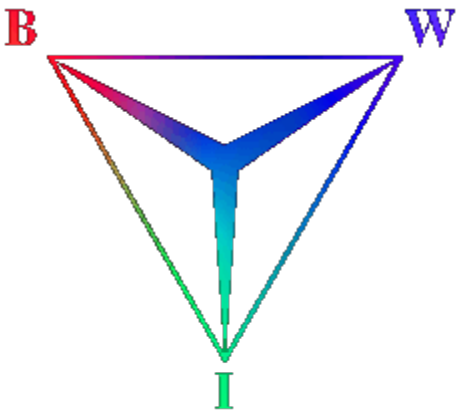
Het Modelleren van de Olieprijzen en de Samenhangende Factoren

Oras Ibrahim

BWI Werkstuk

Januari 2006



Vrije Universiteit
Faculteit der Exacte Wetenschappen
Bedrijfswiskunde en Informatica
De Boelelaan 1081
1081 HV Amsterdam
Nederland

Voorwoord

Dit werkstuk wordt het afstudeervak voor de opleiding Bedrijfswiskunde en Informatica (BWI). Het verrichte onderzoek maakt gebruik van de combinatie van de drie BWI-vakgebieden, waarbij de economische aspecten in de olieprijzenwereld en het mogelijke gebruik van de wiskundige methoden in het modelleren van de economische factoren met behulp van de informatica worden bestudeerd.

Het werkstukonderwerp bespreekt de factoren die het gedrag van de olieprijs beïnvloeden en hoe de afhankelijkheid in een multivariabele olieprijzenwereld in de tijd te kunnen modelleren.

De keuze voor het onderwerp is te wijten aan het feit dat ik mijn stage bij Shell heb gelopen, en zo de wereld van de olieprijs van dichtbij heb leren kennen. De uitdaging is te achterhalen wat nou de echte invloedfactoren op de olieprijs zouden kunnen zijn en of er een nauwkeurig model daarvoor is op te stellen. Bovendien voor iemand van een Iraakse afkomst, speelt de oliewereld een belangrijke rol in het economische en politieke hedendaagse leven en de toekomst van Irak en vooral wereldwijd.

Veel dank aan de behulpzame begeleider Geurt Jongbloed. Dankzij zijn begeleiding is dit werkstuk tot stand gekomen. Rest mij nog te hopen dat dit werk net zo interessant is voor u als lezer...

Samenvatting

Grote oliebedrijven zoals Shell hebben te maken met grote projecten. De besluitvorming om projecten aan te nemen hangt van de opbrengst van het project af. Daarbij kunnen onzekerheden een belangrijke rol spelen. Dit vooral wanneer het om de ontdekking en winning van olie in de olievelden gaat. Zal de olieprijs stijgen? Zo ja, dan is de opbrengst voor het oliewinnende bedrijf hoger. Is er genoeg olie in de aardbodem? Hoe lager de reserves, des te minder olie geproduceerd kan worden. Zullen de kosten zich gedragen zoals van tevoren is berekend? Hogere winningkosten zullen in de olieprijs worden verrekend. Kunnen de olieraffinaderijen de olieproductie aan? Als er meer olie kan worden verwerkt, is de opbrengst hoger, etc…
De olieprijs blijft wel de belangrijkste factor die een rol speelt bij het besluitvormingsproces.

Voor ons als eindgebruikers is het eveneens zeer van belang hoe de olieprijzen zich gedragen. Bijvoorbeeld hogere olieprijzen zijn minder aantrekkelijk en de vraag naar olie of de geraffineerde olieproducten zal dalen en vice versa. Zo oefenen wij ook invloed op de olieprijzen uit, maar in mindere mate.

Veel factoren hangen samen met de olieprijzen. In dit werkstuk zal ik focussen op de gemiddelde jaarlijkse olieprijzen en de andere factoren die ermee samenhangen. Het geheel zal ik modelleren met behulp van multivariate data analyse. De multivariate data kan worden samengevat in een univariate tijdreeks die veel samenhangende eigenschapen van de originele afhankelijkheden bevat. Door de univariate tijdreeks te analyseren, kan er een model worden gemaakt zodat toekomstige olieprijzen en samenhangende factoren voorspeld kunnen worden.

Alle factoren die ik in dit onderzoek gebruik, zullen als variabelen worden gedefinieerd. Deze variabelen hebben jaarlijkse metingen/waarden. Het gedrag van de onderlinge afhankelijkheid zal worden bestudeerd om het geheel van de economische aspecten in model te brengen en zo meer iets te kunnen zeggen over de toekomst.

Het statistische programma R zal worden gebruikt tijdens dit onderzoek. De variabelen worden gedefinieerd als vectoren om vervolgens met behulp van R de toepassing van statistische analyses zoals Principal Component Analysis en Time Series Analysis mogelijk te maken. Het onderzoek met behulp van het programma R zal in grote lijnen de volgende stappen volgen:

1. De economische factoren die met de olieprijs samenhangen bestuderen.
2. De meest belangrijke factoren selecteren en invoeren in het programma R als vectoren.
3. De variabelen analyseren met Principal Component Analysis en Time Series Analysis.
4. De toekomstige prijzen en metingen voorspellen.

Met dit werkstuk wordt er een poging gewaagd om de samenhang tussen de olieprijzen en de oliemarkt te analyseren om op basis van de onderlinge afhankelijkheid voorspellingen te doen.

Inhoudsopgave

Hoofdstuk 1 Introductie

In dit hoofdstuk zal er in het kort worden weergegeven hoe het onderzoek wordt gedaan. Vervolgens zal de structuur van het werkstuk aan bod komen.

1.1 Onderzoek

Er zijn verschillende onderzoeken gedaan naar het gedrag van olieprijzen. Er zijn ook vele methoden en analyses gebruikt om de olieprijzen te voorspellen. Het blijft echter altijd de vraag of de olieprijzen zich gaan gedragen zoals verwacht en dus hoe nauwkeurig de schattingen zijn. Er is nog steeds geen vaste methode om de olieprijzen en de andere factoren in de toekomst met 100% zekerheid te kunnen voorspellen.

Het is dan ook niet zo zeer de bedoeling met dit werkstuk een schatting te maken van de olieprijzen in de toekomst. De nadruk ligt meer op de gehele samenhang van factoren die elkaar beïnvloeden. Het gaat om de olieprijzen en de factoren die daarop betrekking hebben. Voor het analyseren van deze factoren die intern afhankelijk zijn wordt de *multivariate data analyse* gebruikt. Op de multivariate data wordt de *Principal Component Anlaysis* losgelaten. Met de eerste en de tweede principale component kan de informatie die in de multivariate data zit worden samengevat in univariate data.

De univariate data is dan weer een tijdsreeks die met behulp van *Time Seris Analysis* gemodelleerd kan worden. Zo kunnen we de vector, die onze oorspronkelijke data beweegt, analyseren en daar een model voor zien te vinden. Hebben we eenmaal het geschikte model gevonden, dan kunnen er toekomstige waarden worden geschat voor deze tijdreeks. Ik zal genoegen nemen met 1 jaar vooruit.

Met de geschatte waarde van de univariate tijdreeks voor het toekomstige jaar, kan de tijdreeks weer zo veel mogelijk informatie weergeven van het gedrag van de oorspronkelijke variabelen en wel 1 jaar vooruit. En zo kan de univariate tijdreeks weer divergeren naar het oude model met een extra waarde voor alle variabelen.

Met dit werk is er een analyse gevoerd op een multivariate dataset. Er is geprobeerd zo veel van de informatie over de onderlinge afhankelijkheden mee te nemen in het doen van voorspellingen voor het volgende jaar en dat voor zowel de olieprijzen als voor alle andere economische aspecten in het onderzoek.

1.2 Benadering

Met de kennis van de statistische modellen, de wiskundige analyses en het statistische programma R zullen alle variabalen worden ingevoerd als vectoren. Deze vectoren worden in een matrixvorm gebonden om de analyses erop los te laten.

Het onderzoek zal de volgende stappen volgen:

- De olieprijzen bestuderen en waar die van afhangen
- De economische variabelen selecteren die het meest van elkaar afhangen
- De variabelen invoeren als vectoren in een matrixvorm in R
- Principal Component Analysis toepassen
- De eerste en de tweede principale component analyseren en modelleren als een univariate tijdreeks
- Voorspellingen doen met behulp van de Time Series Analysis
- De voorspelde toekomstige waarden voor de oorspronkelijke data reconstrueren
- De resultaten analyseren en conclusies trekken

1.3 Structuur

Het werkstuk begint met de studie naar het gedrag van olieprijzen en de andere economische variabelen die ermee samenhangen en wel in Hoofdstuk 2. Hoofdstuk 3 geeft een theorie uitleg over Principal Component Analysis en de Time Seris Analysis. Het echte werk begint in Hoofdstuk 4 waar de analyses op de data zullen worden toegepast. Als afronding van het onderzoek zullen de resultaten en conclusies in Hoofdstuk 5 aan bod komen.

