronen en verbanden in databases. Tevens is SMR gespecialiseerd in het uitvoeren van database-marketing-analyses, waarin selecties en segmentaties in een database worden uitgevoerd voor marketingdoeleinden.

Klanten van SMR bleken ook geïnteresseerd te zijn in het voorspellen van data. In het bijzonder was het dagblad De Telegraaf geïnteresseerd in een voorspelling van de verkochte advertentieruimte. Op het gebied van voorspellen was nog weinig kennis beschikbaar bij SMR en dit leek mij een interessante uitdaging voor een stageonderzoek.

Omdat SMR vaak op nearest-neighbour gebaseerde technieken gebruikt, ontstond de vraag of deze technieken ook toepasbaar waren voor het voorspellen van tijdreeksen in het algemeen en de advertentieruimte van de Telegraaf in het bijzonder.

Voor dagbladuitgevers is het interessant om een voorspelling te hebben van de verkochte advertentieruimte van een dagblad op een bepaald moment in de toekomst. Een eindejaarsvoorspelling halverwege het jaar is bijvoorbeeld zeer gewenst, evenals een langere termijn voorspelling van één tot zelfs een aantal jaren vooruit. De Telegraaf was al langer op zoek naar een goede voorspellingstechniek maar is daar nooit goed in geslaagd. Intern is ook onderzoek gedaan naar het gebruik van bestaande tijdreeksvoorspellingstechnieken, maar die zijn nooit operationeel geweest.

Voor het voorspellen van tijdreeksen bestaan standaard lineaire voorspellingsmodellen. Voorbeelden hiervan zijn onder andere AR, ARX en Box-Jenkins modellen. Dit zijn statistische modellen waarbij de voorspelling een lineaire functie van de historische data is. Met deze modellen zijn in het verleden goede resultaten behaald.

Neurale netwerken kunnen door hun veelzijdige toepasbaarheid ook gebruikt worden bij het voorspellen van tijdreeksen. Neurale netwerken zijn niet-lineair en de veronderstelling is dat zij een beter resultaat geven bij complexe datasets dan de lineaire modellen.

Daarnaast is het voor SMR interessant om voor de veel gebruikte nearest-neighbourtechnieken de toepasbaarheid op het voorspellen van tijdreeksen te onderzoeken. Deze technieken zijn tevens niet-lineair en hebben een transparante, snelle en robuuste werking.

Vaak wordt bij het voorspellen van tijdreeksen uitsluitend gebruik gemaakt van historische gegevens van de te voorspellen variabele. In dit onderzoek wordt tevens onderzocht of het gebruik van externe variabelen in het voorspellingsmodel kan bijdragen tot een betere voorspelling. Het is bekend dat er verbanden bestaan tussen de advertentievolumes en externe macro- en micro-economische variabelen. Het onderzoek richt zich op het vinden van deze variabelen en het onderzoeken van de bijdrage die deze variabelen kunnen hebben in het voorspellingsproces.

Doelstelling van de stage is het onderzoeken van de toepasbaarheid en prestaties van verschillende technieken voor het voorspellen van tijdreeksen. Tevens wordt de invloed van externe variabelen op de voorspelling onderzocht.

De volgende methoden zullen onderzocht worden:

- Trend-fitting-methode
- Neurale netwerken
- Nearest-neighbour-methoden

De resultaten van deze methoden worden vergeleken met twee gangbare tijdreeks-voorspellingsmethoden:

- ARX
- Box-Jenkins

De hoofdstukindeling is als volgt: hoofdstuk 2 is een inleiding in de theorie van tijdreeksen. Hoofdstuk 3 bevat een beschrijving van de voorspellingstaak. In hoofdstuk 4 worden de diverse voorspellingsmethoden, respectievelijk lineaire modellen, trend-fitting-methode, neurale netwerken en nearest-neighbour-methoden beschreven. Hoofdstuk 5 bevat de experimenten en hoofdstuk 6 de algemene resultaten en de conclusies.

2. Tijdreeksen

2.1 Tijdreeksen

Een tijdreeks is een verzameling observaties van een variabele in de tijd. Met behulp van de tijdreeks kan een model opgesteld worden waarmee toekomstige waarden van de tijdreeks voorspeld kunnen worden.

2.1.1 Beschrijving

Een geobserveerde tijdreeks $\boldsymbol{x}$, bestaande uit t observaties wordt weergegeven als:

$$\boldsymbol{x} = x_1, x_2, \ldots, x_t \quad (x_i \in \mathbb{R})$$

In dit verslag wordt uitgegaan van tijdreeksen waarbij de tijd tussen twee observaties constant is.

2.1.2 Voorspelling

De voorspelling op een willekeurig tijdstip $t+n$ in de toekomst is:

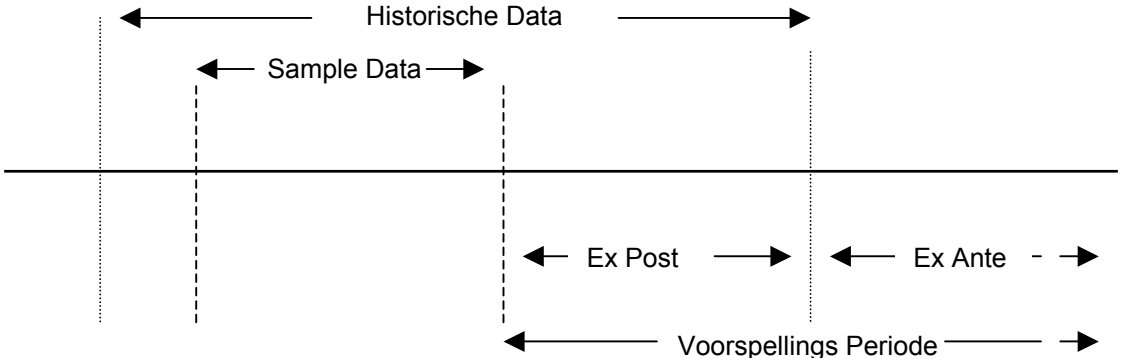
$$\hat{x}_{t+n}$$

Wanneer de waarde van de variabele op het tijdstip $t+n$ beschikbaar is, kan de het verschil met de voorspelde waarde bepaald worden. Deze voorspellingsfout ε_{t+n} is dan:

$$\varepsilon_{t+n} = x_{t+n} - \hat{x}_{t+n}$$

2.1.3 Tijdlijn

Analoog aan Gaynor/Kirkpatrick [Gaynor94] wordt alle data van de geobserveerde variabele de *historische data* genoemd. De *sample data* wordt gebruikt om het model van de tijdreeks te maken en maakt deel uit van de historische data. De *voorspellingsperiode* is verdeeld in twee gescheiden perioden, de *ex post* en de *ex ante* voorspelling. (Zie figuur 2.1)



Figuur 2.1: De tijdlijn.

De *ex post* periode is de periode vanaf de eerste observatie na het einde van de sample periode tot de laatste geobserveerde waarde van de historische data. Ex post voorspellingen kunnen getoetst worden, omdat de werkelijke waarden ook beschikbaar zijn. Op deze manier kan de voorspellingsfout bepaald worden.

De *ex ante* periode begint na de laatste geobserveerde waarde van de historische data. Van deze periode zijn geen observaties beschikbaar en het is deze periode die voorspeld moet worden met het ontwikkelde voorspellingsmodel.

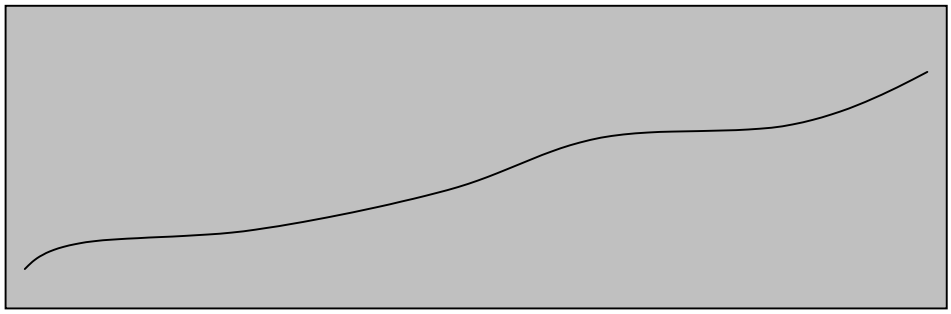
2.1.4 Tijdreekspatronen

In dit verslag wordt gewerkt met tijdreeksen van economische processen. Gaynor/Kirkpatrick [Gaynor94] onderscheiden een aantal standaard patronen in dit type tijdreeksen:

- Trend
- Seizoensvariaties
- Cyclische variaties
- Irreguliere variaties

De gebruikte tijdreeksen bestaan doorgaans uit een combinatie van twee of meer van deze patronen.

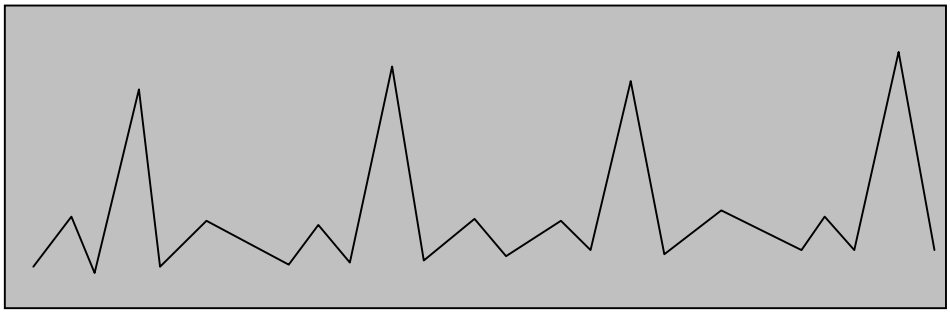
Trend:



Figuur 2.2: Trend in een tijdreeks.

Trend is een aanhoudende op- of neerwaartse beweging in de data. Trend geeft een lange termijn groei of daling weer. Veel economische variabelen bezitten een (vaak stijgende) trend.

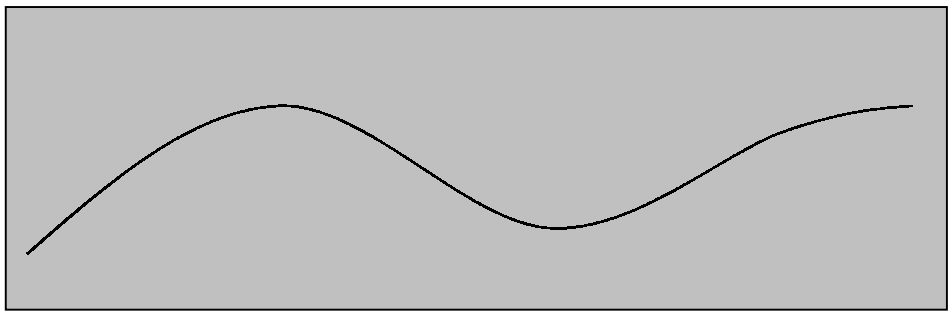
Seizoensvariaties:



Figuur 2.3: Seizoensvariaties in een tijdreeks.

Als een tijdreeks seizoensvariaties heeft, dan is er een herhalend patroon binnen een vaste periode. Dit patroon wordt bijvoorbeeld bepaald door weersinvloeden, vakanties of andere seizoensgerelateerde variabelen..

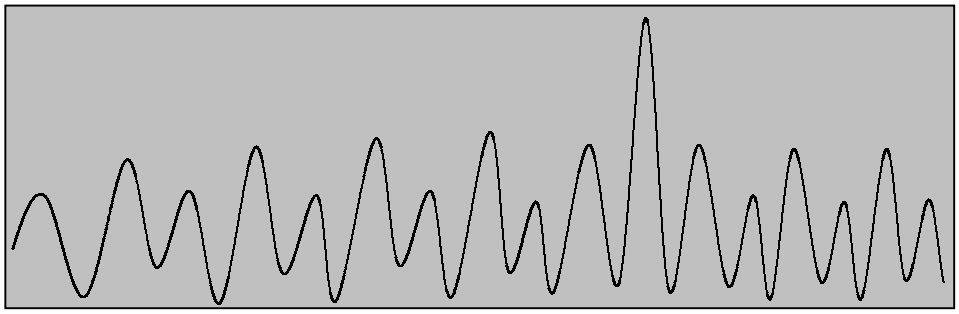
Cyclische variaties:



Figuur 2.4: Cyclische variaties in een tijdreeks .

Er is sprake van een cyclische variatie in een tijdreeks als een patroon bestaat waarbij een globale stijging gevolgd wordt door een periode van daling. Een voorbeeld van een cyclische variatie is het gemiddelde inkomen over een groot aantal jaren. Wanneer periodes van economische groei afgewisseld worden door recessies, ontstaan er cycli in het gemiddelde inkomen.

Irreguliere variaties:



Figuur 2.5: Een irreguliere variatie in een tijdreeks.

Irreguliere variaties hebben geen duidelijk patroon en worden vaak veroorzaakt door externe invloeden. Deze invloeden zijn ook meestal éénmalige gebeurtenissen. Een voorbeeld is een politieke uitspraak die invloed heeft op de olieprijs. Door het willekeurige karakter zijn deze variaties niet te voorspellen.

3. Voorspellingstaak

3.1 Inleiding

Het onderzoek richt zich op het voorspellen van de omvang van de verkochte advertentieruimte door het dagblad De Telegraaf. Deze advertentieruimte wordt in het vervolg advertentievolume genoemd en wordt gemeten in pagina's.

Voor dagbladuitgevers is het belangrijk te weten hoe de markt voor dagbladadvertenties zicht ontwikkelt. Op basis hiervan worden bijvoorbeeld prijsafspraken, afnamegaranties en kortingen afgesproken. Daarom is het voor een dagbladuitgever interessant een inzicht te hebben in het toekomstige verloop van het verkochte advertentievolume.

In dit onderzoek wordt tevens onderzocht of het gebruik van externe variabelen in het voorspellingsmodel kan bijdragen tot een betere voorspelling. Het is bekend dat er verbanden bestaan tussen de advertentievolumes en externe macro- en micro-economische variabelen. Het onderzoek richt zich op het vinden van deze variabelen en het onderzoeken van de bijdrage die deze variabelen kunnen hebben in het voorspellingsproces.

3.2 Domein

Er is een aantal benaderingen mogelijk wat betreft de advertentievolume voorspellingen.

- Voorspellen van de **categoriegroepen.**
  Van alle Nederlandse dagbladen worden de advertentievolumes bijgehouden onderverdeeld in categoriegroepen: nationale, lokale, gerubriceerde en personeelsadvertenties.

- Voorspellen van de volumes van de **advertenties per katern.**
  De Telegraaf brengt speciale katernen uit zoals wonen, reizen en personeel.

- Voorspellen van een enkele **productgroep**.
  Bijvoorbeeld alle advertenties van wasmiddelen.

3.3 Gekozen voorspellingstaak

In overleg met de Telegraaf is gekozen voor het voorspellen van de categoriegroepen (nationaal en lokaal, gerubriceerde en personeelsadvertenties) en het totale advertentievolume. De data is verdeeld in vier datasets:

- Dataset 1 bestaat uit het totale advertentievolume.

- Dataset 2 bestaat uit nationale en lokale advertenties.
  Nationale advertentie*s* zijn van bedrijven/instellingen die landelijk adverteren.
  Lokale advertentie*s* zijn van plaatselijke bedrijven/instellingen.

- Dataset 3 bestaat uit personeelsadvertenties, exclusief de rubrieksadvertenties.

- Dataset 4 bestaat uit gerubriceerde advertenties.
  Rubrieksadvertenties zijn de kleine advertenties gegroepeerd per rubriek.
  (auto's, contactadvertenties ed.)

Op de volgende pagina zijn de datasets weergegeven. (figuur 3.1)

3.3.1 Voorbewerking van de data

De datasets starten in 1981 en eindigen in de laatste periode van 1997. Een jaar bestaat uit 13 vierwekelijkse perioden. Er worden geen maanden gebruikt omdat deze geen gelijk aantal dagen hebben. Perioden van vier weken kunnen beter vergeleken worden. Wanneer in dit verslag maanden genoemd worden, zijn dit vier-wekelijkse perioden.