Achter in het verslag zijn er appedices beschikbaar voor diegenen die geineteresseerd zijn in wat meer details over het modelleren van de univariate data (Appensix I) of in de R codes, die terug zijn te vinden in Appendix II.

Hoofdstuk 2 De Olieprijzen en de Samenhangende Factoren

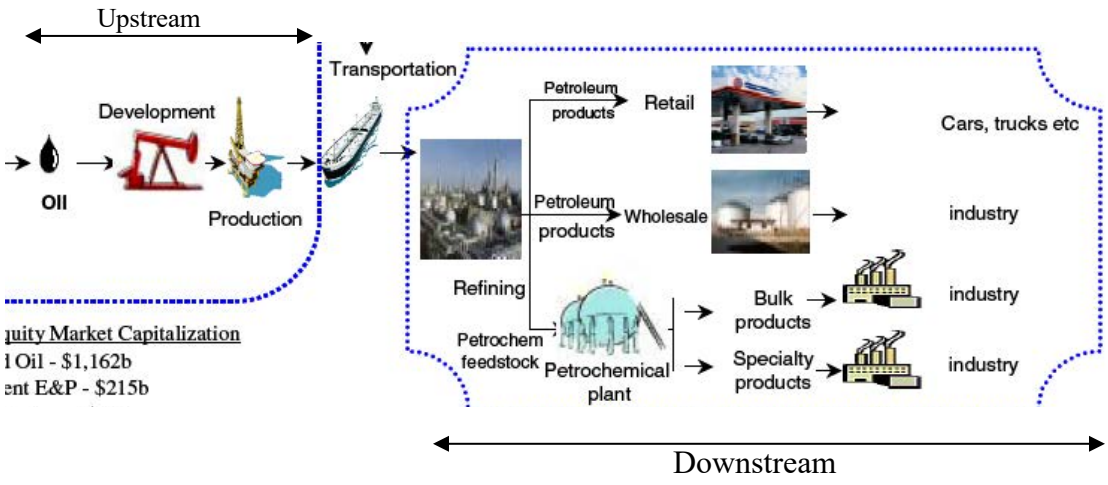
De oliemarkt is een wereldmarkt. Olie wordt in bijna ieder continent geproduceerd. Zo bestaat er een groot complex systeem van transport en raffinaderijen om zowel ruwe olie als olieproducten (benzine, diesel, etc) voor de eindgebruiker beschikbaar te stellen.

Olie wordt overal in Amerikaanse dollars verkocht en verhandeld in de belangrijkste wisselkoersen. Olie, in welk vorm dan ook, wordt gekocht of verkocht door ieder individu, bedrijf of land. Het wordt vervoerd via pijplijnen, tankschepen, en vrachtwagens. Oliemaatschappijen en onafhankelijke raffinaderijen verwerken de ruwe olie tot producten en verkopen die weer in groothandel en detailhandel.

De keten wordt onderscheiden in twee delen:

- Het leveren van ruwe olie (upstream)
- Het leveren van de geraffineerde olieproducten (downstream)

Zie Figuur 1 voor een schematische weergave.



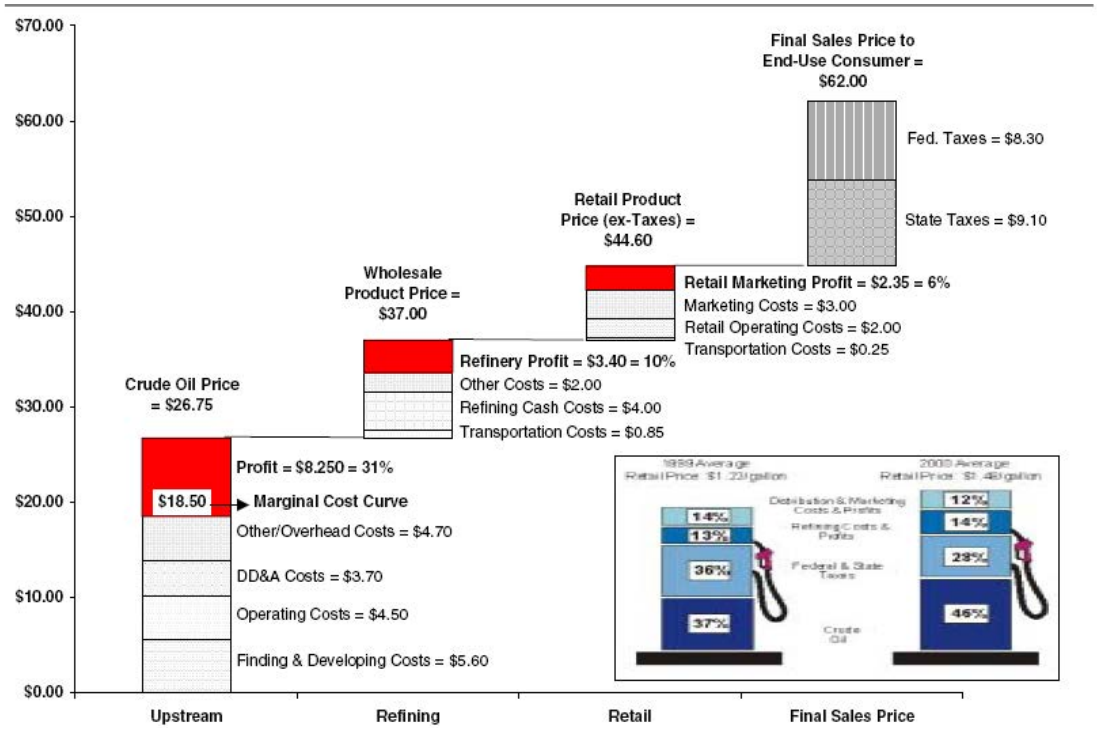
Figuur 1

De olie wordt geproduceerd door de oliemaatschappijen, waarna de raffinaderijen het werk overnemen door olie aan te schaffen van de oliemaatschappijen. Olieraffinaderijen verwerken de ruwe olie en verkopen de olieproducten aan de onafhankelijke distributeurs of detailhandelaars. Deze verkopen de olieproducten weer aan de eindgebruikermarkten (industriële, commerciële, woonwijken en vervoer) voor de uiteindelijke consumptie.

Vanaf de exploratiefase van olie tot aan de verkoop van het eindproduct is er een keten van kosten die gemaakt worden om van ruwe olie een eindproduct te maken. Deze kosten worden in de uiteindelijke olieprijs verrekend en worden door de consument betaald.

Geologisch gezien ligt de olie dichter bij de oppervlakte in het Midden Oosten en vooral in Irak en Saoedi-Arabië. Dit maakt het makkelijker de olie uit de grond te halen. Zo zijn de kosten van oliewinning kleiner dan 10 dollar cent per *barrel* ($cent/bbl). Een barrel is 195 liter. Dit terwijl de productiekosten in de Verenigde Staten of Venezuela waar de olie diep in de grond ligt veel duurder zijn en zo liggen de kosten niet onder de 12 $/bbl. Vanwege het kostenverschil moet er een gelijke basis voor de olieprijs worden gehanteerd wat de *differential rate* voorschrijft. Deze differential rate dwingt af dat de basiskosten gelijk moeten zijn aan de duurdere kosten. Zo is de basisprijs gelijk aan de duurdere kosten en wel beginnend met rond de 12 $/bbl ruwe olie.

De oliemaatschappij die de kosten van de exploratie van olie en de oliewinning betaalt verkoopt de olie door met een bepaalde winst wat rond de 20 $/bbl kan zijn. Raffinaderijen maken hun eigen kosten en sommeren daarbij de benodigde transportkosten van de olievelden naar de raffinaderij om de olieproducten met een bepaalde winst door te verkopen aan de detailhandelaars. Deze detailhandelaars voegen de transportkosten en de marketingkosten toe om zo per barrel een prijs te geven die de kosten dekt en een winst wordt behaald, wat de eindprijs/olieprijs per barrel wordt. Een voorbeeld van deze keten van kosten zal in Figuur 2 worden afgebeeld.



Figuur 2

Uit Figuur 2 kan worden opgemaakt dat de schommeling in de olieprijs kan worden veroorzaakt overal in de keten beginnend met de schommeling in de exploratiekosten, transportkosten tot aan de marketingkosten. Daartussenin kunnen nog vele factoren een rol spelen zoals de rente, de inflatie, etc.

De olieprijs blijft een fenomeen dat moeilijk is te voorspellen. Gaat de olieprijs omhoog of omlaag de komende maanden of blijft hij constant en voor hoelang?
Vele onderzoeken en vele speculaties op dat gebied zijn al gedaan. Voor vele belanghebbenden speelt de hoogte en de fluctuatie van de olieprijs een belangrijke rol. Zo zullen bij hogere olieprijzen door oliemaatschappijen meer investeringen in de oliesector plaatsvinden en tegelijkertijd zullen aandeelhouders van hogere olieprijzen profiteren.