De datasets worden genormaliseerd tussen 0 en 1 op de onderstaande manier, waarbij n de lengte van dataset $\boldsymbol{x}$ is:

$$\forall i : x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, 1 \leq i \leq n$$

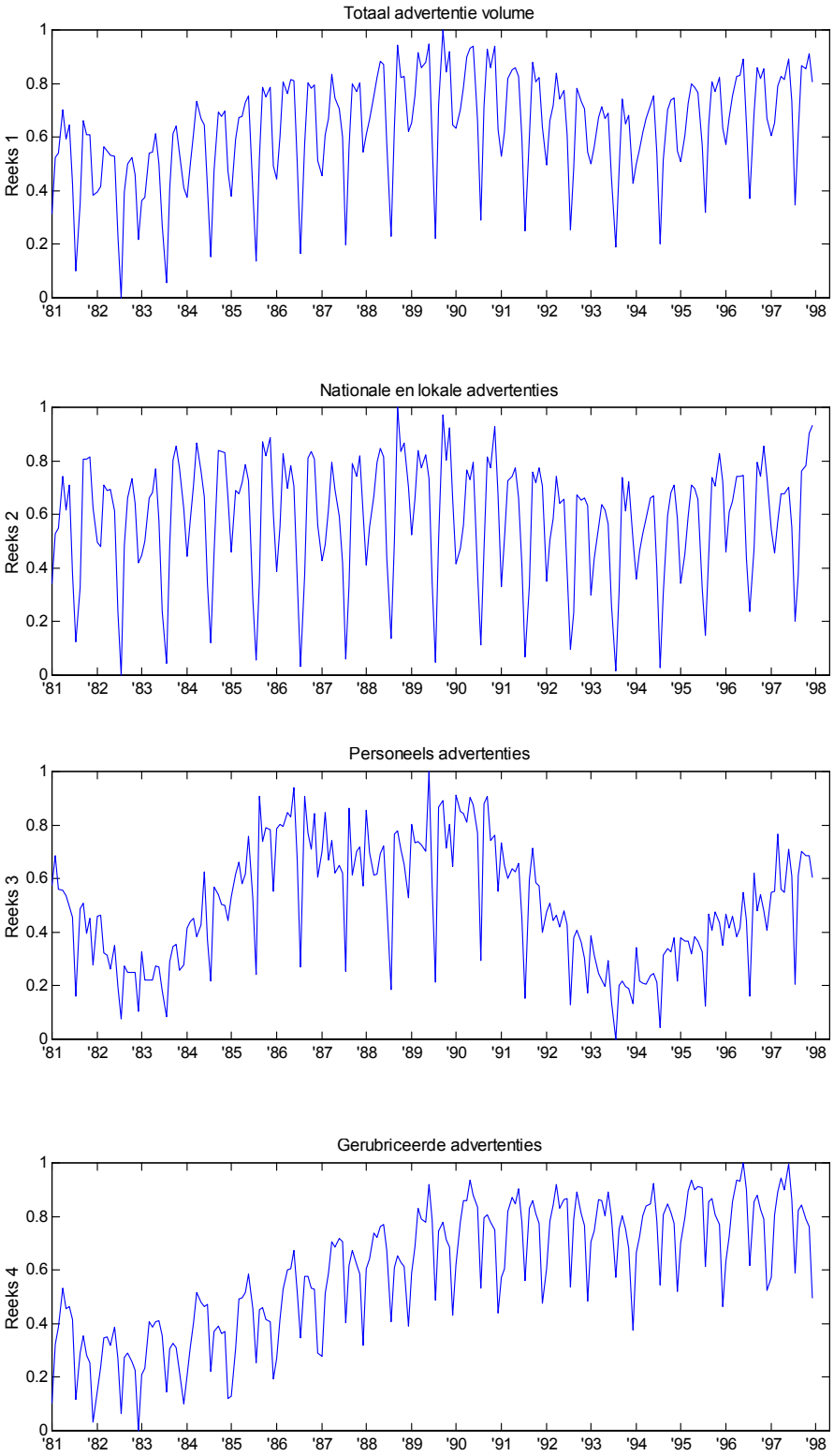
3.4 Externe variabelen

Bij de voorspelling wordt gebruik gemaakt van historische gegevens van de te voorspellen variabele. Tevens kunnen externe variabelen in het voorspellingsmodel opgenomen worden, wanneer de verwachting bestaat dat deze de te voorspellen variabele beïnvloeden.

Elf externe variabelen worden betrokken in het onderzoek. Deze data is afkomstig van het Centraal Bureau voor de Statistiek. (CBS) Gekozen is voor een aantal variabelen met een verondersteld verband met de advertentievolumes en een aantal conjunctuurvariabelen die het economisch klimaat weergeven.

3.4.1 Gebruikte externe variabelen

- Binnenlandse consumptieve bestedingen
- Voedings- en genotmiddelen
- Duurzame consumptiegoederen
- Overige goederen en diensten
- Geregistreerde werklozen
- CBS-koersindex
- Spaartegoeden
- Dollarkoers
- Goudprijs
- Daggeldrente
- Hypotheekrente



Figuur 3.1: De genormaliseerde datasets.

Binnenlandse consumptieve bestedingen

De totale binnenlandse consumptieve bestedingen hebben betrekking op het gebruik van goederen en diensten door gezinshuishoudingen.

Voedings- en genotmiddelen

De binnenlandse consumptie door gezinnen van voedings- en genotmiddelen die in Nederland zijn geproduceerd of ingevoerd.

Duurzame consumptiegoederen

De binnenlandse consumptie door gezinnen van duurzame consumptiegoederen. Duurzame consumptiegoederen zijn gebruiksgoederen die doorgaans langer dan een jaar meegaan. De afbakening 'langer dan een jaar' ligt echter niet zo scherp. Zo horen ook modegevoelige artikelen als kleding en schoenen tot de duurzame consumptiegoederen.

Overige goederen en diensten

De categorie overige goederen en diensten bestaat uit alle diensten plus de consumptiegoederen die niet in een van de andere categorieën zijn opgenomen. Hiertoe behoren de vaste woonlasten, dat wil zeggen de bestedingen aan woningdiensten plus het verbruik van gas, elektriciteit en water. Verder vallen hieronder de bestedingen aan uiteenlopende zaken als motorbrandstoffen, verkeersdiensten en gezondheidsdiensten.

Geregistreerde werkloosheid

Onder geregistreerde werklozen worden verstaan de bij een arbeidsbureau ingeschreven personen van 15 tot 64 jaar die geen betaald werk hebben en beschikbaar zijn voor een functie van twaalf uur per week of meer.

CBS-koersindex

De CBS-koersindex geeft de waardeontwikkeling weer van alle op de beurs genoteerde Nederlandse gewone aandelen, waarbij de invloed van veranderingen als gevolg van kapitaalmutaties is geëlimineerd.

Spaartegoeden

De spaartegoeden zijn de tegoeden op spaarrekeningen en deposito's van particulieren.

Dollarkoers

De wisselkoers van 1 Amerikaanse dollar, uitgedrukt in guldens.

Goudprijs

De prijs van 1 gram fijn goud in guldens.

Daggeldrente

Het tarief voor interbancaire daggeldleningen zonder onderpand. Een indicator voor de korte termijn rente.

Hypotheekrente

Voor de hypotheekrente is het gemiddelde rentepercentage van alle nieuw ingeschreven hypotheken op woonhuizen en combinaties van woonhuis/bedrijfspand opgenomen.

3.4.2 Herbemonstering van de externe datasets

De externe datasets zijn afkomstig van het CBS en bevatten de jaren 1981 tot en met 1997.

De externe datasets zijn op maandbasis, terwijl de te voorspellen datasets 13 vierwekelijkse perioden per jaar hebben. Om de datasets en de externe data goed in een voorspellingsmodel te kunnen gebruiken, moeten ze dezelfde tijdsintervallen hebben. De externe datasets moeten daarom ook 13 perioden per jaar hebben en zijn daartoe herbemonsterd van 12 naar 13 perioden per jaar.

Deze herbemonstering wordt uitgevoerd door een anti-aliasing FIR-filter. [Parks97] Dit filter gebruikt de gewogen som van *n* punten aan beide kanten van de huidige sample van de originele tijdreeks $\boldsymbol{e}^j$ om een punt van de herbemonsterde tijdreeks te bepalen.

De procedure is als volgt:

- Instelling optimale filter parameters *h* door kleinste-kwadraten minimalisatie en gebruikmakend van *n, p en q.* De tijdsinterval van de originele tijdreeks is *p*, de tijdsinterval van de herbemonsterde tijdreeks is *q*.
- Upsampling met waarde *p*
- Toepassing FIR-filter met parameters *h*
- Downsampling met waarde *q*

Wanneer *l* de lengte van $\boldsymbol{e}^j$ is, dan is de lengte *k* van de herbemonsterde tijdreeks:

$$k = l \cdot \frac{p}{q}$$

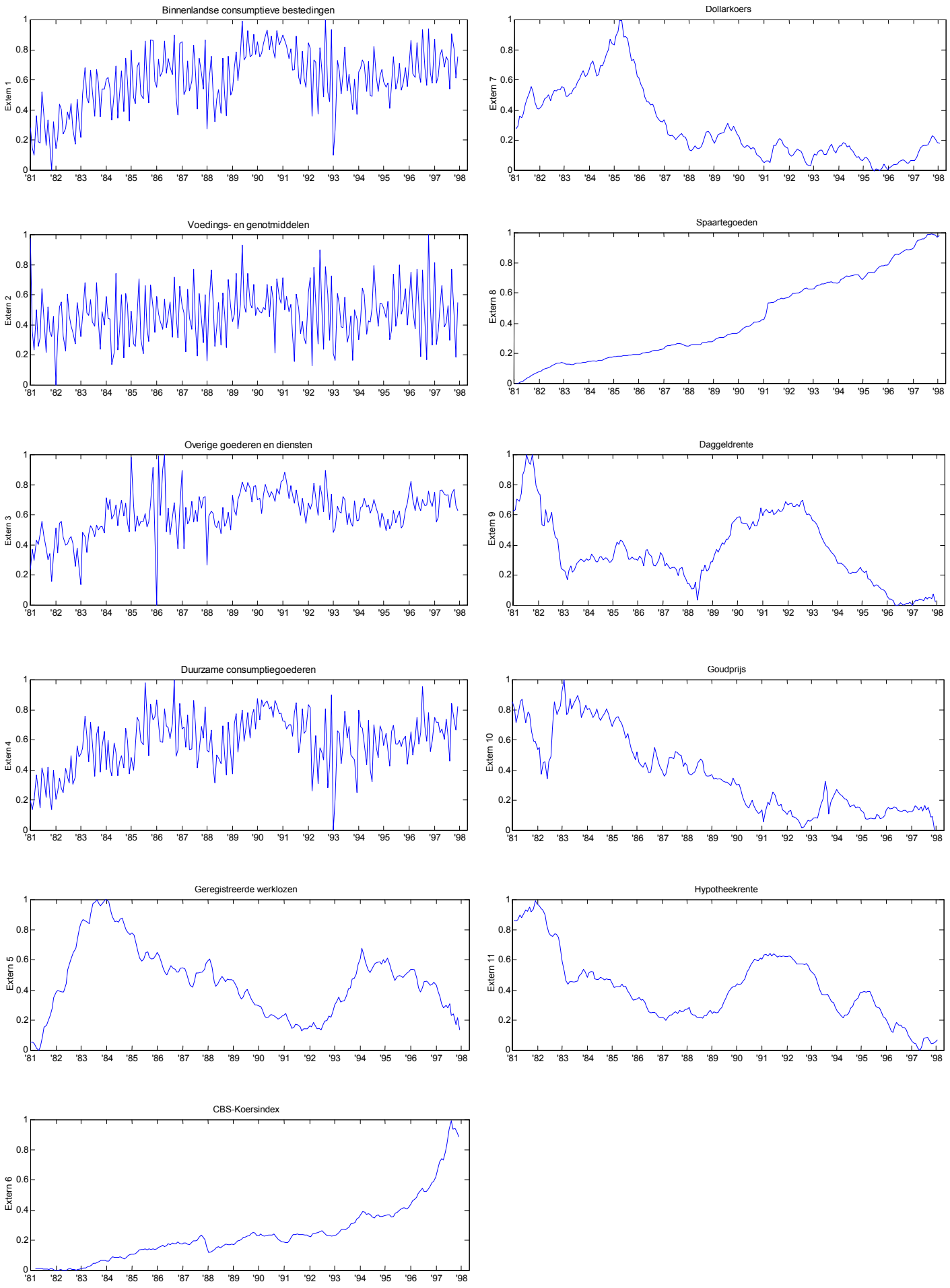
3.4.3 Normalisatie van de externe datasets

De externe datasets worden op dezelfde manier genormaliseerd als de te voorspellen datasets. De datasets worden geschaald tussen 0 en 1 op de onderstaande manier, hierbij is *n* de lengte van de dataset $\boldsymbol{e}^j$ en is *m* het aantal externe variabelen:

$$\forall j \forall i : e_i'^j = \frac{e_i^j - e_{\min}^j}{e_{\max}^j - e_{\min}^j}, 1 \le i \le n, 1 \le j \le m$$

3.4.4 Grafieken van externe variabelen

Op de volgende pagina worden de externe variabelen getoond.



Figuur 3.2: Externe variabelen

4. Voorspellingsmethoden

4.1 Inleiding

In dit hoofdstuk worden de onderzochte modellen beschreven. In paragraaf 4.2 worden lineaire methoden besproken. Paragraaf 4.3 behandelt het trend-fitting-methode. In paragraaf 4.4 worden neurale netwerken beschreven en in 4.5 wordt de nearest-neighbour-methode behandeld.

Traditioneel worden lineaire modellen veel gebruikt bij het voorspellen van tijdreeksen. Voorbeelden van deze modellen zijn onder andere: AR, ARX en Box-Jenkins. Dit zijn statistische autoregressieve modellen waarmee in het verleden goede resultaten behaald zijn.

Neurale netwerken kunnen door hun veelzijdige toepasbaarheid ook gebruikt worden bij het voorspellen van tijdreeksen. Neurale netwerken zijn niet-lineair en de veronderstelling is dat zij een beter resultaat geven bij complexe datasets dan de lineaire modellen.

Mijn stagebedrijf maakt veel gebruik van nearest-neighbour-modellen als voorspellings-techniek. Een deel van mijn onderzoek is gericht op het onderzoeken in hoeverre deze modellen geschikt zijn voor het voorspellen van tijdreeksen.

Daarnaast zal de trend-fitting-methode ontwikkeld worden. Dit is een eenvoudige voorspellingsmethode, die gebruik maakt van een overeenkomstig jaargedrag.

4.1.1 Voorspellingsmethoden

De volgende typen methoden kunnen onderscheiden worden [Gaynor94]:

- Univariate, waarbij voorspeld wordt op basis van slechts historische data.
- Multivariate of causaal, waarbij de voorspelling ook afhangt van externe variabelen.

Tevens kunnen modellen lineair of niet-lineair zijn [Gunst97]:

- Lineair: statistisch model waarbij de voorspelling een lineaire functie van de data is.
- Niet-lineair: de voorspelling is een niet-lineaire functie van de data.

De verschillende methoden worden in het onderstaande overzicht getoond:

	Uni-variate	**Multi-variate**
Lineair	AR	ARX Box-Jenkins
Niet-lineair	Trend-fitting Neuraal, kNN	Neuraal, kNN

Tabel 4.1: Overzicht tijdreeksvoorspellings-methoden.

4.1.2 Onderzochte methoden

In dit onderzoek wordt van een drietal methoden de toepasbaarheid onderzocht voor het gebruik als tijdreeksvoorspellingsmodel. De onderzochte methoden zijn:

- Trend-fitting-methode
- Neurale netwerken
- Nearest-neighbour-methode (kNN)

De resultaten worden vergeleken met standaard tijdreeksvoorspellingsmethoden:

- ARX
- Box-Jenkins

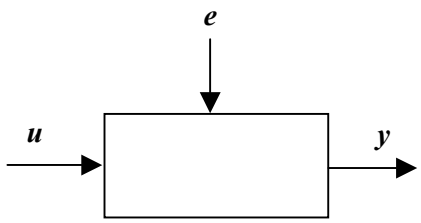
4.2 Lineaire modellen

4.2.1 Inleiding

Een lineaire model maakt een mathematisch model van een dynamisch systeem op basis van gemeten data. Dit wordt gedaan door aanpassing van parameters van een gegeven model totdat de uitvoer van het model de gemeten uitvoer zo dicht mogelijk is genaderd. Het model kan vervolgens getest worden door de uitvoer te vergelijken met de gemeten data. Vervolgens kan het model gebruikt worden om uitvoer van het systeem te voorspellen. In dit verslag worden lineaire modellen gebruikt voor het voorspelen van tijdreeksen met als invoer externe variabelen.

4.2.2 Dynamische systemen

Dynamische systemen [Ran97] worden gekenmerkt door een verzameling variabelen en hun wederzijdse afhankelijkheden in de tijd. In het onderstaande figuur wordt een schematisch overzicht van een dynamisch systeem gegeven.



Figuur 4.1: Dynamisch systeem

In dit systeem is ***u*** de invoer, ***y*** de uitvoer en ***e*** een ruissignaal. Het modelleringsprobleem bestaat uit het vinden op welke wijze deze drie signalen zich tot elkaar verhouden. In het algemeen wordt de uitvoer *y* op een zeker tijdstip bepaald door een lineaire combinatie van aantal historische waarden van de invoer en een aantal historische waarden van de uitvoer. Daarnaast in de uitvoer afhankelijk van de foutbron ***e***.

4.2.3 Lineair model als tijdreeksvoorspeller

Autoregressieve modellen, gebaseerd op een dynamisch systeem kunnen worden toegepast op het voorspellen van tijdreeksen. Hiervoor is een aantal modellen ontwikkeld, waaronder AR, ARX, ARMA, ARIMA en Box-Jenkins. [Gaynor94] Deze modellen hebben het autoregressieve karakter gemeenschappelijk en bezitten al dan niet een moving-average (MA), integrerend (I) of exogeen (X) element.

Het exogene element heeft betrekking op de aanwezigheid van één of meerdere invoer-variabelen in het model. In dit verslag wordt als invoer één of meerdere externe variabelen gebruikt. De voorspelling wordt naast de historische waarden van de doelvariabele door deze externe variabelen bepaald.

De AR, ARX en Box-Jenkins modellen worden hieronder beschreven.

4.2.4 AR-model

Het AR-model is een autoregressief model zonder invoer-variabelen. De uitvoer op een zeker tijdstip wordt slechts bepaald door een aantal historische waarden van de uitvoer zelf en een foutbron. De uitvoer *y* op tijdstip *t* wordt weergegeven door onderstaande formule:

$$y(t) = \sum_{i=1}^{na} \alpha_i y(t-i) + e(t)$$

Waarin *na* de orde van de uitvoer is en α een vector met modelparameters. De vector α wordt bepaald door middel van een kleinste kwadraten optimalisatie. [Gunst95] Dit model kan ook weergegeven worden middels de onderstaande bondige formule:

$$\mathbf{A}(q)y(t) = e(t)$$

4.2.5 ARX-model

Het ARX-model in een autoregressief model met één of meer invoer-variabelen. De uitvoer op een zeker tijdstip wordt bepaald door een aantal historische waarden van de uitvoer zelf, een aantal historische waarden van de invoer en een foutbron. De uitvoer *y* op tijdstip *t* wordt weergegeven door onderstaande formule:

$$y(t) = \sum_{i=1}^{na} \alpha_i y(t-i) + \sum_{j=1}^{nb} \beta_j u(t-j) + e(t)$$

Waarin *na* de orde van de uitvoer is, *nb* de orde van de invoer, *nk* de time-delay en α en β vectoren met modelparameters. De vectoren α en β worden net zoals bij het AR-model bepaald door middel van een kleinste kwadraten optimalisatie.

Dit ARX-model kan ook weergegeven worden door de onderstaande formule:

$$\mathbf{A}(q)y(t) = \mathbf{B}(q)u(t-nk) + e(t)$$

In deze formule wordt het autoregressieve (AR) element gerepresenteerd door **A***(q)y(t)* en het exogene deel wordt weergegeven door **B***(q)u(t-nk)*.

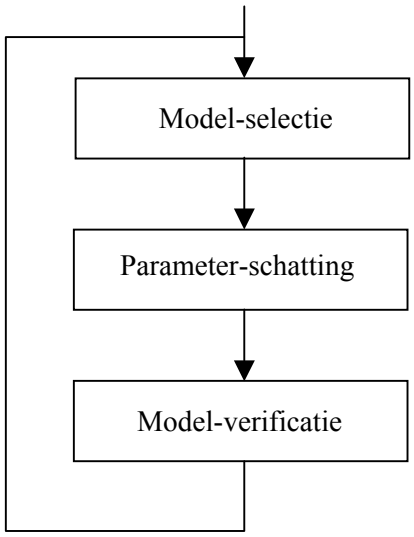
4.2.6 Box-Jenkins

Box-Jenkins [Gaynor94] is een afleiding van een autoregressief, integrerend, moving-average model (ARIMA) en wordt weergegeven door de onderstaande formule:

$$\mathbf{y}(t) = \frac{B(q)}{F(q)}\mathbf{u}(t-nk) + \frac{C(q)}{D(q)}\mathbf{e}(t)$$

Het Box-Jenkins-model voert de onderstaande stappen uit in een iteratieve procedure:

- **Model-selectie**: Hierin wordt gekozen voor één of meerdere componenten van het ARIMA-model.
- **Parameter-schatting**: De optimale model-parameters worden ingesteld.
- **Model-verificatie**: Het model en de parameters worden getest.



Figuur 4.2: Box-Jenkins modellerings diagram

Na de model-verificatie kan het proces zonodig herhaald worden om tot een optimaal model te komen.(zie figuur 4.2)

4.3 Trend-fitting-methode

4.3.1 Inleiding

Bij deze methode wordt uitgegaan van een tijdreeks waarvan het jaargedrag een grote mate van constantheid vertoont. De voorspelling van de tijdreeks wordt bepaald door de voorspelling van de trend te combineren met een gewogen gemiddeld jaargedrag.

4.3.2 Methode beschrijving

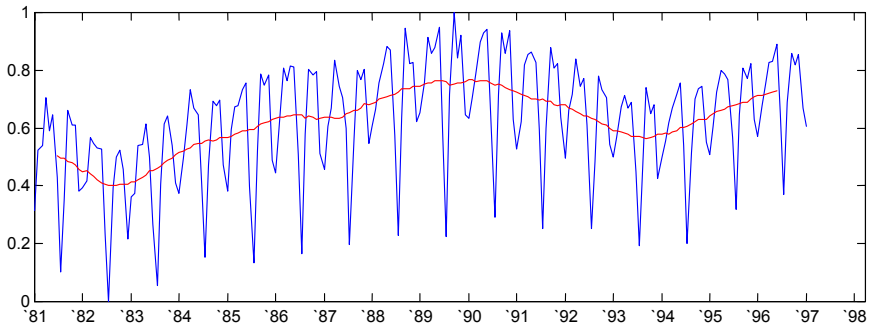
De globale trend van de tijdreeks wordt bepaald door de moving-average van de tijdreeks te berekenen. Deze trend word vervolgens voorspeld door middel van curve-fitting. Tevens wordt het gemiddelde jaargedrag bepaald en deze wordt gecombineerd met de trend-voorspelling tot de eigenlijke voorspelling van de tijdreeks.

4.3.3 Moving-average berekening

Er wordt gebruik gemaakt van de gecentreerde moving-average [Gaynor94] van tijdreeks $\boldsymbol{x}$ op tijdstip t met oneven averaging-periode L en deze wordt als volgt berekend:

$$CMA_t = \frac{1}{L}(x_{t-((L-1)/2)} + .. + x_t + .. + x_{t+((L-1)/2)})$$

Van de tijdreeks wordt de gecentreerde moving-average bepaald door voor elk punt het gemiddelde te berekenen van zichzelf en de *(L-1)/2* voor- en achterliggende punten. Hierdoor kan van dit aantal punten aan het begin en eind van de tijdreeks geen juiste moving-average berekend worden. (zie figuur 4.3) De berekende moving-average geeft de globale trend weer.



Figuur 4.3: De tijdreeks en de moving-average.

4.3.4 Trend-voorspelling

De trend wordt voorspeld met behulp van een polynomen-curve-fitting. Het onderstaande polynoom van orde *n* wordt gefit aan een deel van de moving-average.

$$p_1 t^n + p_2 t^{n-1} + \ldots p_n t + p_{n+1}$$

Hierbij is p_i een polynoomcoëfficiënt en n is de orde van het polynoom. De lengte van het deel van de moving-average waarop de curve-fitting gebaseerd wordt, is bepaald door de parameter *fitlengte*.

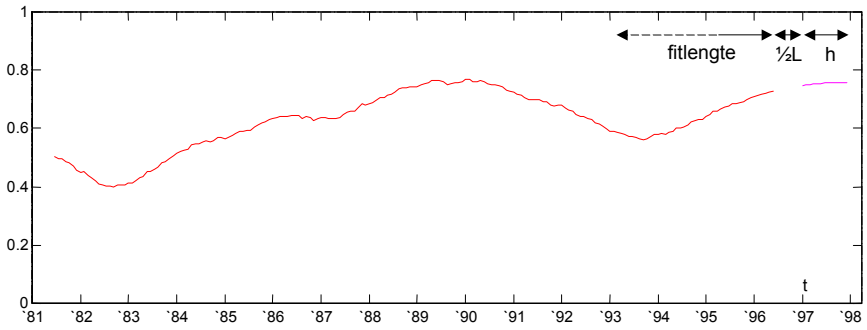
De curve-fitting bestaat uit het bepalen van de polynoomcoëfficiënten. De coëfficiënten-vector $\boldsymbol{p}$ wordt berekend door een kleinste kwadraten schatter (LSE) [Gunst95]:

$$\boldsymbol{p} = LSE\left(CMA_{(t-\frac{L}{2})},\ CMA_{(t-\frac{L}{2})-1},\ \ldots,\ CMA_{(t-\frac{L}{2})-fitlengte} \right)$$

Vervolgens wordt deze vector gebruikt om de trendvoorspelling te berekenen. De voorspelling van de trend (*PRED*) op tijdstip *i* is als volgt:

$$\forall i:\ PRED_i = p_1 i^n + p_2 i^{n-1} + \ldots p_n i\ + p_{n+1},\quad t \le i \le t + h$$

Hierbij is *h* de lengte van de voorspelling. In het onderstaande figuur is een voorspelling te zien met een tweede graads polynoom.



Figuur 4.4: Voorspelling van de trend door curve-fitting met n=2.

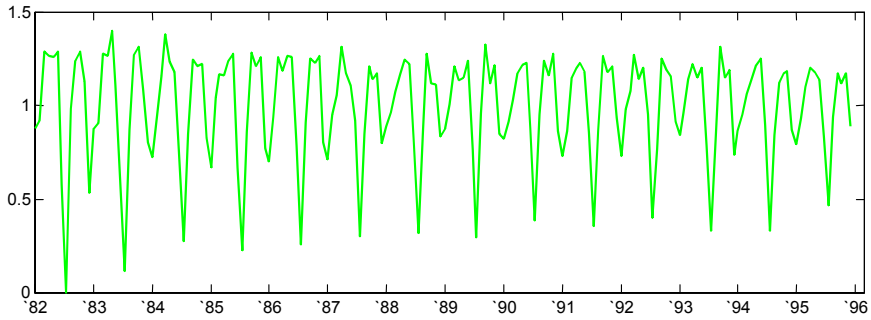
4.3.5 Jaargemiddelde

Om het constante jaarpatroon uit de tijdreeks te isoleren, wordt de globale trend van de tijdreeks verwijderd. Hiervoor wordt één van onderstaande methoden gebruikt:

Quotiënt-methode: $$x'_t = \frac{x_t}{CMA_t}$$

Verschil-methode: $$x'_t = x_t - CMA_t$$

In de experimenten wordt onderzocht welke methode de beste resultaten geeft. In deze paragraaf wordt uitgegaan van het gebruik van de quotiënt-methode. Het resultaat van trendverwijdering is een tijdreeks met slechts irreguliere- en seizoensvariaties waarvan het verloop binnen één periode is redelijk constant is. (zie figuur 4.5)



Figuur 4.5: Tijdreeks zonder trend

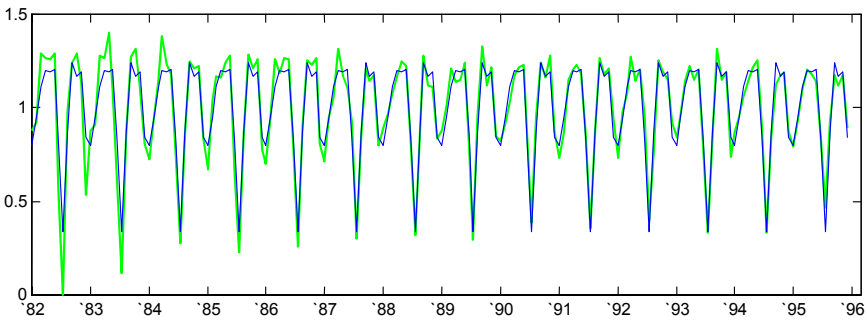
Vervolgens wordt het gemiddelde jaargedrag bepaald door de perioden te middelen over de jaren, hetzij lineair of gewogen. Bij het gewogen gemiddelde tellen de latere jaren zwaarder mee. Het voordeel hiervan is dat de kans groot is dat het jaargedrag van de te voorspellen periode de grootste overeenkomst heeft met recente jaren.

De gemiddelden worden op de onderstaande manier berekend, waarbij n het aantal jaren, j het jaarnummer, p het aantal perioden en m het periodenummer is:

$$\forall m : AVG_m = \frac{1}{n}\sum_{j=0}^{n-1} x_{jp+m}, \quad 1 \le m \le p$$

$$\forall m : AVGW_m = \frac{\sum_{j=0}^{n-1} j x_{jp+m}}{\sum_{i=1}^{n} i}, \quad 1 \le m \le p$$

In onderstaand figuur is het berekende gemiddelde jaargedrag en de tijdreeks zonder de trend getoond.



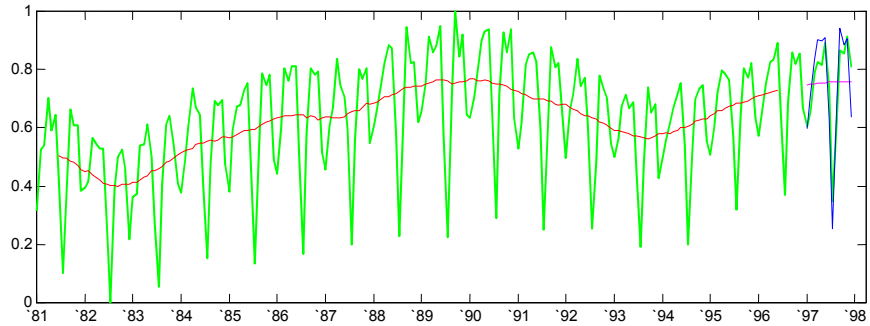
Figuur 4.6: Gemiddeld jaargedrag en tijdreeks zonder trend

4.3.6 Voorspelling

Tenslotte wordt de voorspelling verkregen van de originele tijdreeks $\boldsymbol{x}$ op tijdstip i door het gemiddelde jaargedrag te vermenigvuldigen met de voorspelde moving-average.

$$\hat{x}_i = AVG_i \times PRED_i$$

Dit leidt tot onderstaande grafiek, waarin de originele tijdreeks, de trend, de trend-voorspelling en de tijdreeksvoorspelling zijn weergegeven.



Figuur 4.7: Voorspelling van de tijdreeks door combinatie van trend-voorspelling met het berekende gemiddelde jaargedrag.

4.3.7 De trend-fitting-methode als tijdreeksvoorspeller

Bij het gebruik van de trend-fitting-methode moeten de volgende parameters bepaald worden:

- average-periode L; dit is het aantal punten waarop de berekening van de moving-average gebaseerd wordt.
- fitorde n; n-de graad polynoom.
- fitlengte; dit deel van de originele data wordt gebruikt voor de curve-fitting.
- quotiënt- of verschilmethode bij de trendverwijdering.
- jaargemiddelde; lineair of gewogen.