De laatste jaren werd de oliemarkt gekarakteriseerd door de hoge ongebruikelijke combinatie van enerzijds de verzwakking van grondbeginselen van olieprijzen op korte termijn en anderzijds de stijging van de lange termijn prijzen. De stijging in de lange termijn prijzen heeft geleid tot hogere kosten in de olie-industrie. De markt gaat namelijk over van de winningfase met de daarbij behorende uitputting van de beschikbare capaciteit, naar een nieuwe investeringsfase waarbij naar nieuwe velden worden gezocht. De grotere vraag naar nieuwe velden terwijl de bronnen zijn gelimiteerd; reserves zijn onbereikbaar, werkers, installatie en technologie zijn allemaal de reden voor de hoger wordende productiekosten.

Onderzoeken wijzen erop dat de olieprijzen in de toekomst verder zullen stijgen. Deze stijging is te wijten aan de oplopende onderliggende fundamenten voor een hogere olieprijs. Een warme winter of een dalende economische groei kunnen op de korte termijn hun invloed op de olieprijs uitoefenen met dalende prijzen. Echter de grondbeginselen drijven de olieprijzen omhoog over de tijd heen en niet de speculaties.

Zoals gezegd wijzen de onderliggende grondbeginselen voor de oliesector naar hogere olieprijzen voor de komende jaren. In de toekomst zullen er zeker periodes zijn waarin de volatiliteit van de prijs omlaag gaat, maar er zal uiteindelijk sprake zijn van een totale stijging. Dit wordt ondersteund door het feit dat:

- De groei in de vraag naar olie laat een lage elasticiteit met de olieprijs zien.
- De reserve capaciteit van OPEC is gelimiteerd.
- De geopolitieke onrust blijft hoog.

Naast de vraag naar olie en de reserve capaciteit zijn er andere factoren die van invloed kunnen zijn op de prijs van olie. Deze factoren veranderen van plaats tot plaats en kunnen zowel locaal als wereldwijd van betekenis zijn. De belangrijkste factoren zijn:

- De olieproductie
- De productiecapaciteit van olie
- De reserve capaciteiten
- De economische groei
- De transportcapaciteit

- De staalprijzen
- De capaciteit van de raffinaderijen:
    1. Er vinden geen grote investeringen plaats in de bouw van de raffinaderijen
    2. De beperkingen van het milieu aspecten; er kunnen niet overal raffinaderijen worden gebouwd
- De onverwachte politieke calamiteiten
- De onverwachte natuurrampen
- De macht van de oliemaatschappijen
- De belastingen
- Het gemiddelde consumptieniveau om bepaalde levensnormen te behouden; in Amerika is het consumptieniveau hoger dan het Europese niveau (de autosoorten, huizen, etc)
- Het onder controle houden van de olieproductie; soms wordt de productie gestopt in bepaalde gebieden om grotere hoeveelheden vanuit andere gebieden te produceren waar de productiekosten lager zijn.

Er zijn nog vele factoren die een directe en een indirecte invloed kunnen uitoefenen de olieprijs. Sommige factoren zijn sterk afhankelijk van de hoogte van de olieprijs en andere weer in mindere mate. En sommige factoren zijn ook onderling afhankelijk en sommigen ook weer niet. Er zijn ook factoren die af te leiden zijn uit andere factoren. Het is moeilijk alle factoren te kennen, maar de belangrijkste zijn zeker genoemd. Voor dit onderzoek zullen een paar factoren onder de loep worden genomen. Deze factoren zullen voor onze statistische analyses als vectoren worden gezien.

2.1 Onderzoeksfactoren

Het is nu zaak, de samenhangende factoren te bekijken en te zoeken naar een manier om de olieprijs en de samenhangende factoren te modelleren.

Het is moeilijk alle factoren in beschouwing te nemen. Uiteindelijk zijn niet alle factoren te kwantificeren en ze zijn ook niet allemaal even belangrijk. Het einde van de Amerikaanse presidentiele verkiezingscampagne en de komende verkiezingen in Irak voorspellen een zekere rust in de politiek en zo een betere controle over de olieprijs. Zulke geopolitieke gebeurtenissen zijn echter onvoorspelbaar, niet duurzaam en ook nog niet te kwantificeren. Verder is de macht dat OPEC kan uitoefenen op de olieproductie ook niet te kwantificeren en sommige factoren zijn nog minder van belang. Bovendien beschouw ik in mijn onderzoek de jaargemiddelden. Zo zullen onverwachte gebeurtenissen beperkte invloed hebben vanwege het effect van het nemen van gemiddelden.

In dit werkstuk zullen de factoren in beschouwing worden genomen die meer betekenisvol zijn. Verder hebben onderzoeken aangetoond dat er fundamentele veranderingen zijn in de recente jaren die niet vergeten mogen worden, zoals:

- de grotere vraag naar olie in China, tegenover Amerika aan de andere kant

- de stijgende upstream kosten, wat een omkeer vormt van de dalende trend in de kosten van de jaren negentig
- de stroefheid in de complexe raffinaderijencapaciteit die in staat is de hoeveelheid olie te verwerken en de bottlenecks in de scheepvaart

Aan hand van de bovengenoemde belangrijke veranderende hedendaagse veranderingen en de vele artikelen en onderzoeken die zijn gedaan, zal ik voor dit onderzoek de volgende factoren onder de loep nemen en dat op wereldbasis:

1. De olieprijzen ($/bbl)
2. De reserves (miljoen barrels)
3. De olieproductie (1000 barrels per dag)
4. De olie export (1000 barrels per dag)
5. De raffinaderijencapaciteit (1000 barrels per dag)
6. De productie van geraffineerde olieproducten (1000 barrels per dag)
7. De consumptie van geraffineerde olieproducten (1000 barrels per dag)

Bij dit onderzoek zijn de jaarlijkse metingen voor de jaren 1960-2004 in beschouwing genomen. Zo hebben we voor iedere factor 45 waarnemingen wat niet al te veel is maar wel voldoende om wat statistische technieken erop los te laten.

Alle factoren zullen worden gebruikt voor de analyses en de modellering. De metingen van deze factoren worden onze data waarop de analyses op worden gedaan. Dit en meer in de volgende hoofdstukken.

Hoofdstuk 3 Multivariate Data Analyses

Multivariate data bestaan uit waarnemingen van verschillende variabelen voor een aantal individuen of objecten. Zulke data komen in alle taken in de wetenschap voor, van psychologie tot aan biologie. Methodes om de multivariate data te analyseren vormen een toenemend belangrijk gebied in de statistiek.

Er zijn verschillende voorbeelden van multivariate data:

1. Tentamenresultaten: elke student heeft een cijfer voor verschillende tentamens en zo hebben verschillende studenten verschillende cijfers voor verschillende tentamens. Hier zijn de variabelen de verschillende vakken waarvoor een tentamen wordt afgelegd en de individuen zijn dan de studenten die de tentamens afleggen.
2. Zo kunnen er ook medische resultaten en archeologische resultaten etc. worden samengevat en weergegeven in tabellen als multivariate data.

Er bestaan verschillende technieken voor het analyseren en het modelleren van multivariate data. De keus voor de meest geschikte methode hangt van het type data af, van het type probleem en van het soort object waar de analyses op worden losgelaten. De bedoeling van deze technieken is vereenvoudiging; grootschalige gegevens samen te vatten in relatief weinig parameters. De multivariate technieken zijn tevens onderzoekend; ze genereren hypotheses meer dan ze te testen.

Het is ook van belang een verschil te maken tussen de technieken, die gericht zijn op het analyseren van de relatie tussen de variabelen, en de technieken die juist gericht zijn op het analyseren van de relatie tussen de individuen. Het zou voor het bovengenoemde voorbeeld betekenen dat er verschil is wanneer we twee studenten met elkaar willen vergelijken of juist meer iets willen zeggen over de resultaten van twee verschillende vakken.

De factoren die beschreven zijn in het vorige hoofdstuk zullen hier verder worden verwerkt. Deze factoren kunnen worden gezien als multivariate data. Deze data bestaan uit waarnemingen van verschillende variabelen gemeten ieder jaar. In dit geval zijn er voor iedere variabele 45 waarnemingen (metingen vanaf het jaar 1960 tot en met het jaar 2004) Dit is tevens het grootste aantal verkrijgbare waarnemingen die te vinden zijn. En ieder jaar heeft dus 7 waarnemingen, iedere variabele heeft 45 waarnemingen.