Deze laatste parameter bepaalt op welke manier het gemiddelde jaargedrag berekend wordt. Bij ‘lineair’ hebben alle maanden een gelijk aandeel. In het geval van ‘gewogen’ tellen de laatste jaren zwaarder mee, daar de laatste jaren waarschijnlijk representatiever zijn voor het voorspelde jaar.

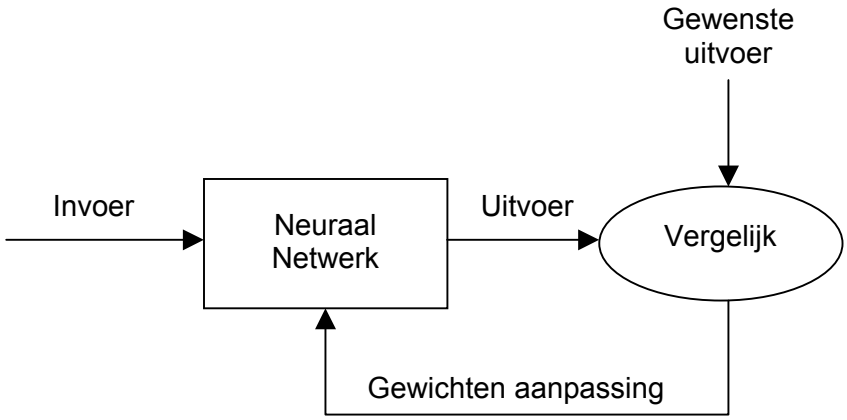
In paragraaf 4.1 is deze methode geclassificeerd als een niet-lineair methode. Dit is correct voor een fitorde $n > 1$. Als $n = 1$ bestaat de trend-fitting uit een lineaire fit waardoor de methode als lineair gezien kan worden.

4.4 Neurale netwerken

4.4.1 Inleiding

Neurale netwerken bestaan uit eenvoudige rekenelementen die parallel werken. Deze elementen zijn geïnspireerd op biologische zenuwcellen. Net zoals in de natuur wordt de netwerkfunctie voor een groot deel bepaald door de verbindingen tussen de elementen. Het netwerk kan getraind worden door de gewichten van de verbindingen tussen de elementen aan te passen.

Het netwerk wordt getraind zodat een bepaalde invoer leidt tot een gewenste uitvoer. Er zijn veel invoer/gewenste-uitvoer-paren nodig om het netwerk te trainen.



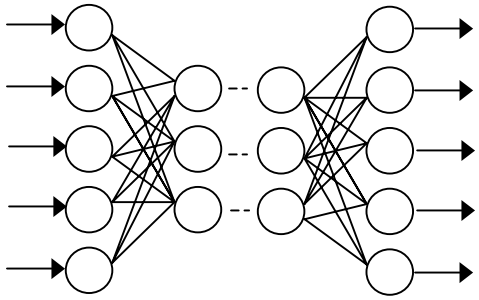
Figuur 4.8: Schematisch overzicht van de werking van een neuraal netwerk.

Generalisatie is een belangrijke eigenschap van een neuraal netwerk. Dit betekent dat een goed getraind netwerk een redelijke uitvoer zal geven op een invoer die niet getraind is. De uitvoer zal overeenkomen met de uitvoer behorende bij een vergelijkbare invoer die wel getraind is. Neurale netwerken kunnen gebruikt worden voor herkennen, classificeren of voorspellen op allerlei gebieden.

4.4.2 Multilayer-feedforward neuraal netwerk

Een veel gebruikte neuraal netwerk topologie is het multilayer-feedforward-netwerk [Kröse93]. Het bestaat uit een invoerlaag, één of meer tussenlagen (hidden layers) en een uitvoerlaag. Elke laag bestaat uit een aantal neuronen die een signaal doorgeven aan neuronen in de volgende laag. (zie figuur 4.9)

De verbindingen tussen de neuronen zijn gewogen en kunnen zowel positief als negatief zijn. De netwerken worden feedforward genoemd, omdat de signalen slechts in één richting kunnen gaan, namelijk van de invoerlaag via de tussenlaag naar de uitvoerlaag.



Figuur 4.9: Neuraal netwerk met één invoerlaag, één of meer hidden lagen en één uitvoerlaag.

Het netwerk wordt op twee manieren gebruikt. Tijdens het voorspellen of classificeren wordt een invoervector aan het netwerk aangeboden. Via de gewogen verbindingen propageert het signaal verder naar de tussenlaag (of lagen) en bereikt daarna de uitvoerlaag. Tijdens de propagatie wordt van elk neuron de activatiewaarde berekend, die de som is van de gewogen invoer. De uitvoer is tenslotte de vector met activatiewaarden van de uitvoerneuronen.

Het netwerk moet echter getraind worden voordat een goede voorspelling gedaan kan worden. Dit trainen van het netwerk en bestaat uit het aanpassen van de gewichten.

4.4.3 Trainen

Het netwerk wordt getraind via het backpropagation-algoritme [Kröse93]. Een set van invoervectoren waarvan de uitvoer bekend is wordt aan het netwerk gepresenteerd. De invoervector propageert door het netwerk en geeft een bepaalde uitvoervector. Deze uitvoervector wordt vergeleken met de gewenste uitvoer en dat leidt tot een fout voor elk uitvoerneuron. Vervolgens worden deze fouten terug gepropageerd (backpropagation) naar voorliggende lagen. Aan de hand van de fouten worden de gewichten aangepast.

Op analoge wijze worden alle trainingsvectoren geleerd. Het trainen van het netwerk is een iteratieve parameterschattingsmethode die de optimale gewichten van het netwerk zoekt waarvoor de fout tussen de uitvoer en de gewenste uitvoer minimaal is.

De set van trainingsvectoren wordt veelal gedeeld in twee sets: de trainset en de testset. De trainset wordt gebruikt om het netwerk te trainen. Daarna wordt met de testset de prestatie van het netwerk getest. Omdat de testset niet gebruikt is om het netwerk te trainen, wordt op deze manier het generaliserend vermogen van het netwerk aangesproken. Dit vermogen is belangrijk voor het netwerk als classificerend of voorspellend systeem.

4.4.4 Trainingsfuncties

Diverse backpropagation trainingsfuncties zijn ontwikkeld, onder andere:

- Gradient descent
- Gradient descent met momentum
- Conjugate gradient
- Levenberg-Marquardt

- Gradient descent [Kröse93] is de eenvoudigste backpropagation trainingsfunctie. Het verandert de gewichten in de richting waarin de fout het snelste afneemt. De gewichtsverandering is proportioneel met de negatieve gradiënt van de fout.

- Aan deze trainingsfunctie kan een zogenaamd momentum [Kröse93] worden toegevoegd. De gewichten worden dan niet slechts aangepast naar de negatieve gradiënt, maar ook naar het recente verloop in het foutlandschap. De gewichtenaanpassing is dan tevens afhankelijk van de vorige gewichtenaanpassing. Op deze manier is een snellere training mogelijk en wordt de kans verkleind dat het trainen eindigt in een lokaal minimum.

- Conjugate gradient [Hagan96] en Levenberg-Marquardt [Hagan94] zijn trainingsfuncties die ontwikkeld zijn met het doel het trainen van het netwerk te versnellen. De conjugate gradient trainingsfunctie gaat bij de gewichtenaanpassing niet uit van één richting zoals gradient descent, maar van een verzameling richtingen. Deze verzameling heeft dezelfde oorsprong en de minimalisatie in een richting wordt niet belemmerd door die in één van de andere richtingen.

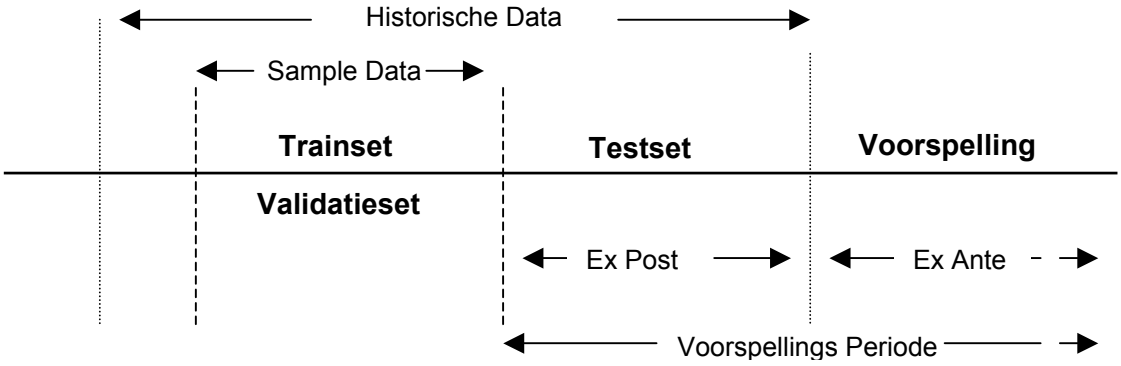
- De Levenberg-Marquardt trainingsfunctie is een variant van de Newton methode [Demuth97]. De Newton trainingsfunctie is een snelle optimalisatiemethode waarvoor de Hessian-matrix met tweede afgeleiden van de netwerkfouten berekend moet worden. Deze berekening kost echter veel rekenwerk. Levenberg-Marquardt benadert de Hessian-matrix, wat minder rekenwerk kost en wordt daarom ook een quasi-Newton methode genoemd. De Levenberg-Marquardt trainingsfunctie gebruikt bij aanvang gradient descent en schakelt zo snel mogelijk over op de Newton methode, omdat deze sneller en nauwkeuriger is in de buurt van een minimum. Deze trainingsfunctie kan 10 tot 100 maal sneller zijn dan standaard gradient descent. Een nadeel is dat veel geheugen nodig is om de Hessian-matrix te benaderen.

4.4.5 Overfitting en ‘early stopping’

Een belangrijke eigenschap van een neuraal netwerk is het generaliserende vermogen. Het netwerk wordt getraind met een set historische gegevens en de fout op deze set wordt geminimaliseerd. Doorslaggevend is echter de fout op nieuwe gegevens waarmee niet getraind is en het netwerk moet generaliseren.

De prestaties van het neurale netwerk zijn onder andere afhankelijk van de hoeveelheid trainingsgegevens en de netwerkgrootte. Als het netwerk te groot is en er te veel parameters (gewichten) geoptimaliseerd moeten worden, kan overfitting optreden. Als het netwerk echter te klein is, zal het niet goed kunnen voorspellen. Bij overfitting zal de fout op de trainset heel klein worden, maar wanneer met nieuwe gegevens voorspeld moet worden is de fout (op de testset) groot. Het netwerk heeft een functie geleerd die specifiek is voor de trainingsset, maar niet voor de testset en is dus niet in staat goed te generaliseren.

Een methode om overfitting te voorkomen en de generalisatie te verbeteren is ‘early stopping’. De beschikbare gegevens worden dan niet zoals gewoonlijk in een trainingsset en een testset verdeeld, maar tevens in een validatieset. (zie figuur 4.10) Het netwerk wordt getraind met data uit de trainingsset en tijdens het trainen wordt de fout op de validatieset gecontroleerd.



Figuur 4.10: Gebruikte datasets door het neurale netwerk

Tijdens de training zal de fout op de validatieset net als de fout op de trainingsset afnemen. Als het netwerk echter begint te overfitten, zal de fout op de trainingsset blijven afnemen, terwijl de fout op de validatieset toeneemt. Als deze stijging voortzet, zal het trainen stoppen en de gewichten gekozen worden behorende bij het minimum van de fout op de validatieset. Met de testset wordt vervolgens de prestatie van het netwerk getest.

4.4.6 Neurale netwerken als tijdreeksvoorspeller

Neurale netwerken kunnen gebruikt worden bij het voorspellen van tijdreeksen. Een aantal zaken is dan van belang:

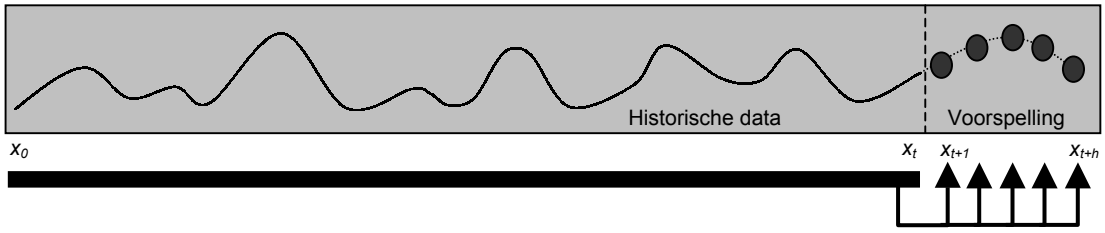
- **voorspellingslengte;** de lengte *h* van het te voorspellen stuk tijdreeks
- **voorspellingstype**; per punt of per periode
- **trainingsmethode**; direct of iteratief (in het geval van een puntvoorspelling)
- **invoer**; welke tijdreekspunten als invoer te gebruiken
- **externe variabelen**; welke externe variabelen te gebruiken

4.4.6.1 Voorspellingstype

Wanneer een stuk tijdreeks met lengte *h*, bestaande uit een aantal punten (x_{t+1}, x_{t+2},...,x_{t+h}) voorspeld moet worden, kan het netwerk getraind worden om in één keer de gehele periode of punt voor punt te voorspellen. In het eerste geval heeft het netwerk een aantal uitvoerneuronen dat gelijk is aan de voorspellingslengte. In het geval van de periodevoorspelling heeft het netwerk één uitvoerneuron en moeten *h* deelvoorspellingen uitgevoerd worden om de gehele voorspelling te doen.

4.4.6.2 Periode-voorspelling

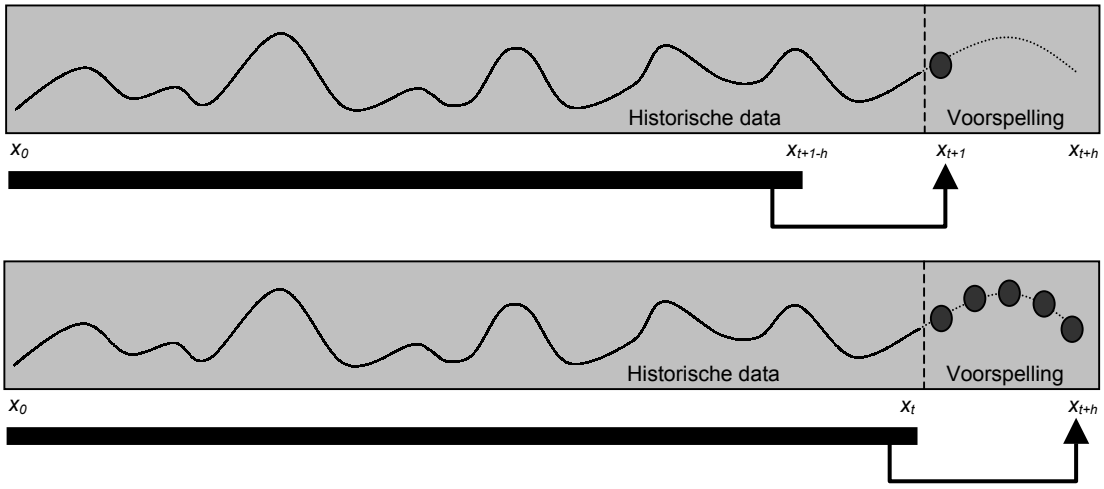
In het geval van de periode-voorspelling wordt alle data van x_0 tot en met x_t gebruikt voor de voorspelling in één keer van de gehele periode. In het onderstaande figuren wordt de gebruikte data grafisch weergegeven door de zwarte balk.



Figuur 4.11: Gebruikte historische data bij periode-voorspelling.