Wanneer we maar twee variabelen hebben, bijvoorbeeld de olieprijzen en de olieproductie, dan zou een mogelijke analyse bijvoorbeeld uit het kijken naar de correlatie tussen de twee variabelen kunnen zijn. Een beter toepasbare analyse is te doen met behulp van de *Principal Component Analysis*. Met deze techniek kan dan ook een 2D visualisatie worden nagebootst. Met meerdere variabelen, zoals in dit geval, is de techniek Principal Component Analysis zeer nuttig. De vraag wat deze techniek inhoudt en wat haar toepassingen zijn, zal in het volgende hoofdstuk worden beantwoord.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) heeft als doel de afhankelijke waarnemingen te transformeren in een nieuwe set van onafhankelijke variabelen (principal components genoemd), geordend in een afdalende mate van belang. Het voornaamste doel is de dimensionaliteit van het probleem te reduceren en nieuwe variabelen te vinden, de scores op de principale componenten, zodat de data beter valt te begrijpen (met score plots en biplots).

Deze nieuwe ongecorreleerde variabelen zijn genormeerde lineaire combinaties van de originele variabelen, geordend in dalende orde van variantie. Deze lineaire combinatie zou uit gewichten bestaan die voor ieder individu (jaar) alle gerelateerde combinaties uit de data een nieuwe waarde geeft. Zodat bijvoorbeeld de eerste principale component het meeste verklaart van de variatie in de originele data. En zo bevat de eerste principale component de meeste informatie uit de originele data. De PCA levert namelijk een set van genormeerde eigenvectoren als *loadings* op. De *scores* van de data op deze principaple componenten worden de nieuwe matrix met ongecorreleerde vectoren.

Om dit nader toe te lichten is het handig eerst wat notaties uit de lineaire algebra te demonstreren. Deze basiskennis is nodig om de werking van de PCA wiskundig weer te geven.

Om de PCA toe te passen worden de waarnemingen in een matrixvorm gezet.
Laat X_t^j de jaarlijkse waarnemingen van de j^{de} vector zijn. Dan is $[\, X_t^1 \ \ldots \ X_t^j \,]$ een t x j matrix van waarnemingen voor t = 1, 2, …, 45 en j = 1, 2, …, 7

Aangezien de variabelen in verschillende grootheden zijn gemeten (in $/bbl of in 1000 bbl/dag), moet er standaardisatie worden uitgevoerd alvorens de PCA uit te voeren. Zo worden alle variabelen geschaald op gemiddelde 0 en variantie 1. Laat $\widetilde{Y}_t^j$ de gestandaadiseerde j^{de} vector zijn.

Nu is het doel van de PCA om een orthogonale p x p matrix (de loadings)

$$L = [L_1 \ldots L_p] \text{ en de scores } S = [S_{\mathbf{1}}, \ldots, S_t^7]$$

te vinden, die de verandering bepaalt in de variabele, $\widetilde{Y} = SL$, of eigenlijk

$$\begin{bmatrix} \widetilde{y}_1^j \\ \widetilde{y}_2^j \\ \ldots \\ \ldots \\ \widetilde{y}_t^j \end{bmatrix} = [S_{\mathbf{1}}, \ldots, S_t^7] \begin{bmatrix} l_1^j \\ l_2^j \\ \ldots \\ \ldots \\ l_p^j \end{bmatrix}$$

Waarbij
$\widetilde{y}_t^j$ = de gestandaardiseerde waarneming van de j[de] vector (coördinaten van $\widetilde{Y}_t^j$)
S_t^j = de scores op de j[de] prinsipale component
l_p^j = de loadings op de j[de] principale component voor p = 1, …, 7
Met de belangrijkste eigenschap dat de scores ontgecorreleerd/onafhankelijk zijn en dat $\widetilde{Y}$ een lineaire combinatie is van de scores met de loadings als gewichten.

De unit vectoren S_1, … , S_7 zijn niets anders dan de scores van de originele data. En de eerste principale component is dan de lineaire combinatie die wordt gedefinieerd door de eigenvector, dat correspondeert met de hoogste eigenwaarde van de covariantie matrix, de tweede principale component correspondeert met de tweede hoogste eigenwaarde, enz.

Zo kan de eerste principale component de gestandiseerde vector $\widetilde{Y}_t^j$ bepalen.
Laat l_1, … , l_p de entries (loadings) zijn van het eerste principale component zijn. Dan is $\widetilde{y}_t^j$ de waarneming met betrekking tot $\widetilde{Y} = S\mathrm{L}$ als volgt uiteen te zetten:

$$\widetilde{y}_t^j = s_t^1 l_p^1 + s_t^2 l_p^2 + \ldots + s_t^7 l_p^7$$

Zoals eerder is besproken, heeft de PCA een nieuwe variabele aangemaakt die een lineaire combinatie is van de originele waarnemingen, gebruikmakend van de eigenvector L_1 als gewichten. Op dezelfde manier bepaalt L_2 de gestandaardiseerde variabele $\widetilde{Y}_t^j$, enzovoorts.

3.2 Time Series Analysis

Vaak bestaat een tijdreeks uit herhaalde metingen van hetzelfde object zoals in dit onderzoek het geval is. Zo is de gebruikelijke aanname van onafhankelijkheid niet van toepassing voor tijdreeksen. Time Series Analysis heeft betrekking op statistische analyses van een reeks van afhankelijke variabelen.

Het doel om analyses uit te voeren met time series voor dit onderzoek, is om de verkregen scores als een tijdreeks te modelleren. Met zo'n model kan deze univariate data in de tijd worden gemodelleerd. Bovendien kunnen de individuele variabelen ook als tijdreeksen worden gezien. Met een goede benadering kunnen toekomstige waarden worden voorspeld wat veel kan zeggen over het model dat we hebben gemaakt met principal component analysis. Deze benadering kan de verandering die onze afhankelijke data meemaakt in de tijd visualiseren om een beter inzicht te verschaffen en uiteraard de meest belangrijke informatie weer te geven in de voorspellingen voor de toekomst.

Voordat de analyses uitgevoerd kunnen worden is het wel van belang de data te filteren van de mogelijke aanwezigheid van de trend. Dit is nodig om stationaire data te genereren die gemodelleerd kan worden. Het filteren is met *differencing* te doen, wat

in het volgende hoofdstuk aan bod zal komen (zie 4.1)

Voor de scores die we willen benaderen kunnen we een AR(p), MA(q) of een ARMA(p,q) model aannemen. Er zijn methoden (zie Appendix I) die helpen bij het kiezen van het meest geschikte model. Het blijft wel de vrije keus van de gebruiker die naar de beste benadering kan kijken en voor een bepaald model kan gaan.

In dit onderzoek zal het AR(p) proces worden gebruikt om de scores te benaderen. Het model ziet er dan als volgt uit:

$$S_t = \alpha S_{t-1} + \ldots + \alpha_p S_{t-p} + Z_t$$

Waarbij
S_t = score in jaar t
S_{t-1} = score in jaar t-1
α = een constante
Z_t = een white noise proces met variantie σ^2

Dit AR model is gekozen, omdat het logischerwijs is aan te nemen, dat de huidige waarde van de olieprijzen of de olieproductie of ieder andere variabele, afhangt van de meest recente waarde van p jaren, plus een bepaalde ruis. Meestal is de hoogte van de huidige olieprijs grotendeels te wijten aan de hoogte van de olieprijs van een jaar eerder of meer. Het is aan de gebruiker de keus in hoeverre de voorspelling van het komend jaar te laten afhangen van de eerdere jaren en wat voor gewichten daaraan weer toe te kennen.

Dit model representeert als meeste de wereld van onze economische factoren. Nader onderzoek heeft ook getoond dat de ruis die de voorspelling in zich heeft niet veel verschilt wanneer we voor andere tijdreeksmodellen kiezen. Zo blijft het AR(p) model een geschikte keus.

Nu een korte uitleg over de methoden en hun toepassing is gegeven, zal Hoofdstuk de analyses representeren. Er zal dan ook meer uitleg worden gegeven over de methoden en hun toepassingen en de volgorde van gebruik.

Hoofdstuk 4 Toepassing

Allereerst worden alle variabelen als vectoren ingevoerd. Iedere vector bestaat uit 45 waarnemingen. Vervolgens maken we een matrix aan van alle 7 variabelen die voor de analyse gebruikt zullen worden.

Deze variabelen kunnen ook als time series worden gezien aangezien de metingen in de tijd zijn genomen en juist het gedrag van de metingen in de tijd geanalyseerd moet worden. De afhankelijkheid tussen de variabelen is van groot belang. De metingen zijn jaarlijkse gemiddelden.

De toepassing van de uitgelegde theorieën op de data heeft tot een bepaald stappenplan geleid. Zo zal het werk ook in die stappen worden gedemonstreerd en wel in de volgende stappen (zie Appendix II voor de gebruikte R code):

4.1 Stationaire data genereren met differencing

Om stationaire data te genereren, moet er eerst de trend en seizoen effecten worden weggehaald. In de olieprijzen en de relaterende variabelen zit er een zekere trend in de tijd (zie hoofdstuk 1). Seizoeninvloeden kunnen voorkomen in bepaalde maanden, maar op het jaargemiddelde is er geen invloed te merken. Bij een warmere winter bijvoorbeeld is de vraag naar olie lager maar zal het jaargemiddelde niet beïnvloeden.