4.4.6.3 Punt-voorspelling (directe methode)

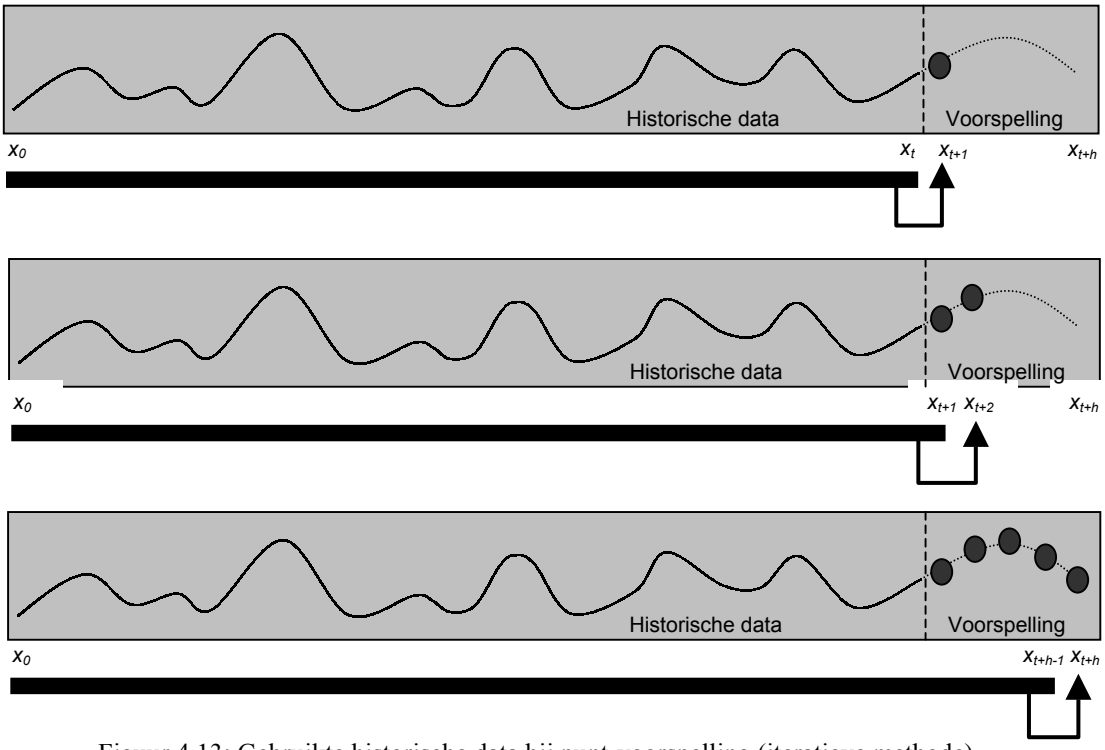
Bij het punt voor punt voorspellen kan een directe of iteratieve methode gebruikt worden. Beide methodes hebben h deelvoorspellingen nodig voor de complete voorspelling. De directe methode voorspelt x_{t+1} gebruikmakend van data tot en met x_{t+1-h}. Het laatste punt van de voorspelling x_{t+h} wordt voorspeld met data tot en met x_t . Dit betekent dat steeds h punten vooruit voorspeld worden. (zie figuur 4.12)



Figuur 4.12: Gebruikte historische data bij punt-voorspelling (directe methode).

4.4.6.4 Punt-voorspelling (iteratieve methode)

De iteratieve methode voorspelt x_{t+1} gebruikmakend van data van x_0 tot en met x_t.
Vervolgens wordt deze voorspelling ook gebruikt om het volgende punt x_{t+2} te voorspellen.
Op deze manier wordt elk punt van de periode voorspeld, waarbij het laatste punt van de voorspelling x_{t+h} wordt voorspeld met data van x_0 tot en met x_{t+h-1} .

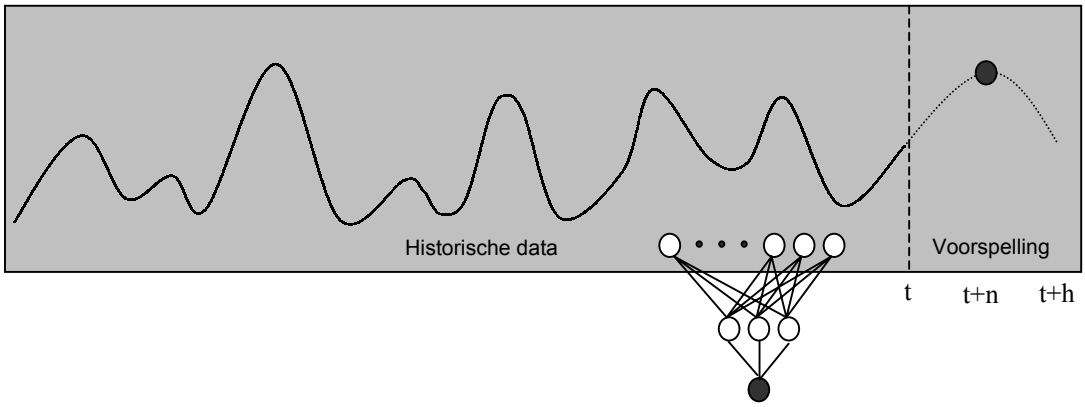


Figuur 4.13: Gebruikte historische data bij punt-voorspelling (iteratieve methode).

Bij deze methode wordt dus steeds één punt voorspeld. Deze laatste voorspelling is bijna geheel gebaseerd op eerdere voorspellingen, wat de betrouwbaarheid van de voorspelling kan beïnvloeden. De directe voorspellingsmethode gebruikt alleen historische data, maar deze methode heeft als nadeel dat *h* punten vooruit voorspeld worden en dus niet altijd de meest recente data gebruikt wordt.

4.4.6.5 Invoer

Bij het voorspellen worden historische waarden als invoer gebruikt. Bepaald moet worden welke en hoeveel van deze punten gebruikt gaan worden. In het onderstaande figuur is een aantal opeenvolgende maanden gebruikt, deze representeren het recente verloop van de tijdreeks.

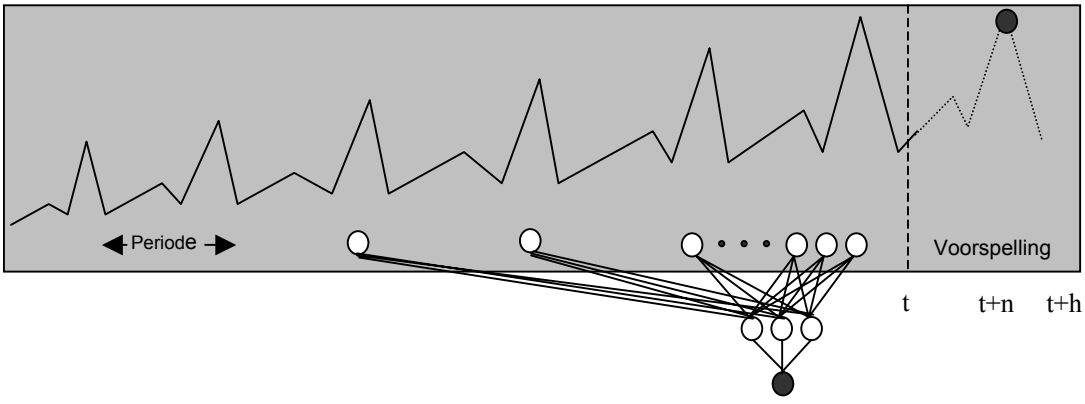


Figuur 4.14: Een aantal opeenvolgende maanden als invoer voor het netwerk.

Het neurale-netwerk-model *f* voor een willekeurig punt $\hat{x}_{t+n}$ uit de voorspellingsperiode is dan als volgt, waarbij *m* het aantal gebruikte maanden en *h* de lengte van de voorspellingsperiode is:

$$\hat{x}_{t+n} = f(x_{(t+n-h)}, x_{(t+n-h)-1}, x_{(t+n-h)-2}, \ldots x_{(t-+n-h)-(m-1)})$$

Naast een aantal opeenvolgende maanden kan ook een aantal maanden van voorgaande jaren als invoer gebruikt worden. Deze punten vertegenwoordigen de globale trend in de reeks en worden vanaf nu de 'gebruikte jaren' genoemd.



Figuur 4.15: Een aantal opeenvolgende maanden en twee voorgaande jaren als invoer voor het netwerk.

Het neurale-netwerk-model *f* voor een willekeurig punt $\hat{x}_{t+n}$ uit de voorspellingsperiode is dan als volgt:

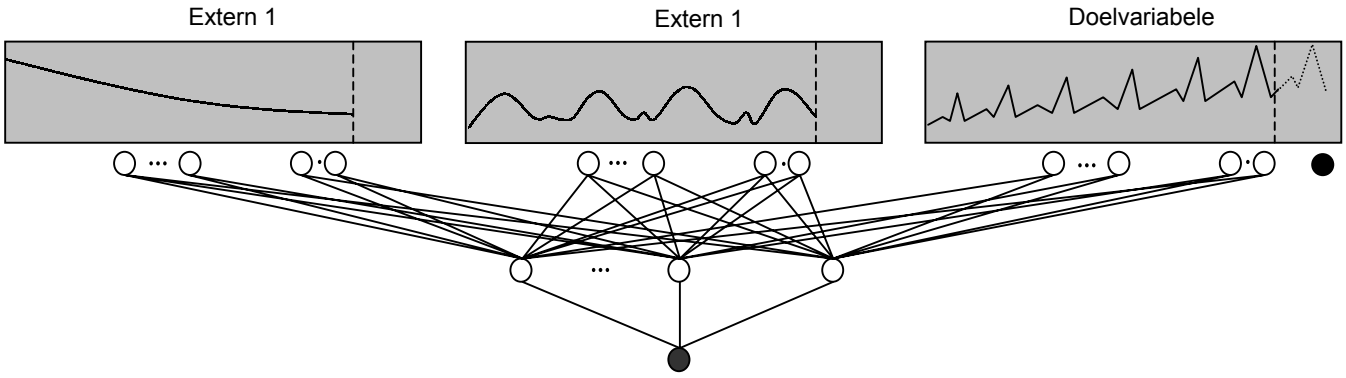
$$\hat{x}_{t+n} = f(x_{(t+n-h)}, x_{(t+n-h)-1}, x_{(t+n-h)-2}, \ldots x_{(t-+n-h)-(m-1)}, x_{(t-+n-h-m)}, x_{(t-+n-h-m)+p}, x_{(t-+n-h-m)+2p}, \ldots x_{(t-+n-h-m)+jp})$$

Hierbij is *m* het aantal gebruikte maanden, *j* het aantal gebruikte jaren, *h* de lengte van de voorspellings-periode en *p* de periodelengte.

4.4.6.6 Externe variabelen

Bij het voorspellen kan gebruik gemaakt worden van externe variabelen, waarvan verwacht wordt dat ze een verband hebben met de te voorspellen tijdreeks. De verwachting is dat deze variabelen kunnen bijdragen aan de voorspelling.

Van de variabelen wordt op dezelfde wijze als hierboven een aantal opeenvolgende punten en perioden als invoer van het netwerk gebruikt. Dit leidt tot het onderstaande netwerk.

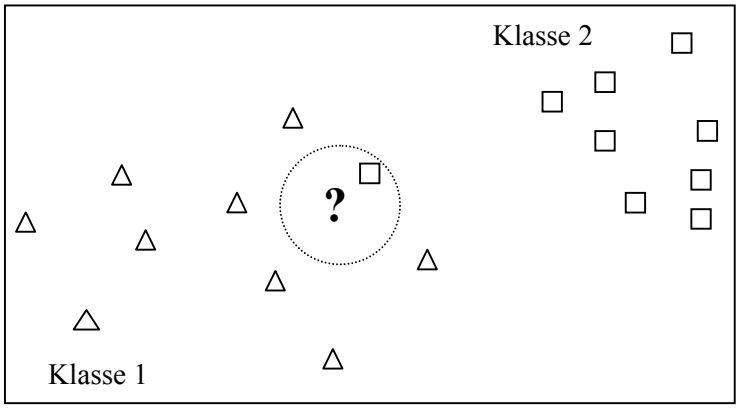


Figuur 4.16: Externe variabelen als extra invoer voor het netwerk.

4.5 Nearest-neighbour-methode

4.5.1 Inleiding

De nearest-neighbour-methode [Rich91] is een eenvoudige classificatietechniek. Aan een ongeclassificeerd object wordt de klasse toegekend van het dichtstbijzijnde object gemeten met een bepaalde afstandsmaat.



Figuur 4.17: Nearest-neighbour-classificatie van object **?**.

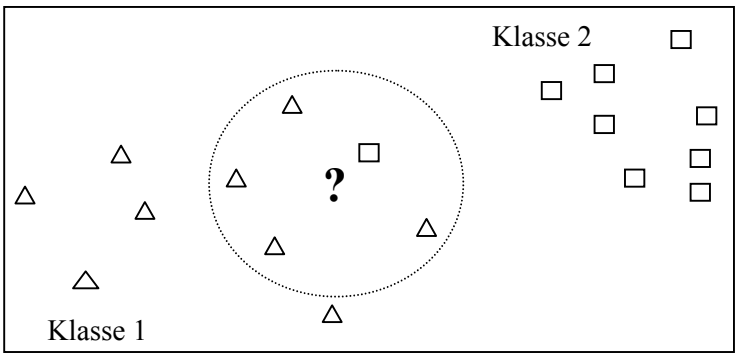
In het bovenstaande figuur wordt aan het ongeclassificeerde object (**?**) de klasse 2 toegekend, omdat zijn naaste buur deel uit maakt van klasse 2.

De methode heeft geen trainingsfase zoals bijvoorbeeld neurale netwerken, maar gebruikt de beschikbare gegevens direct bij de classificatie. Het voordeel van deze methode is dat er geen rekenintensieve trainingsfase uitgevoerd moet worden. Nadeel is dat bij elke classificatie de gehele dataset doorgerekend moet worden om de naaste buur te vinden.

4.5.2 k-nearest-neighbour-methode

In het bovenstaande voorbeeld wordt het ongeclassificeerde object de klasse 2 toegekend omdat zijn naaste buur deel uit maakt van deze klasse. Gezien de spreiding van de twee klassen zou het voor de hand liggen juist klasse 1 aan dit object toe te kennen omdat het zich in de cluster van klasse 1 bevindt.

Door gebruik te maken van de k-nearest-neighbour-methode kan aan het ongeclassificeerde object klasse 1 worden toegekend. Met deze methode wordt de klasse gekozen die in de verzameling k beste buren de grootste vertegenwoordiging heeft. (zie figuur 4.18) In het voorbeeld is k gelijk aan 5. Als maat voor de naaste buur geldt de Euclidische afstand van een object tot het ongeclassificeerde object.



Figuur 4.18: k-Nearest-neighbour-classificatie van object **?** met k=5.

In het bovenstaande voorbeeld worden de objecten vergeleken op basis van één kenmerk; de ligging ten opzichte van elkaar in een tweedimensionale ruimte. In het algemeen geldt voor de afstandsmaat *d* van twee objecten *a, b* met *j* kenmerken p_i:

$$d(a,b) = \sqrt{\sum_{i=1}^{j} \left(p_i^a - p_i^b\right)^2}$$

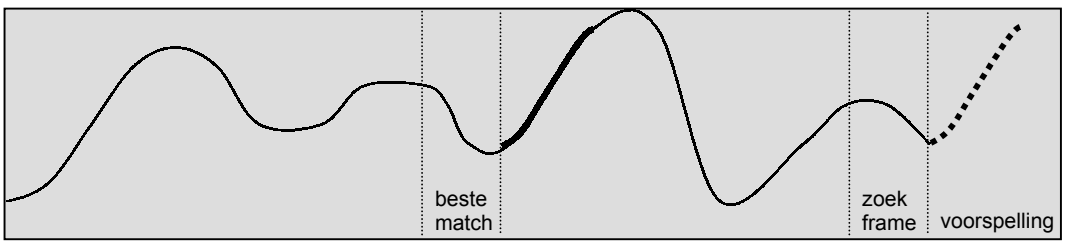
De afstand wordt berekend op basis van *j* kenmerken in kenmerkvector $\boldsymbol{p}$. De waarden van deze kenmerken moeten genormaliseerd worden, omdat ze anders een ongelijk aandeel hebben in de afstandsmaat.