In dit onderzoek wordt er dus aangenomen dat iedere variabele

$$X_i = m_i + Y_i$$

waarbij
X_i = waarnemingen van de vector i
m_i = de (deterministische) trend
Y_i = de nieuwe waarnemingenvector

Hier proberen we de trend weg te halen door voor iedere vector de jaarlijkse verschillen weg te halen (vandaar Differencing). Laat het verschil ∇X_i van orde k (k = 1, 2, …) zijn. Dan is

$$Y_t^j = \nabla X_t^j = X_t^j - X_{t-k}^j$$

Waarbij
X_t^j = waarneming van de j^{de} variabele gemeten in jaar t
X_{t-k}^j = waarneming van de j^{de} variabele gemeten in jaar t-k

De orde van k is door de gebruiker zelf te bepalen. Er wordt vaak volstaan met de orde van een of twee. In dit geval zullen de jaarlijkse verschillen worden genomen (dus orde 1).

Iedere vector Y_t^j bestaat uit de verschillen tussen opeenvolgende jaren. Zo is iedere vector gedifferenceerd en dus verlost van de trendinvloeden. De aanname kan dan ook worden gemaakt dat deze vectoren stationair zijn. Onze nieuwe matrix wordt een *44* x *7* matrix.

4.2 De PCA uitvoeren

Alvorens de principale componenten uit te rekenen, moeten eerst de vectoren worden gestandaardiseerd. Dit is nodig omdat iedere vector waarden bevat van verschillende grootheden. Zo zijn de metingen van de olieprijzen in $/bbl en de olie export in duizenden barrels per dag (zie 2.1)

Laat $\widetilde{Y}_t^j$ de gestandaardiseerde vector van Y_t^j zijn, dan:

$$\widetilde{Y}_t^j = \frac{Y_t^j - \overline{Y}^j}{\sigma_y^j}$$

Waarbij

Y_t^j = vector j
$\overline{Y}^j$ = het gemiddelde van Y_t^j
σ_y^j = de standaarddeviatie van Y_t^j

De PCA techniek gaat nu de *principale componenten*, de *loadings* (de loadings van de oorspronkelijke data op de principale componenten en de *scores* (de jaarlijkse scores op de principale componenten) uitrekenen. Voor het gemak nomen we de PCA hier pcM (de principal component analysis van matrix M)

Om een goed inzicht te verkrijgen over de berekende principale componenten, worden er de varianties van de principale componenten met elkaar vergeleken (zie Figuur 3)

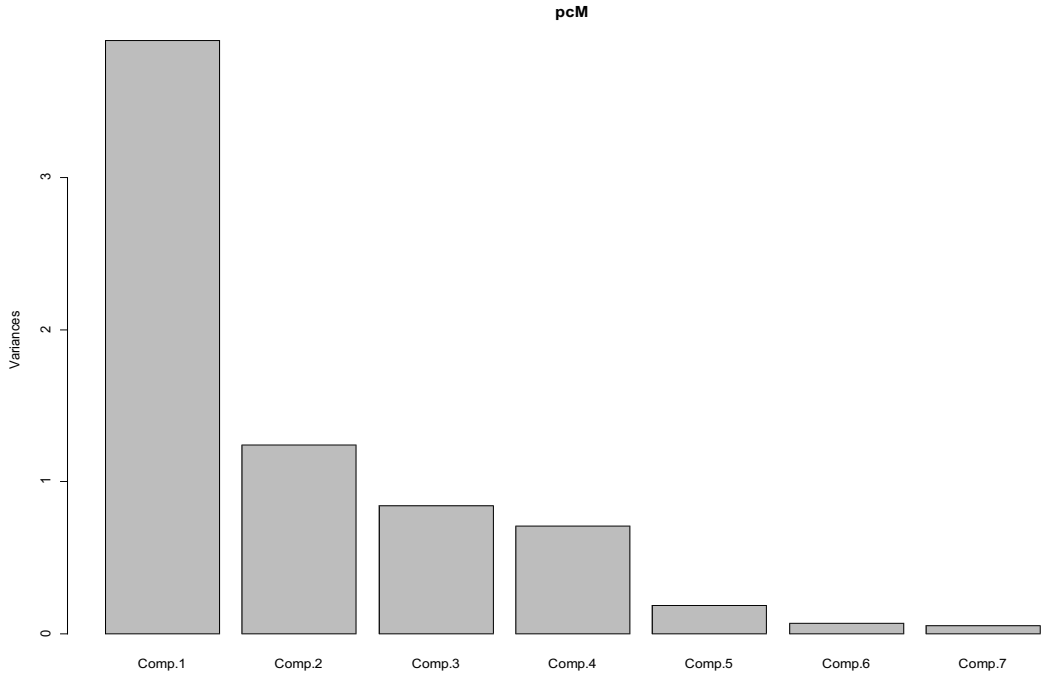


Figure 3

De plot laat zien dat de eerste component uiteraard de grootste variantie heeft en zo ook de belangrijkste is (zie hoofdstuk 2) De eerste principale component bevat ook de meeste informatie over de oorspronkelijke data (ongeveer 55% verklaarde variantie). Zo kunnen met deze principale component de oorspronkelijke data weer worden geschat. De tweede component zou ook belangrijke informatie bevatten maar dan in mindere mate dan de eerste component. Het is ook interessant om deze tweede principale component te analyseren.

De loadings van de oorspronkelijke data op de eerste twee principale componenten nemen de volgende waarden aan:

	Eerste component	Tweede component
Oil Prices		0.662
Resevers		-0.650
Production	-0.480	
Refinery Capacity	-0.303	0.346
Oil Exports	-0.462	
Consumption Refined Products	-0.473	
Production Refined Products	-0.486	-0.103

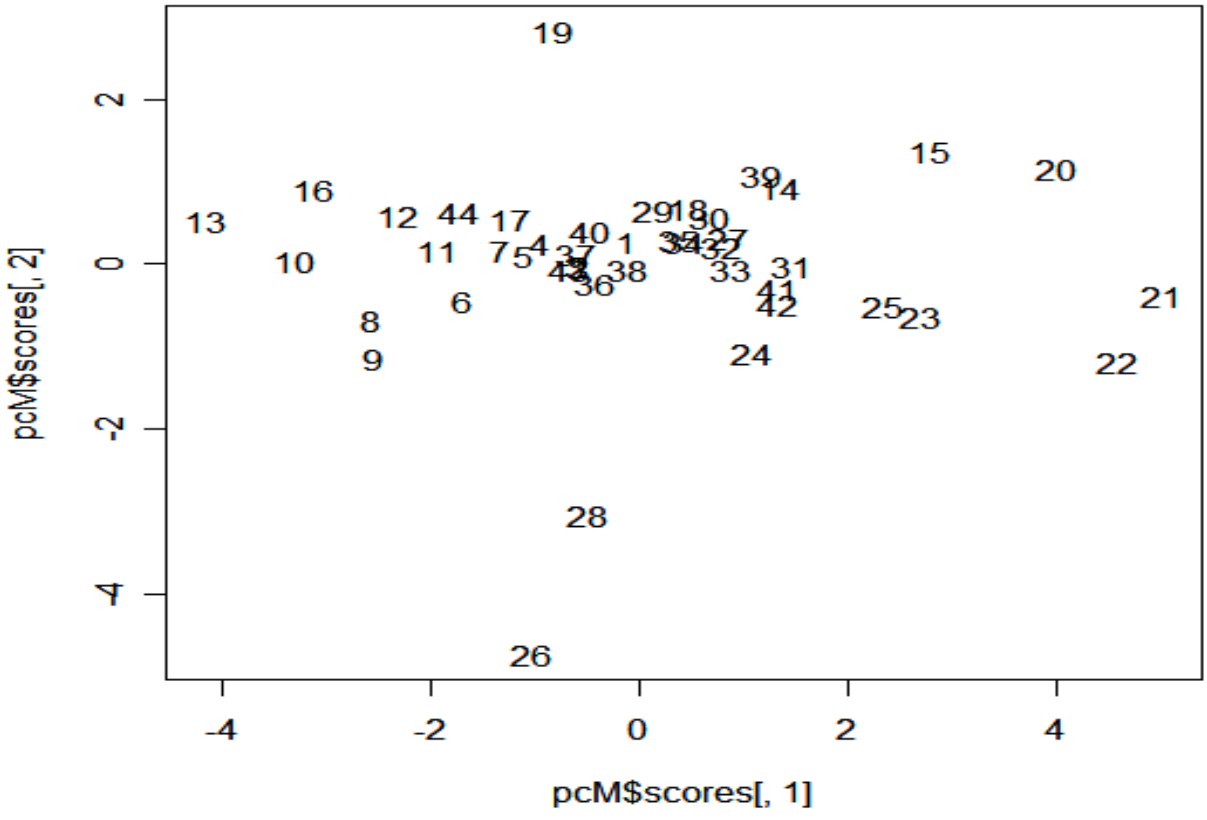
Figuur 4

Figuur 4 laat zien dat de oorspronkelijke data negatieve loadings hebben op de eerste principale component behalve de olieprijzen en de reserves. De laatst genoemde variabelen hebben hoge loadings op de tweede principale component. De olieprijzen

hebben een positieve loading terwijl de reserves juiste een negatieve loading hebben. De tegenovergestelde loadings zouden kunnen betekenen dat er een omgekeerde verband is in de oorspronkelijke data tussen de olieprijzen en de reserves. Dit terwijl de negatieve loadings van de andere variabelen op de eerste principale component zouden kunnen betekenen dat al deze variabelen meer aan elkaar zijn gerelateerd. Een daling in de productie kan bijvoorbeeld een daling in alle andere variabelen betekenen.
Zo kunnen er ook meerdere uitspraken worden gemaakt.