Het is ook mogelijk om aan elk kenmerk een wegingsfactor w_i te koppelen. Op deze manier kan een bepaald kenmerk een groter of juist kleiner effect hebben op de afstandsmaat en dus op de classificatie van een object. Dit leidt tot de volgende afstandsmaat:

$$d(a,b) = \sqrt{\sum_{i=1}^{j} w_i \left(p_i^a - p_i^b\right)^2}$$

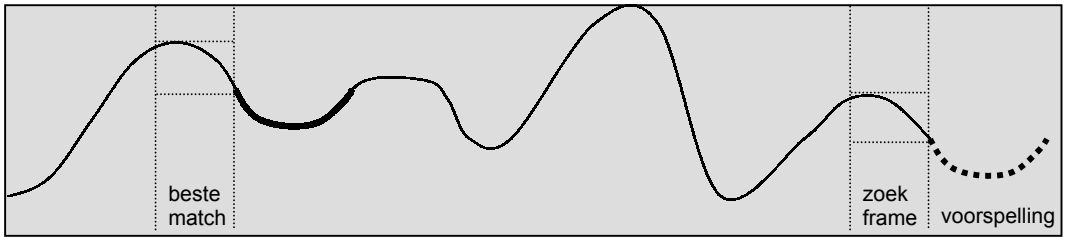
4.5.3 kNN voor tijdreeks voorspellingen

De k-nearest-neighbour-methode (kNN) kan gebruikt worden voor tijdreeks-voorspellingen. In de tijdreeks wordt dan gezocht naar een stuk tijdreeks, vanaf nu een frame [Das98] genoemd dat de grootste overeenkomst vertoont met het frame dat net voor de voorspelling ligt, het zogenaamde zoekframe. Als voorspelling wordt dan het stuk tijdreeks genomen dat na het gevonden frame (beste match) ligt.



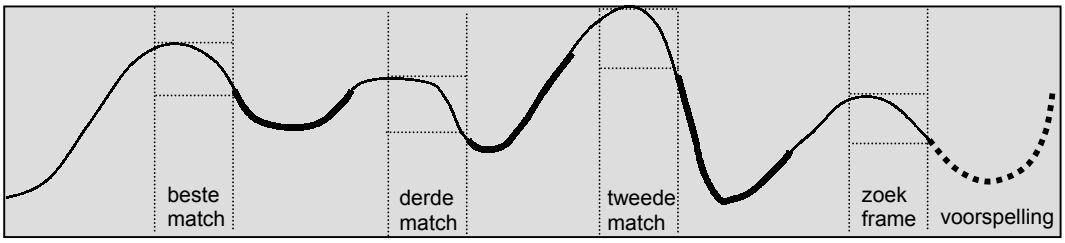
Figuur 4.19: kNN voor tijdreeksvoorspellingen. Zoekframe en beste-match.

In het figuur 4.19 is als kenmerkvector de verzameling punten uit de frames genomen. Het nadeel van deze methode is dat frames slechts overeenkomen als ze op dezelfde hoogte liggen. Als de tijdreeks bijvoorbeeld continu stijgend is, zal nooit een goede match gevonden worden. Een oplossing is om de frames te normaliseren tussen hun minimale en maximale waarde. Op deze manier kan een overeenkomst op vorm gevonden worden in plaats van op ligging. Dit leidt tot het onderstaande figuur.



Figuur 4.20: De frames zijn genormaliseerd zodat de match op basis van vorm is in plaats van ligging.

Als de nearest-neighbour parameter k groter is dan 1, worden meerdere frames gezocht. De voorspellingen van elk van de frames worden gemiddeld naar de matchfout; de afstand tussen een frame en het zoekframe. In het onderstaande figuur is k gelijk aan drie.

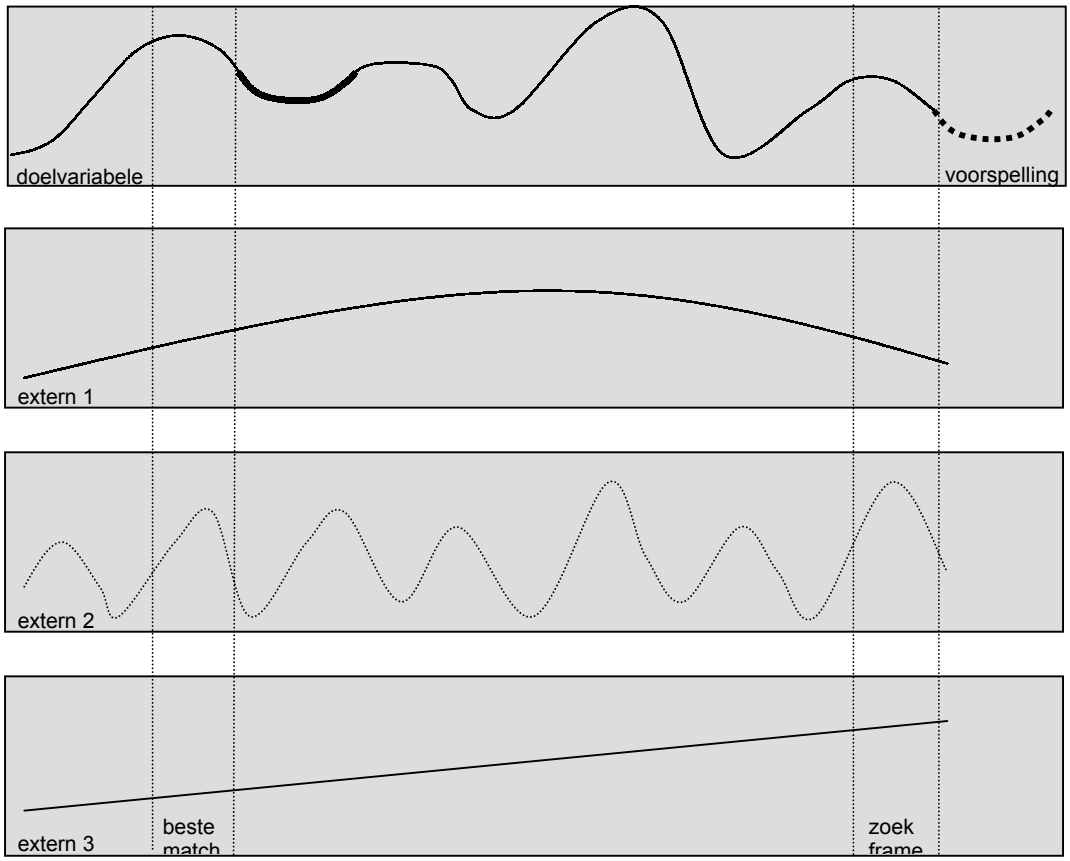


Figuur 4.21: kNN-voorspelling met k=3 en de rangorde van matches.

Als bij de voorspelling tevens externe variabelen gebruikt worden, maken deze variabelen ook deel uit van de kenmerkvector. Er wordt nu gezocht naar de beste match gemeten over meerdere tijdreeksen, waarbij aan elke reeks een bepaald gewicht gekoppeld wordt. (zie figuur 4.22) Op deze manier kan ingesteld worden dat de doelvariabele een grotere invloed heeft op de voorspelling dan de externe variabelen.

De te voorspellen tijdreeks is $\boldsymbol{x}$, de externe variabelen die bij de voorspelling gebruikt worden zijn $\boldsymbol{e}^1$ .. $\boldsymbol{e}^n$ en de breedte van het frame is b. De kenmerkvector $\boldsymbol{p}$ is dan:

$$\mathbf{p} = \begin{pmatrix} x_1 \\ \vdots \\ x_b \\ e^1_1 \\ \vdots \\ e^1_b \\ \vdots \\ e^n_1 \\ \vdots \\ e^n_b \end{pmatrix}$$



Figuur 4.22: kNN-voorspelling met gebruik van externe variabelen.

Bij het voorspellen van tijdreeksen met knn-methoden is een aantal zaken van belang:

- De breedte van het zoekframe *b*
- De lengte van de voorspelling *h*
- De grootte van de nearest-neighbour parameter *k*
- Directe of iteratieve voorspelling
- Externe variabelen
- Weging van de verschillende variabelen

De keuze van de breedte van het zoekframe is belangrijk voor de voorspelling. Wordt deze te klein gekozen (bijvoorbeeld 3 punten), dan wordt een beste match gevonden die niet specifiek is voor het zoekframe. Wanneer de breedte van het zoekframe echter te groot gekozen wordt (bijvoorbeeld de helft van het aantal punten in de tijdreeks), zal ook geen goede matchframe gevonden worden.

De gekozen lengte van de voorspelling moet in overeenstemming zijn met de breedte van het zoekframe. Wanneer voor een kleine zoekframebreedte gekozen wordt, is het niet verstandig een grootte voorspellingslengte te kiezen. (bijvoorbeeld 5 maal de zoekframebreedte), omdat door het kleine zoekframe slechts naar een lokale overeenkomst gezocht is.

De nearest-neighbour parameter k bepaalt het aantal frames dat gezocht wordt en dus van hoeveel frames de voorspelling gemiddeld wordt. Als k=1 wordt slechts één frame gebruikt voor de voorspelling. Het model wordt hierdoor te gevoelig voor fouten. De k moet echter ook niet te groot gekozen worden, omdat de voorspelling dan te algemeen wordt.

De voorspellingsmethode kan direct of iteratief zijn. De directe methode voorspelt in één keer de gehele voorspellingsperiode. De iteratieve methode voorspelt het eerste punt van de voorspellingsperiode, daarna verschuift het zoekframe één punt en bevat het zojuist voorspelde punt. Met dit zoekframe wordt vervolgens het volgende punt voorspelt totdat de gehele voorspellingsperiode is voorspeld. De directe methode is sneller dan de iteratieve, maar deze laatste kan als voordeel hebben dat slecht één punt vooruit voorspeld wordt, waardoor de voorspelling beter wordt.

Het voorspellingsmodel kan gebruik maken van externe variabelen. In dat geval moet het aantal en de soort variabelen gekozen worden. Het heeft alleen zin variabelen te gebruiken die een correlatie hebben met de doelvariabele. Aan elke variabele kan een gewicht gekoppeld worden, dat het aandeel in de voorspelling bepaalt. Op deze manier kan de bijdrage van de doelvariabele aan de voorspelling groter gemaakt worden dan die van de externe variabelen. Tussen de externe variabelen onderling kan tevens onderscheid gemaakt worden.

5. Experimenten

5.1 Inleiding

In dit hoofdstuk worden de experimenten en de resultaten van de verschillende voorspellingstechnieken beschreven. Paragraaf 5.2 beschrijft de algemene experimentele opzet. In paragraaf 5.3 tot en met 5.5 worden de experimenten en resultaten van respectievelijk de trend-fitting-methode, nearest-neighbour-methode en neurale netwerken beschreven. Paragraaf 5.6 geeft een overzicht van de resultaten.

5.2 Experimentele opzet

De technieken worden getest zijn op alle vier de datasets, maar voor de overzichtelijkheid worden de resultaten slechts uitgebreid getoond voor de eerste dataset: het totale advertentievolume van De Telegraaf. In het volgende hoofdstuk met de conclusies wordt een overzicht gegeven van de resultaten van alle datasets.

5.2.1 Doel van de experimenten

In de experimenten worden de trend-fitting-methode, neurale netwerken en de nearest-neighbour-methode getest en vergeleken met de standaard lineaire voorspellingstechnieken ARX en Box-Jenkins.

De voorspellingsmethoden gebruiken elk een aantal parameters die het voorspellingsmodel instellen. Tijdens de experimenten zijn deze geoptimaliseerd door met verschillende parameters een groot aantal modellen te maken en de voorspellingsfout daarvan te berekenen.

Tevens wordt bij de neurale netwerken en de nearest-neighbour-methode de invloed onderzocht van externe variabelen op de voorspelling.

5.2.2 Evaluatiemaat

De kwaliteit van een voorspellingsmodel wordt weergegeven door de voorspellingsfout. In elke methode worden steeds 13 punten voorspeld, waarvan de gemiddelde kwadratische fout (MSE) [Gaynor94] wordt berekend. Omdat de doelvariabele bij alle methoden geschaald is tussen 0 en 1, zijn de MSE's van de verschillende methoden te vergelijken. De MSE wordt op de onderstaande manier berekend, waarbij x_i een punt uit de dataset is, $\hat{x}_i$ de voorspelling en n het aantal datapunten van de voorspelling is.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2$$

Naast de MSE wordt ook de variantie (VAR) berekend om inzicht te verkrijgen in de spreiding binnen de foutverzameling. Hierbij is ε_i de fout en $\bar{\varepsilon}$ de gemiddelde fout.

$$VAR = \frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2$$

5.2.3 Externe variabelen

Bij het voorspellen met neurale netwerken en de nearest-neighbour-methode wordt ook gebruik gemaakt van externe variabelen. Getest wordt of deze variabelen door hun economische samenhang de voorspelling van de advertentievolumes positief kunnen beïnvloeden.

In de resultaten van de experimenten worden de volgende nummers van externe variabelen gebruikt:

1. Binnenlandse consumptieve bestedingen
2. Voedings- en genotmiddelen
3. Duurzame consumptiegoederen
4. Overige goederen en diensten
5. Geregistreerde werklozen
6. CBS-koersindex
7. Spaartegoeden
8. Dollarkoers
9. Goudprijs
10. Daggeldrente
11. Hypotheekrente

5.3 Trend-fitting-methode

5.3.1 Inleiding

Bij het gebruik van de trend-fitting-methode moeten de volgende parameters bepaald worden:

- Experiment 1:Averaging-periode L; dit is het aantal punten waarop de berekening van de moving-average gebaseerd wordt.
- Experiment 2: fitorde n; n-de graad polynoom
- Experiment 3: fitlengte; dit deel van de originele data wordt gebruikt voor de curve-fitting
- Experiment 4: quotiënt- of verschilmethode bij de trend-verwijdering.
- Experiment 5: jaargemiddelde; lineair of gewogen

5.3.2 Experimenten

De verschillende parameters zijn getest door middel van een groot aantal voorspellingen. Voor elke parameter wordt een interval bepaald waarbinnen deze getest wordt. Vervolgens worden alle mogelijke combinaties van parameters binnen deze intervallen getest. Hierna kan een optimale combinatie bepaald worden.

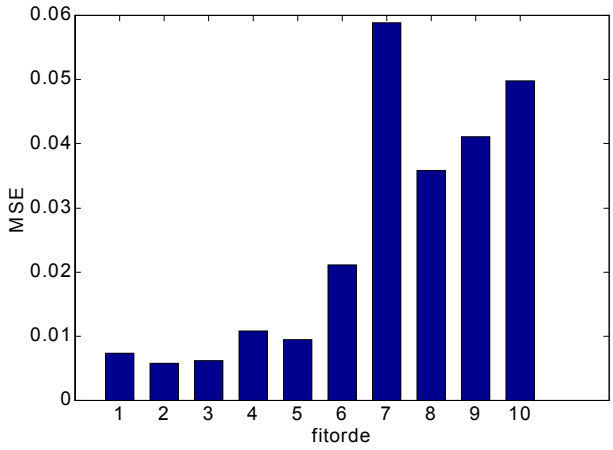
5.3.2.1 Experiment 1: Averaging-period

Gekozen is voor een averaging-period van 13. Deze is gelijk aan de periodelengte van de advertentievolume-dataset die bestaat uit 13 vierwekelijkse perioden. De moving-average wordt hierdoor steeds bepaald door de voorgaande en volgende 6 perioden.

Zou de averaging-period te klein worden gekozen, dan blijven de seizoensvariaties aanwezig. Wordt de averaging-period te groot genomen, dan wordt de moving-average te vlak waardoor hij de originele tijdreeks niet goed volgt, om deze te kunnen gebruiken bij de voorspelling.

5.3.2.2 Experiment 2: Fitorde

De fitorde *n* bepaalt de orde van het polynoom dat gebruikt wordt om de moving-average te fitten. Gekozen is voor een test-interval voor *n* van [1:10]

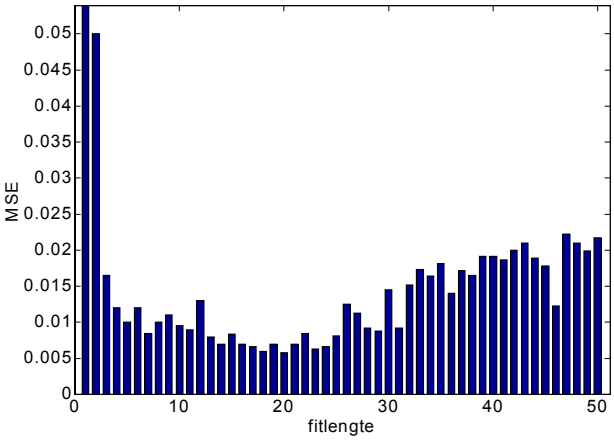


Figuur 5.1: MSE bij verschillende waarden van de fitorde.