De berekeningen van de loadings door R laten ook zien dat de eerste en de tweede component naar schatting 75% van de variantie in de originele data verklaren. Dit is dan ook de reden om de eerste twee componenten te gebruiken voor nader onderzoek. De jaarlijkse scores op de eerste en tweede component zijn in Figuur 5 te zien.

De scores laten zien dat er outliers zijn en voornamelijk de jaren 19, 26 en 28 wat overeenkomen met de jaren 1978, 1985 en 1987. De reden voor het jaar 1978 zou bijvoorbeeld liggen aan de machtige invloed van de OPEC organisatie op de olieprijzen (waar de scores op de tweede component hoog zijn) voor de eerste jaren na de opzetting van de OPEC in 1960.

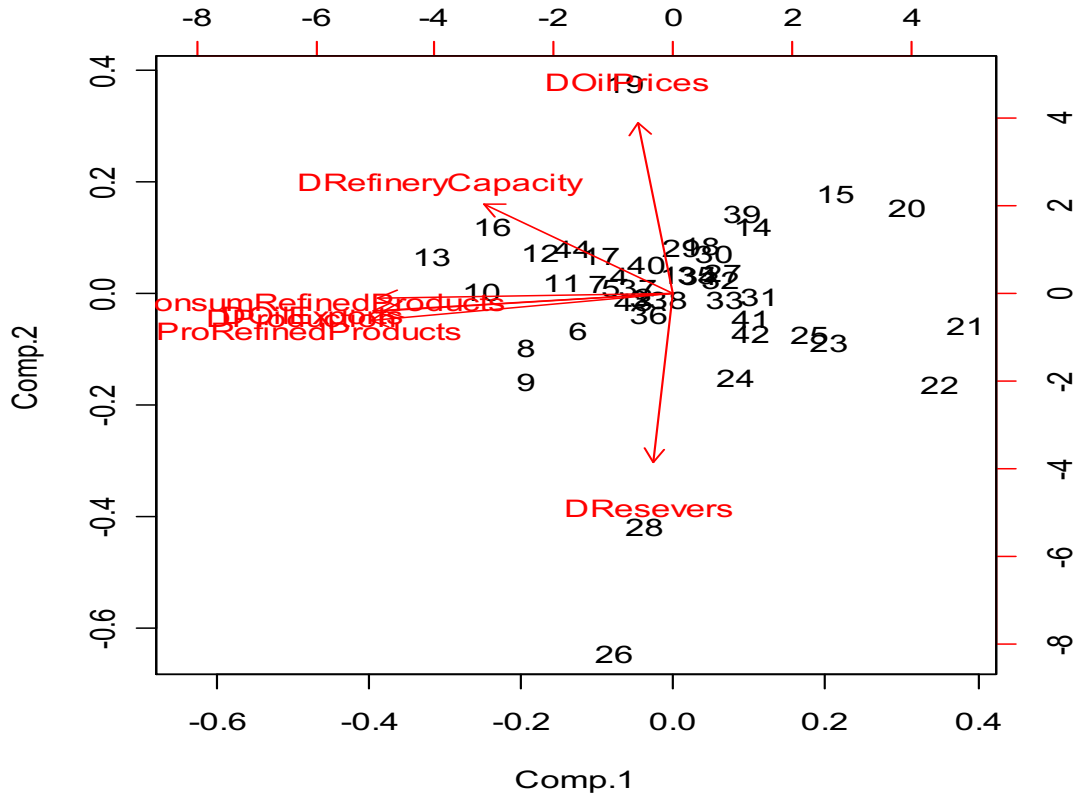


Figuur 5

De jaren die voor outliers zorgen kunnen worden verwijderd uit de data. Echter nadere onderzoek heeft getoond dat het verwijderen van deze outliers niet tot betere plotjes zal leiden en dat de outliers wel zullen blijven bestaan maar dan net door andere jaren. Het verwijderen van bepaalde jaren kan voor inconsistenties zorgen in de data. Dit omdat de vectoren gedifferenceerd zijn door verschillen van steeds 1 jaar te nemen (4.1). Door het wegnemen van sommige jaren zullen er gaten in de data ontstaan, wat de tijdreeksanslyse in de volgende stappen zal bemoeilijken. Bovendien geef ik de voorkeur aan het behouden van het aantal waarnemingen in het onderzoek. Er zal in dit geval meer dan 3 jaren worden verwijderd wat erg veel is.

Verder laat Figuur 6 duidelijk zien dat de loadings van de oorspronkelijke data op de eerste principale component dicht bij elkaar zijn en dezelfde kant op gaan. Dit terwijl de loadings op de tweede principale component juist het tegenovergestelde richting op gaan. Alleen de variabele Refinery Capacity heeft zowel loadings op de eerste als op de tweede princiaple component en in beide richtingen (positieve en negatieve).

Dit is ook in de loadings terug te zien (zie Figuur 4) waar de data en voornamelijk de olieprijzen hoge positieve loadings en de reserves hoge negatieve loadings hebben op de tweede component. Tegelijkertijd hebben alle andere variabelen hoge negatieve loadings op de eerste component.

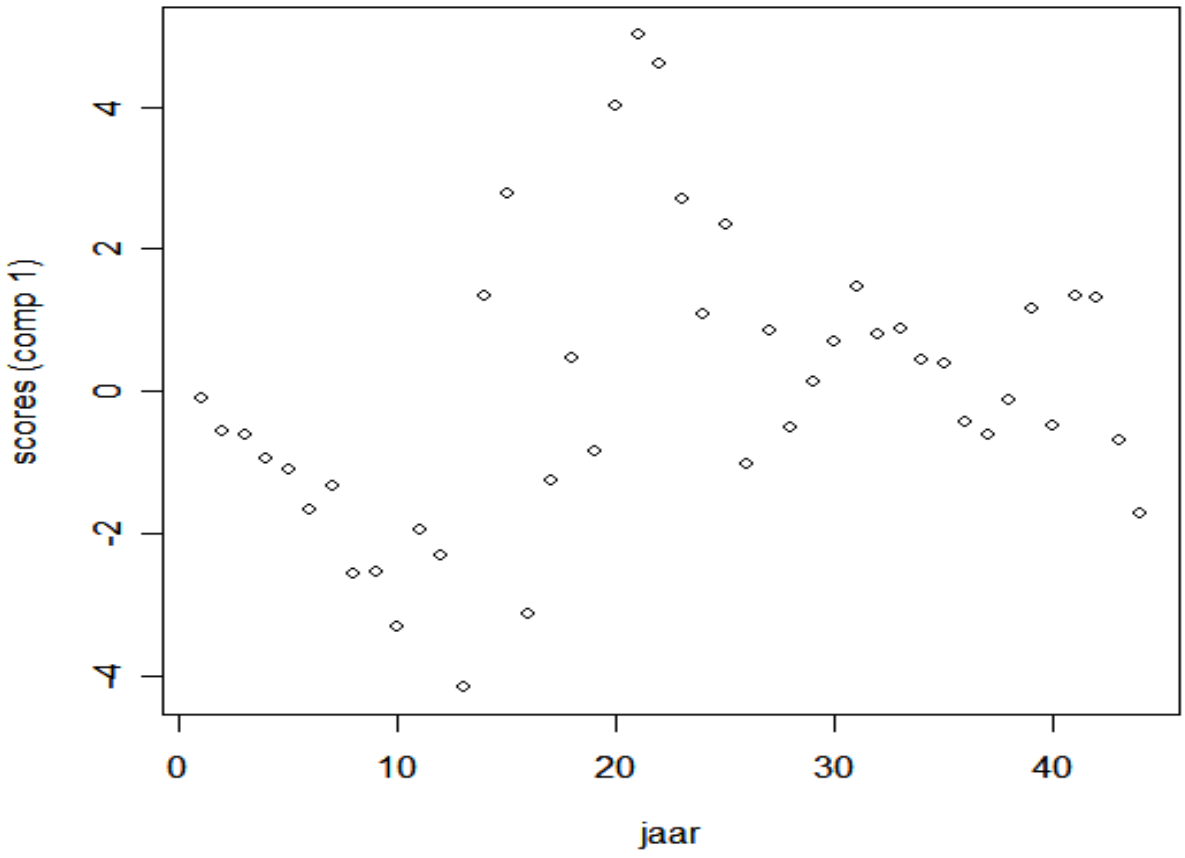


Figuur 6

Zo zal dit onderzoek eerst modelleren met de eerste principale component. Vervolgens zal ik logischerwijs met de eerste en de tweede component de originele data schatten om betere voorspellingen te verkrijgen.

4.3 Het modelleren m.b.v. Time Series Analysis

De scores van de oorspronkelijke variabelen op de ladingsfactors zijn in een matrixvorm uitgerekend. Zoals gezegd beperken we ons tot de scores van de eerste en de tweede principale componenten. De scores op ieder component worden nu gezien als univariate data gerekend in de tijd. Deze nieuwe vector is dan een tijdreeks waar we een model voor proberen te vinden. Het is wel handig eerst de eerste principale component te plotten (zie Figuur 7).

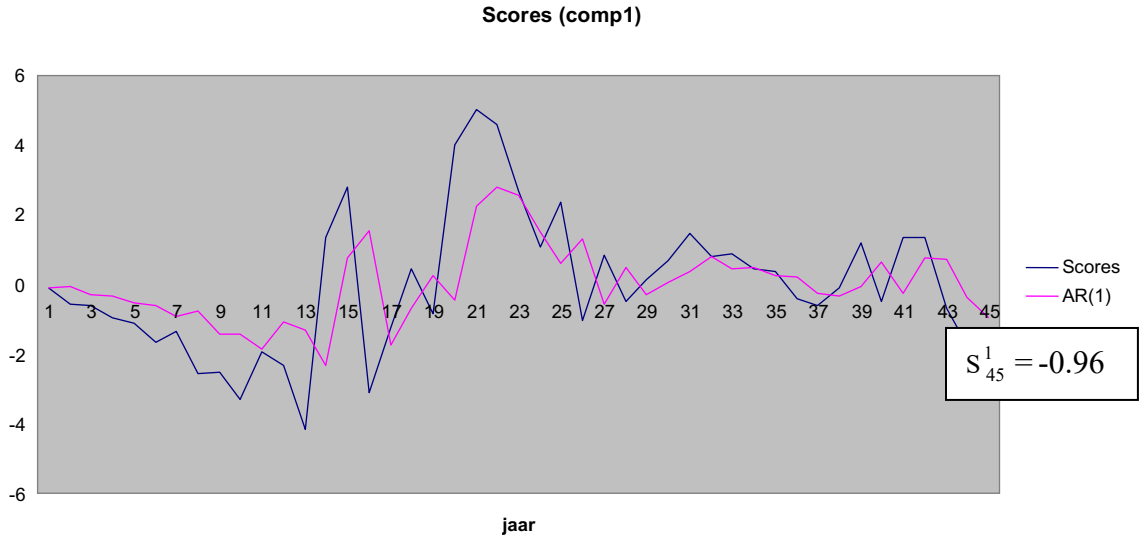


Figuur 7

De scores zullen worden gemodelleerd met een AR(p) model (Zie 3.2). De berekeningen geven aan dat de scores (S_t) met een AR(1) gefit kunnen worden en wel met de volgende waarden:

$$S_t = 0.5581 * S_{t-1} + Z_t, \qquad Z_t \sim N(0, 2.810)$$

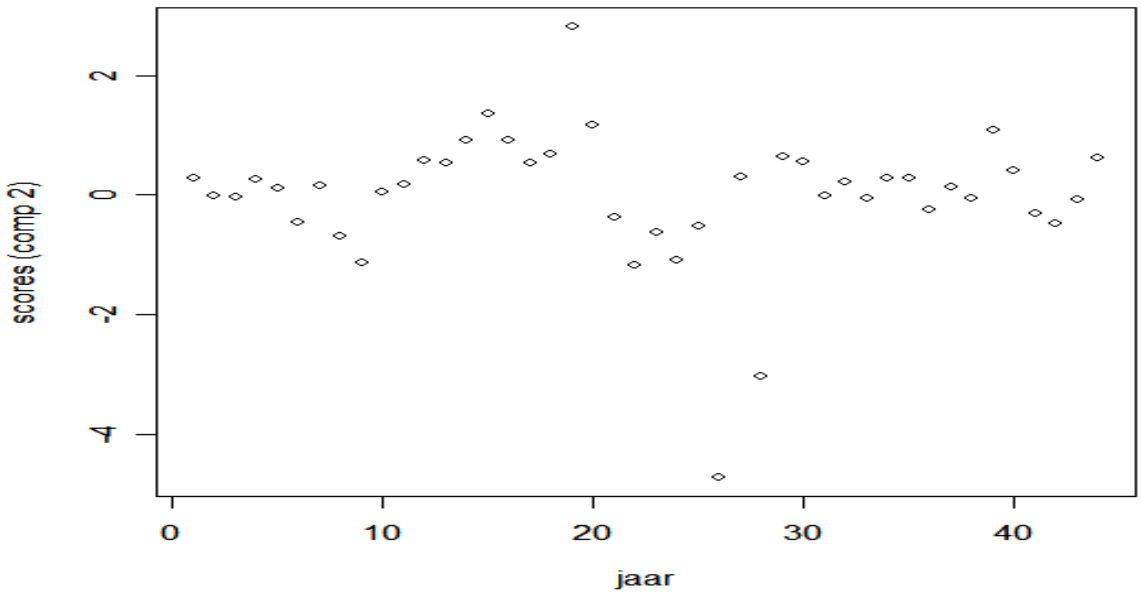
De benadering van de scores is te zien in Figuur 8. De voorspelde score (score 45)op de eerste principale component, S^1_{45}, is gelijk aan -0.96. De ruis is een random trekking uit de normale verdeling met een standaardafwijking van 1.68. Dit houdt in dat er met een betrouwbaarheidsinterval van 95% de voorspelling zal liggen in [-1.68, +1.86]. Volgens het model kan S^1_{45} met 95% in [-1,86 -0.96, -0.96+1.86] of beter gezegd in het gebied [-2.64, 0.72] zich bevinden.



Figuur 8

De breedte van de voorspelling die op ons model gebaseerd is, benadrukt dat de variantie van de ruis een zekere onzekerheid van de voorspelling benadrukt. Dit was al te verwachten, want onze data bevatten vele onderlinge afhankelijkheden. De scores op de eerste component verklaren maar 55% van de oorspronkelijke informatie en bovendien bestaat er geen model die perfect past bij de scores die grote verspreiding vertonen en van variabelen afkomstig zijn die moeilijk voorspelbaar zijn.

Met dezelfde beredenering modelleren we de scores op de tweede component. Hieronder zal de tijdreeks van de tweede component worden geplot (zie Figuur 9).

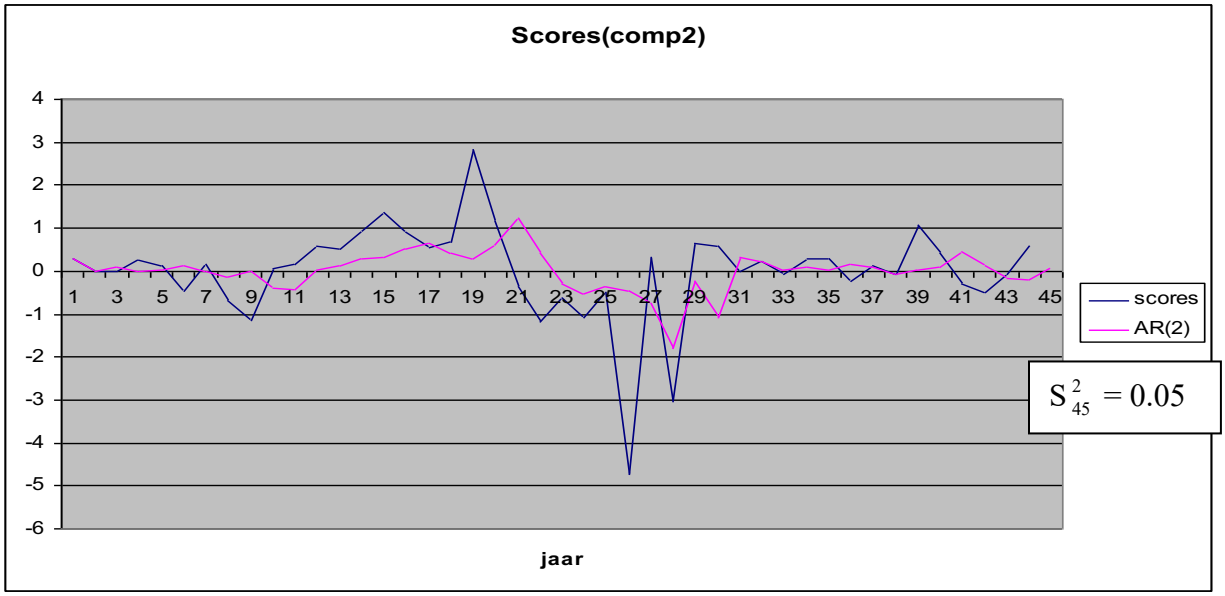


Figuur 9

Laat de scores van jaar t op de tweede component S_t zijn, dan is

$$S_t = 0.1205*S_{t-1} + 0.3851* S_{t-2} + Z_t, \qquad Z_t \sim N(0, 1.091)$$

Het AR model is zo gekozen dat de voorspelling meer afhangt van twee jaar eerder dan van het vorig jaar. Daar bovenop is er altijd een ruis die de onzekerheid in de voorspelling beschrijft. De grafiek hieronder geeft de nauwkeurigheid van het gekozen model weer (zie Figuur 10).