Een fitorde van 2 geeft de beste resultaten. Dit betekent dat een tweedegraads functie gebruikt wordt bij de polynomenfit.

5.3.2.3 Experiment 3: Fitlengte

De fitlengte *n* bepaalt het deel van de moving-average waarop de fit gebaseerd wordt. Gekozen is voor een test-interval voor de fitlengte van [1:50]



Figuur 5.2: MSE bij verschillende waarden van de fitlengte.

Een fitlengte van 20 geeft de beste resultaten. Dit betekent dat de 20 laatste punten van de moving-average gebruikt worden voor de trendvoorspelling.

5.3.2.4 Experiment 4: Quotiënt- of verschilmethode

Om de trend te verwijderen wordt de quotiënt- of de verschilmethode gebruikt. Een reeks voorspellingen is uitgevoerd met verschillende instellingen voor de overige parameters, waarvan de berekende MSE gemiddeld wordt. In onderstaande tabel wordt een overzicht gegeven van deze gemiddelde MSE voor beide methoden.

Methode	MSE
Quotiënt-methode	0.015
Verschil-methode	0.017

Tabel 5.1: Gemiddelde MSE voor twee trendverwijderings-methoden.

Het verschil tussen beide methoden is klein. Gekozen is voor de quotiënt-methode omdat deze de laagste gemiddelde MSE heeft.

5.3.2.5 Experiment 5: Jaargemiddelde

Het jaargemiddelde kan lineair zijn of gewogen worden. Het gemiddelde is lineair als alle jaren even zwaar wegen in het gemiddelde. Bij een gewogen gemiddelde wegen de laatste jaren zwaarder mee.

Een reeks voorspellingen wordt wederom uitgevoerd met verschillende instellingen voor de overige parameters, waarvan de berekende MSE gemiddeld wordt. In onderstaande tabel wordt een overzicht gegeven van deze gemiddelde MSE voor beide methoden.

Methode	MSE
Lineair gemiddelde	0.021
Gewogen gemiddelde	0.009

Tabel 5.2: Gemiddelde MSE voor lineair- en gewogen gemiddelde.

Het gewogen gemiddelde geeft de laagste MSE. Met deze methode worden de laatste jaren zwaarder gewogen. De voorspelling heeft blijkbaar de meeste overeenkomsten met de laatste jaren.

5.3.3 Resultaten

De kleinste voorspellingsfout wordt verkregen met de volgende parameters:

- Averaging-period L=13
- fitorde n=2
- fitlengte = 20
- trendverwijdering: quotiënt-methode
- jaargemiddelde: gewogen

De voorspellingsfout van deze methode is:

- MSE = 0,0058
- VAR = 0,0062

5.4 Neurale netwerken

5.4.1 Inleiding

Bij het gebruik van neurale netweken als tijdreeksvoorspeller moeten de volgende instellingen bepaald worden:

Netwerk-instellingen:

- Welke trainingsfunctie
- Netwerktopologie; aantal tussenlagen
- Experiment 1: netwerktopologie; aantal neuronen in de tussenlagen

Model-parameters:

- Experiment 2: voorspelling per punt of periode
- Experiment 3: trainingsmethode direct of iteratief
- Experiment 4: welke externe variabelen
- Experiment 5: welke tijdreekspunten als invoer

Een aantal van deze instellingen en parameters wordt gekozen en de rest wordt bepaald in de experimenten.

5.4.2 Gekozen netwerk-instellingen/parameters

De Levenberg-Marquardt trainingsfunctie is gekozen omdat deze methode zeer snel is en toch goede resultaten geeft [Demuth97].

Gekozen is voor één verborgen laag, omdat volgens Demuth en Beale [Demuth97] meer dan één verborgen laag niet noodzakelijk is.

5.4.3 Experimenten

De verschillende parameters zijn getest door middel van een groot aantal voorspellingen.
Voor elke parameter wordt een interval bepaald waarbinnen deze getest wordt. Vervolgens worden alle mogelijke combinaties van parameters binnen deze intervallen getest. Hierna kan een optimale combinatie bepaald worden.

5.4.3.1 Experiment 1: Aantal neuronen in tussenlaag

Een basis netwerk van 13 invoerneuronen en één uitvoer neuron wordt gebruikt om het optimale aantal verborgen neuronen te bepalen. Het netwerk wordt getraind om met de voorgaande 13 perioden één jaar te voorspellen. Bij 10 verborgen neuronen presteert het netwerk goed. Bij een groter aantal neemt de kans op overfitting te veel toe. Gekozen is voor een test-interval voor het aantal neuronen in de tussenlaag van [1:20]

5.4.3.2 Experiment 2: Voorspelling per punt of periode

Met dit basis netwerk, nu met 10 verborgen neuronen wordt getest met een voorspelling per punt en per periode. De voorspelling per punt geeft de kleinste voorspellingsfout en deze methode wordt gekozen.

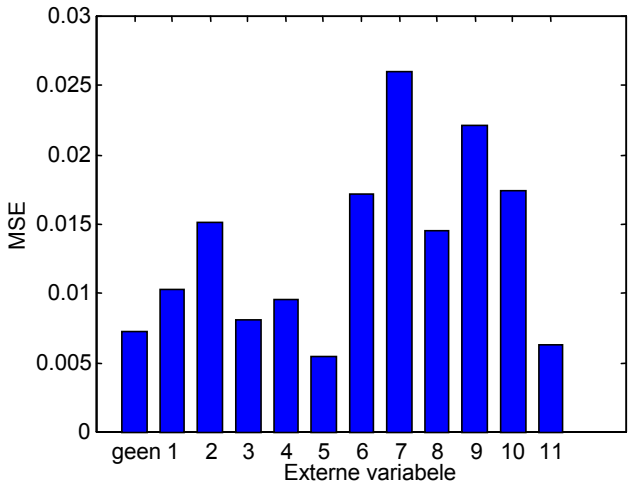
5.4.3.3 Experiment 3: Trainingsmethode; direct of iteratief

Vervolgens wordt gekozen voor een directe of iteratieve trainingsmethode. Beide methoden scoren vergelijkbaar. Gekozen wordt voor de directe methode, omdat bij de iteratieve methode voorgaande voorspellingen gebruikt worden waarvan de fouten cumuleren.

5.4.3.4 Experiment 4: Externe variabelen

In dit experiment wordt getest welke externe variabelen de beste resultaten geven. Een netwerk wordt gebruikt met 10 verborgen neuronen en één uitvoerneuron. Het netwerk wordt getraind om steeds één punt één jaar vooruit te voorspellen.

De invoer bestaat uit 13 punten (één jaar) aan historische data en 13 punten van één externe variabele. Voor de controle is ook getest met een netwerk zonder externe variabelen.
(zie figuur 5.3) De nummers in onderstaande grafiek corresponderen met de nummers van de externe variabelen zoals beschreven in paragraaf 5.2.3.

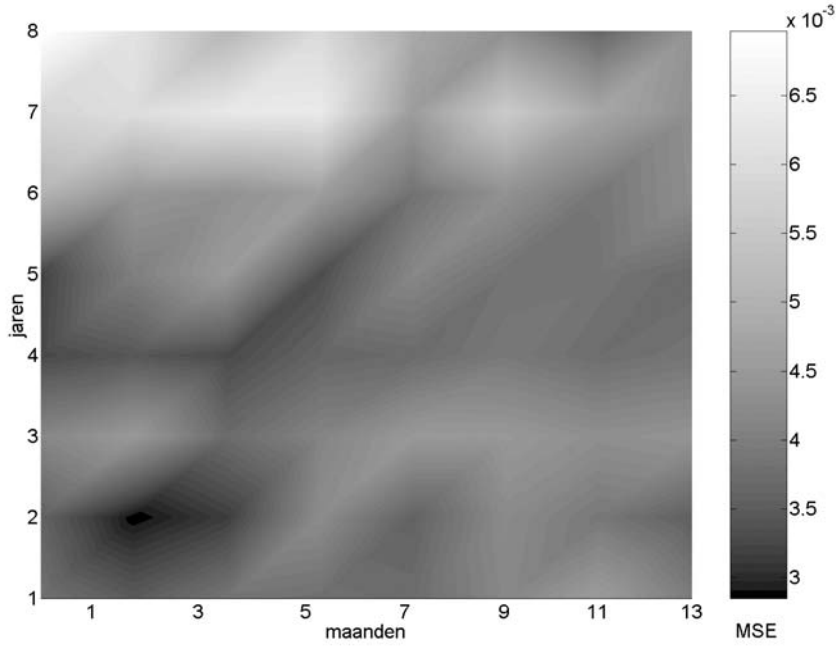


Figuur 5.3: MSE bij verschillende externe variabelen.

Het netwerk heeft de kleinste fout bij het gebruik van externe variabelen 5 en 11. (geregistreerde werklozen en hypotheekrente) Bij het gebruik van de andere externe variabelen wordt de fout groter dan wanneer geen externe variabelen gebruikt worden.

5.4.3.5 Experiment 5: Keuze van de tijdreekspunten als invoer

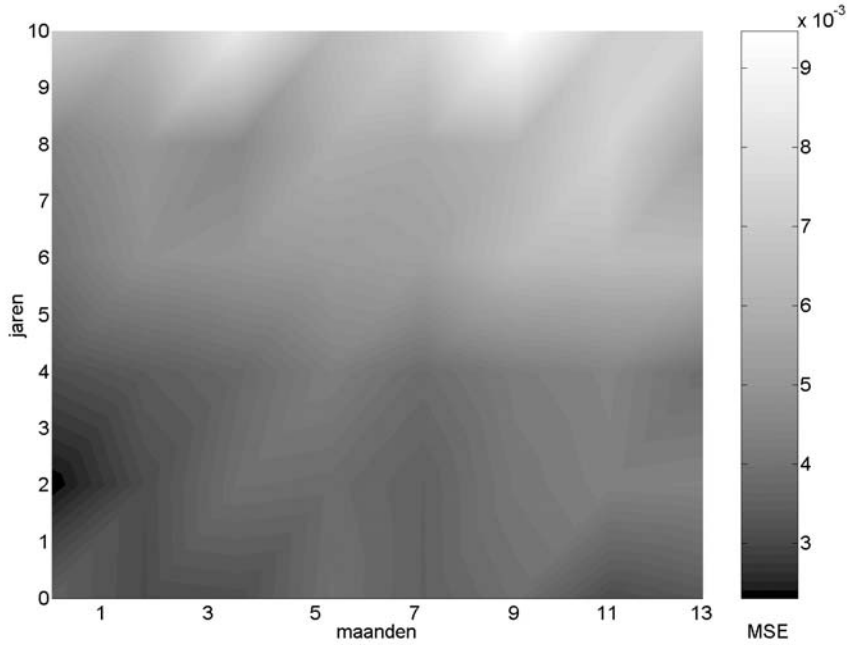
Vervolgens wordt gebruikmakend van externe variabelen 5 en 11 getest welke maanden en jaren van de doelvariabele en externe variabelen gebruikt moeten worden. De resultaten zijn staan in figuur 5.4, waarin de voorspellingsfout is uitgezet tegen het aantal maanden en jaren. Voor het aantal maanden en jaren is respectievelijk gekozen voor een test-interval van [1:13] en [1:10]



Figuur 5.4: MSE bij verschillende maanden en jaren van de doelvariabele.

Het optimale aantal maanden en jaren van de doelvariabele dat gebruikt wordt is 2.

In dit experiment is ook het optimale aantal jaren en maanden van de externe variabelen getest. Deze zijn in onderstaand figuur te zien:



Figuur 5.5: MSE bij verschillende maanden en jaren van de externe variabelen.

Het optimale aantal jaren van de externe variabelen is 2 en het optimale aantal maanden is 0.

5.4.4 Resultaten

Het optimale neurale netwerk heeft de volgende parameters:

- 1 verborgen laag met 10 neuronen
- 2 externe variabelen (5 en 11)
- 2 jaren en 2 maanden van de doelvariabelen
- 2 jaren en 0 maanden van elk van de externe variabelen

De voorspellingsfout van dit netwerk is:

- MSE = 0,0024
- VAR = 0,0004

5.5 Nearest-neighbour-methode

5.5.1 Inleiding

Bij het gebruik van de nearest-neighbour-methode als tijdreeksvoorspeller moeten de volgende instellingen bepaald worden:

- Experiment 1: Directe of iteratieve voorspelling
- Experiment 2: Keuze van de externe variabelen
- Experiment 3: Nearest-neighbour parameter *k*
- Experiment 4: De breedte van het zoekframe
- Experiment 5: Gewichten van de externe variabelen

5.5.2 Experimenten

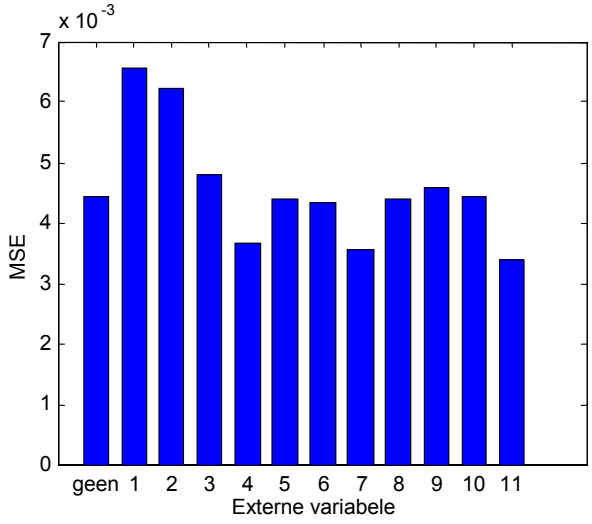
De verschillende parameters zijn getest door middel van een groot aantal voorspellingen. Voor elke parameter wordt een interval bepaald waarbinnen deze getest wordt. Vervolgens worden alle mogelijke combinaties van parameters binnen deze intervallen getest. Hierna kan een optimale combinatie bepaald worden.

5.5.2.1 Experiment 1: Directe of iteratieve voorspelling

De resultaten van de directe en iteratieve voorspellingsmethode vergelijkbaar. Er is gekozen voor de directe methode omdat deze sneller is.

5.5.2.2 Experiment 2: Keuze van de externe variabelen

In dit experiment wordt per externe variabele de invloed daarvan op de voorspelling getest. Voor de controle is ook getest met een model dat geen externe variabele gebruikt. De nummers in onderstaande grafiek corresponderen met de nummers van de externe variabelen zoals beschreven in paragraaf 5.2.3.

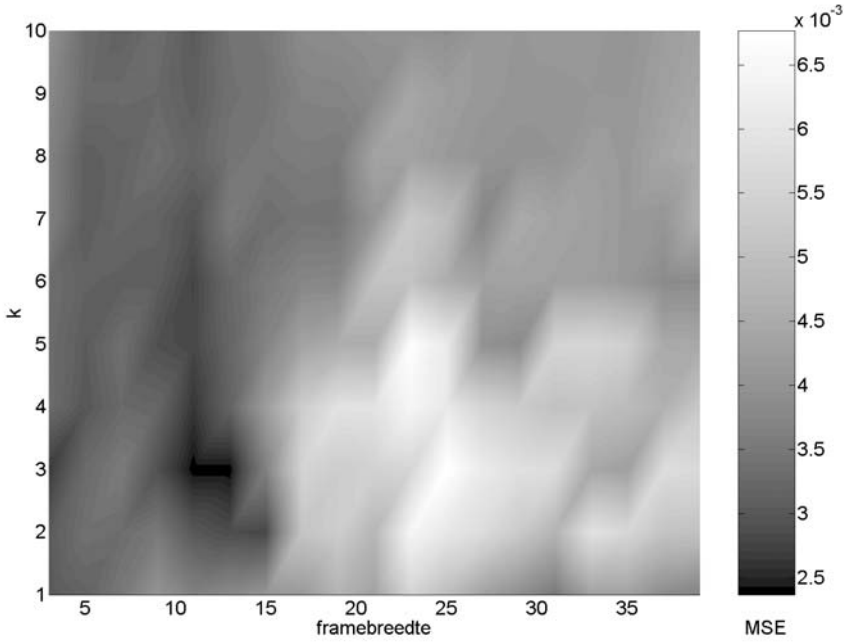


Figuur 5.6: MSE bij verschillende externe variabelen.

Het model heeft de kleinste voorspellingsfout bij het gebruik van externe variabelen 7 en 11. (Dollarkoers en hypotheekrente) Deze externe variabelen worden in de volgende experimenten gebruikt.

5.5.2.3 Experiment 3 & 4: Optimale k en zoekframebreedte

In dit experiment worden de optimale *k* en zoekframebreedte gezocht. Deze worden in figuur 5.7 uitgezet tegen de voorspellingsfout. Voor de nearest-neighbour-parameter *k* en de zoekframebreedte is respectievelijk gekozen voor een test-interval van [1:10] en [1:40]

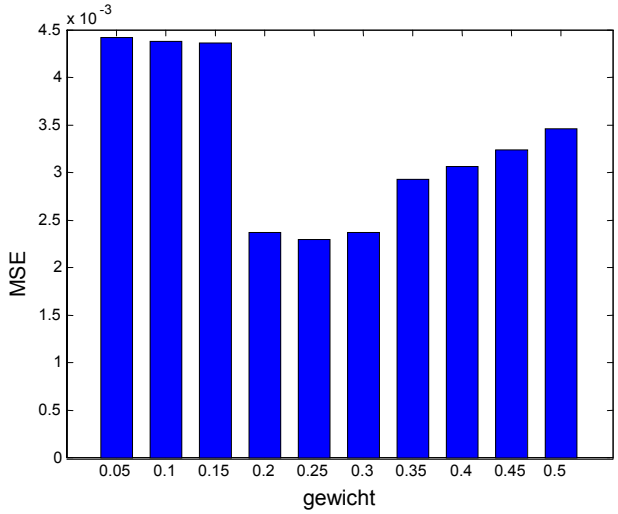


Figuur 5.7: MSE bij verschillende waarden voor k en framebreedte.

De optimale *k* is 3. De optimale framebreedte is 13 punten en dat is precies één jaar.

5.5.2.4 Experiment 5: Gewichten van de externe variabelen

Tenslotte worden de optimale gewichten van de externe variabelen bepaald.
Hiervoor wordt een testinterval gebruikt van [0:0.5] in stappen van 0.05.



Figuur 5.8: MSE bij verschillende gewichten van de externe variabelen.

Het optimale gewicht van de externe variabelen is 0.25 bij het gebruik van twee externe variabelen. Dit wil zeggen dat de externe variabelen beide een gewicht van 0.25 hebben en dat de doelvariabele een gewicht heeft van (1 – 2*0.25) = 0.5 .

5.5.3 Resultaten

De optimale kNN-methode heeft de volgende parameters:

- 2 externe variabelen (7 en 11)
- *k*=3
- framebreedte = 13
- gewicht externe variabelen = 0.25

De voorspellingfout van dit model is:

- MSE = 0,0023
- VAR = 0,0024

5.6 Overzicht van de resultaten voor dataset-1

De onderstaande tabel geeft een overzicht van de geteste voorspellingsmethoden voor dataset-1: het totale advertentie volume.

Methode	MSE	VAR	Externe variabelen
Trend-fitting	0,0058	0,0062	nvt
Neurale netwerken	0,0024	0,0004	5,11
kNN	0,0023	0,0024	7,11

Tabel 5.3: Overzicht resultaten van de geteste voorspellingsmethoden.

De neurale netwerken en nearest-neighbour-methode presteren vergelijkbaar en de Trend-fitting-methode blijft duidelijk achter.

De externe variabele 11 ‘hypotheekrente’ draagt zowel bij de neurale netwerken als nearest-neighbour-methode bij aan een betere voorspelling.

Het optimale neurale netwerk gebruikt naast deze externe variabele ook de variabele ‘geregistreerde werklozen’.

De voorspelling van de nearest-neighbour-methode wordt naast de ‘hypotheekrente’ positief beïnvloed door de ‘Dollarkoers’.

6. Algemene resultaten en conclusies

6.1 Inleiding

Dit hoofdstuk bevat een overzicht van de resultaten en de conclusies. Paragraaf 6.2 geeft de resultaten van de experimenten op alle datasets. In paragraaf 6.3 en 6.4 volgen de conclusies over respectievelijk de voorspellingsmethoden en de externe variabelen. Paragraaf 6.5 bevat het onderzoeksresultaat en in paragraaf 6.6 worden aanbevelingen voor verder onderzoek gedaan.

6.2 Overzicht resultaten

De resultaten van de trend-fitting-methode, neurale netwerken en nearest-neighbour-methode worden vergeleken met standaardmethoden voor het voorspellen van tijdreeksen, ARX en Box-Jenkins.

In deze paragraaf 6.2.1 worden de resultaten van de experimenten getoond zoals beschreven in hoofdstuk 5. De gebruikte externe variabelen kunnen per methode verschillen. Om de methoden goed te kunnen vergelijken zijn deze ook getest zonder gebruik te maken van externe variabelen. Deze resultaten worden gepresenteerd in paragraaf 6.2.2.

6.2.1 Resultaten met gebruik van externe variabelen

In het onderstaande overzicht zijn de resultaten verzameld van de in hoofdstuk 5 beschreven experimenten en deze worden vergeleken met de gebruikte standaardmethoden ARX en Box-Jenkins.

Bij deze methoden is gebruik gemaakt van externe variabelen. Per dataset is bepaald welke combinatie van twee variabelen de beste resultaten gaf. Deze variabelen zijn vermeld in tabel 6.2.

dataset:	Totaal		Nat/lok		Personeel		Rubriek	
methode:	MSE	VAR	MSE	VAR	MSE	VAR	MSE	VAR
Trend-fitting	0,0058	0,0062	0,0149	0,0091	0,0065	0,0067	0,0012	0,0007
Neurale netwerken	0,0024	0,0004	0,0059	0,0032	0,0039	0,0018	0,0021	0,0031
kNN	0,0023	0,0024	0,0075	0,0061	0,0062	0,0026	0,0015	0,0013
ARX	0,0027	0,0024	0,0085	0,0042	0,0091	0,0054	0,0025	0,0024
Box-Jenkins	0,0023	0,0018	0,0069	0,0044	0,0064	0,0056	0,0025	0,0015

Tabel 6.1: Voorspellingsfout van alle methoden op alle datasets

In het onderstaande overzicht worden de gebruikte externe variabelen getoond. De trend-fitting-methode maakt geen gebruik van externe variabelen.

methode/dataset	Totaal	Nationaal/lokaal	Personeel	Rubriek
Trend-fitting	nvt	nvt	nvt	nvt
Neurale netwerken	5,11	3,9	5,9	7,10
kNN	7,11	3,9	3,11	7,11
ARX	8,10	3,5	5,9	8,10
Box-Jenkins	8,10	3,5	5,9	8,10

Tabel 6.2: Gebruikte externe variabelen bij alle methoden en alle datasets

6.2.2 Resultaten zonder gebruik van externe variabelen

In tabel 6.3 worden de resultaten getoond van de experimenten waarbij geen gebruik gemaakt is van de externe variabelen. Daar de trend-fitting-methode geen gebruik maakt van externe variabelen zijn de resultaten van deze methode gelijk aan die van tabel 6.1.

dataset:	Totaal		Nat/lok		Personeel		Rubriek	
methode:	MSE	VAR	MSE	VAR	MSE	VAR	MSE	VAR
Trend-fitting	0,0058	0,0062	0,0149	0,0091	0,0065	0,0067	0,0012	0,0007
Neurale netwerken	0,0034	0,0042	0,0108	0,0088	0,0059	0,0052	0,0043	0,0031
kNN	0,0029	0,0019	0,0103	0,0073	0,0056	0,0062	0,0034	0,0037
ARX	0,0027	0,0024	0,0109	0,0080	0,0209	0,0084	0,0040	0,0037
Box-Jenkins	0,0024	0,0012	0,0083	0,0088	0,0150	0,0070	0,0025	0,0015

Tabel 6.3: Voorspellingsfout van alle methoden op alle datasets
zonder gebruik van externe variabelen.

6.3 Conclusies voorspellingsmethoden

6.3.1 Algemeen

Gezien de beperkte omvang van de gebruikte datasets is het niet verstandig harde conclusies te trekken omtrent gebruikte methoden. Toch kunnen de volgende zaken opgemerkt worden:

De resultaten van de nearest-neighbour-methode zijn vergelijkbaar met die van de standaard tijdreeks-voorspellingsmethoden Box-Jenkins en ARX. Een belangrijke doelstelling van dit onderzoek is daarmee vervuld. Voor mijn stagebedrijf is het belangrijk uitsluitsel te krijgen over de toepasbaarheid van nearest-neighbour-methoden voor het voorspellen van tijdreeksen. Gezien de resultaten kan opgemerkt worden dat deze methoden zeker geschikt zijn voor die taak.

De resultaten van de neurale netwerken komen voor de meeste datasets overeen met die van de lineaire methoden en de nearest-neighbour-methode.

De neurale netwerken hebben echter betere resultaten bij het voorspellen van complexe datasets zoals de personeels-dataset.

De resultaten van de trend-fitting-methode zijn slechter dan die van de andere onderzochte methoden. Een uitzondering vormt de rubrieks-dataset, die een heel constant jaargedrag heeft. De trend-fitting-methode verondersteld de aanwezigheid van een constant jaargedrag en het was dus te verwachten dat de methode op deze dataset goed zou presteren.

De volgende opmerkingen gemaakt worden over onderzochte methoden:

6.3.2 Lineaire methoden (Box-Jenkins/ARX)

De Box-Jenkins methode blijkt een snelle en robuuste voorspellingsmethode te zijn. Een nadeel kan zijn dat de invoerdata beperkingen heeft, de datapunten moeten namelijk opeenvolgend zijn. Er kan niet zoals bij de neurale of kNN-methode gebruikt gemaakt worden van een aantal maanden en een aantal jaren als invoer.

6.3.3 Trend-fitting-methode

De intuïtieve trend-fitting-methode presteert goed bij het voorspellen van datasets met een constant jaargedrag zoals de rubrieksadvertenties, maar blijft duidelijk achter bij het voorspellen van complexe datasets.

6.3.4 Neurale netwerken

Neurale netwerken zijn een goede keuze bij complexe datasets waar een niet-lineair verband verondersteld wordt. Een voordeel is ook dat er geen beperkingen zijn ten aanzien van de invoerdata. Nadelen kunnen zijn de ondoorzichtige werking en omvangrijke trainingstijd. Tevens moet voldoende data beschikbaar zijn om het model te trainen.

6.3.5 Nearest-neighbour-methode

De nearest-neighbour-methode heeft als voordeel een transparante, snelle en robuuste werking en er zijn geen beperkingen aan de invoerdata. Een probleem kan optreden bij het voorspellen van kleine complexe datasets, omdat er dan geen goede match gevonden kan worden in de historische gegevens.

6.4 Conclusies externe variabelen

6.4.1 Algemeen

Het gebruik van externe variabelen blijkt een positieve invloed te hebben op de voorspellingsresultaten. De resultaten van de experimenten met externe variabelen zijn in bijna alle gevallen beter of gelijk aan de resultaten van de experimenten zonder externe variabelen.

De ARX en Box-Jenkins methoden hebben in het geval van een complexe dataset duidelijk voordeel bij het gebruik van externe variabelen.

De neurale netwerken hebben voor alle datasets ongeveer een twee keer zo kleine fout bij het gebruik van externe variabelen.

De beste externe variabelen hebben een globaal verloop en lijken bij te dragen aan de voorspelling van de globale trend. Zo doen de algemene conjunctuur-variabelen als ‘geregistreerde werklozen’, ‘hypotheekrente’ en ‘Dollarkoers’ het goed.

6.4.2 Het totale advertentie volume

De voorspelling van dataset 1 met het totale advertentie volume wordt verbeterd door het gebruik van de algemene conjunctuur-variabelen zoals ‘hypotheekrente’ en ‘goudprijs’.

6.4.3 Nationaal en Lokaal

Dataset 2 met de nationale en lokale advertentievolumes heeft een duidelijk verband met de externe variabele ‘overige goederen en diensten’.

6.4.4 Personeel

De personeelsadvertentievolumes in dataset 3 zijn gerelateerd aan de ‘geregistreerde werklozen’ en dat ligt voor de hand. Bij een lage werkloosheid is het volume aan personeelsadvertenties waarschijnlijk hoog in verband met een grote vraag op de arbeidsmarkt.

6.4.5 Rubrieksadvertenties

De vierde dataset met rubrieksadvertenties heeft het meest constante verloop van de onderzochte datasets. Het combineert een globale stijgende trend met een redelijk constant jaargedrag. De dataset is goed te voorspellen zonder gebruik te maken van externe variabelen, bijvoorbeeld met de trend-fitting-methode. De andere methoden hebben op Box-Jenkins na, voordeel bij het gebruik van de algemene conjunctuur-variabelen zoals ‘goudprijs’ en ‘Dollarkoers’.

6.5 Onderzoeksresultaat

Het resultaat van het onderzoek is dat Sentient Machine Research besloten heeft een applicatie te ontwikkelen voor het voorspellen van tijdreeksen. Gekozen is voor de nearest-neighbour-methode in verband met de transparante, snelle en robuuste werking.

Na mijn stage heb ik een demo-programma ontwikkeld die de mogelijkheden van deze methode in een interactieve applicatie laat zien.

6.6 Verder onderzoek

Verder onderzoek zou gedaan kunnen worden naar de volgende zaken:

- Resultaten op andere en omvangrijkere datasets.
  De gebruikte datasets in dit onderzoek zijn niet groot. Door het ontbreken van voldoende data, is het mogelijk dat de modellen niet optimaal zijn.

- Automatische gewichtenafstelling in de kNN-methode.

- Automatische methode-selectie bij een gegeven tijdreeks:
  Een systeem dat de te voorspellen tijdreeks analyseert en de optimale voorspellings-methode geeft.

- Automatische parameter-optimalisatie bij een gegeven methode:
  Een systeem dat de optimale parameters van de gegeven voorspellingsmethode bepaalt gebruikmakend van de te voorspellen tijdreeks en eventuele kennis over de voor-spellingsmethode.

7. Bibliografie

[Das98] *Rule Discovery from Time Series*, G. Das, K. Lin, H. Mannila, G. Renganathan, P. Smyth, Knowledge Discovery and Data Mining (KDD-98)

[Demuth97] *Neural Network Toolbox User's Guide Version 3.0*, H.B. Demuth, M.H. Beale, Mathworks, 1997

[Gaynor94] *Introduction to Time-series Modeling and Forcasting in Business and Economics*, P.E. Gaynor, R.C. Kirkpatrick, McGraw-Hill, Inc., 1994

[Gunst95] *Statistische Data Analyse*, M.C.M. de Gunst, A.W. van der Vaart, Vrije Universiteit van Amsterdam, 1995

[Gunst97] *Statistical Models*, M.C.M. de Gunst, Vrije Universiteit van Amsterdam, 1997

[Hagan94] *Training Feedforward Networks with the Marquardt algorithm*, M.T. Hagan, M. Menhaj, IEEE Transactions on Neural Networks vol.5 no. 6 p989-993,1994

[Hagan96] *Neural Network Design*, M.T. Hagan, H.B. Demuth, M.H. Beale, PWS Publishing, 1996

[Kröse93] *An Introduction to Neural Networks*, B.J.A. Kröse, P.P. van de Smagt, Universiteit van Amsterdam, 1993

[Parks87] *Digital Filter Design*, Parks, T.W., and C.S. Burrus, John Wiley & Sons New York, 1987. Pgs. 54-83.

[Ran97] Mathematische Systeem Theorie, A.C.M. Ran, Vrije Universiteit van Amsterdam, 1997

[Rich91] *Artificial Intelligence*, E. Rich, K. Knight, Mc.Graw-Hill, 1991

[Tang91] *Feed-forward Neural Nets as Models for Time Series Forecasting*, Z. Tang, P.A. Fishwick, University of Florida, 1991