Figuur 10

De onzekerheid in de benadering geeft weer aan hoe slordig het modelleren kan zijn. Dit is te wijten aan de outliers in de scores en aan de moeilijkheidsgraad van het fitten van de tijdreeks.

Hier geeft de voorspelling van de nieuwe score een hele lage score aan, namelijk 0.05. De 95% betrouwbaarheidsinterval van de voorspelling ligt in [-0.99, 1.09]. Als de voorspelling voor de nieuwe score aan de hogere grens lag, dan zou de voorspelling voor de oorspronkelijke data in de komende stappen hoger uitkomen. Maar de lage score op de tweede principale component zal eigenlijk het gebruik van de tweede component bijna onnodig maken. Dit zal in de volgende stappen worden gedemonstreerd.

4.4 Het voorspellen van de nieuwe waarden

Na de nieuwe jaarlijkse score op de eerste twee componenten te hebben berekend, kunnen de waarden van de gestandaardiseerde variabelen worden teruggerekend. De berekende waarden van de scores zullen gebruikt worden. De boven- en ondergrens van het betrouwbaarheidsinterval zullen we nu niet meer gebruiken. Er kunnen wel uitspraken worden gemaakt hierover wanneer de voorspellingen van het nieuwe jaar voor de oorspronkelijke data zijn gedaan.

4.4.1 Het voorspellen met de eerste principale component

Het voorspellen met alleen de eerste principale component gaat als volgt:
De nieuwe metingen $\widetilde{Y}^{j}_{45}$ zijn niets anders dan (zie 3.1)

$$\widetilde{Y}^{j}_{45} = S^{1}_{45} * L^{1}$$

Waarbij
$\widetilde{Y}_{45}^{j}$ = de 45ste gestandaardiseerde waarde van de de j^{de} vector
S_{45}^{1} = de nieuw voorspelde score op de eerste component
L^{1} = loadings van de data op de eerste component.

Nu de $\widetilde{Y}_{45}^{j}$ berekend zijn voor iedere vector, moet de standaardisering worden opgeheven. Dat is als volgt te doen (zie 4.2):

$$\mathrm{Y}_{45}^{j} = \sigma_{y}^{j} * \widetilde{Y}_{45}^{j} + \overline{Y}^{j}$$

Rest ons nog de voorspelling te doen van het nieuwe jaar X_{45}^{j} voor iedere vector of beter gezegd de voorspelling van het jaar 2005. Hiervoor moet de differencing worden opgeheven (zie 3.1).

$$\mathrm{X}_{45}^{j} = \mathrm{Y}_{45}^{j} + \mathrm{X}_{44}^{j}$$

Voor iedere variabele is nu de nieuwe waarde voor het jaar 2005 uitgerekend.

4.4.2 Het voorspellen met de eerste en de tweede principale component

Hier is er voor het voorspellen van de scores van het nieuwe jaar, $\widetilde{Y}_{45}^{j}$, de nieuwe scores S_{45}^{1} en S_{45}^{2}. De nieuwe gestandaardiseerde variabele zal daar een combinatie van moeten bevatten. Zoals in Hoofdstuk 3 is besproken, is $\widetilde{Y}_{45}^{j}$ gelijk aan

$$\widetilde{Y}_{45}^{j} = S_{45}^{1} * L^{1} + S_{45}^{2} * L^{2}$$

Waarbij
L^{1} = loadings van de data op de eerste component
L^{2} = loadings van de data op de tweede component

Op dezelfde manier gaan we terugrekenen door eerst de standaardisering op te heffen met

$$\mathrm{Y}_{45}^{j} = \sigma_{y}^{j} * \widetilde{Y}_{45}^{j} + \overline{Y}^{j}$$

en vervolgens de differencing met

$$\mathrm{X}_{45}^{j} = \mathrm{Y}_{45}^{j} + \mathrm{X}_{44}^{j}$$

Nu zijn weer de nieuwe waarden van de variabelen berekend. Aan de hand van de voorspellingen kunnen conclusies worden getrokken.

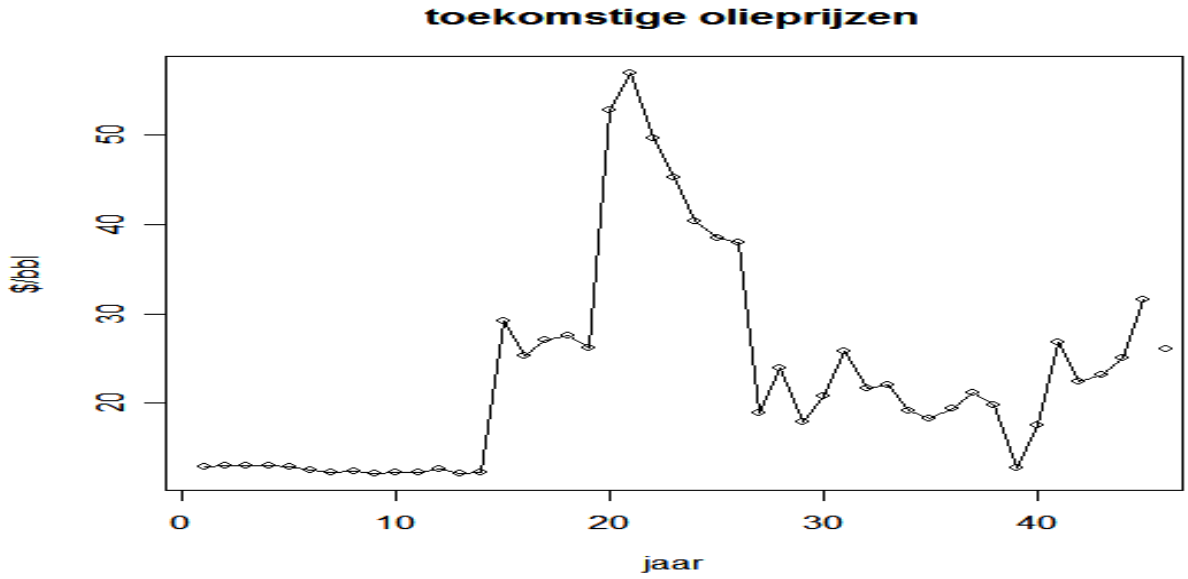
Hoofdstuk 5 Resultaten en Conclusies

Dit onderzoek heeft de olieprijzen en de samenhangende factoren onder de loep genomen. Het idee is om voorspellingen te doen van de olieprijzen, olieproductie en alle andere variabelen die bij dit onderzoek zijn betrokken. Bij het analyseren van deze multivariate data is de aandacht gegaan naar de onderlinge afhankelijkheden tussen de variabelen. Er werd met behulp van de principal component analysis en de tijdreeks analyse gezocht naar een tijdreeks die de meeste informatie in zich bevat over de oorspronkelijke data. Door het benaderen van de tijdreeks kan de voorspelling worden gedaan. De uitspraken die er dan worden gedaan over de toekomstige waarden voor iedere variabele hebben rekening gehouden met de invloed van de andere variabelen/economische factoren

Dit hoofdstuk zal een paar resultaten bespreken en conclusies trekken.

5.1 Resultaten

Het voorspellen van de toekomstige olieprijzen is uiteraard het meest interessant. Volgens dit onderzoek en de gebruikte methoden zal de olieprijs een jaargemiddelde hebben van 25.8 $/bbl (voorspellen met de eerste principale component) of een gemiddelde van 26.1 $/bbl (voorspellen met de eerste en de tweede principale component). Dit houdt in dat de voorspelling met beide manieren geen grote verschillen vertoont. In Figuur 11, is de voorspelling van de olieprijs te zien als het losse punt voor het jaar 2005.

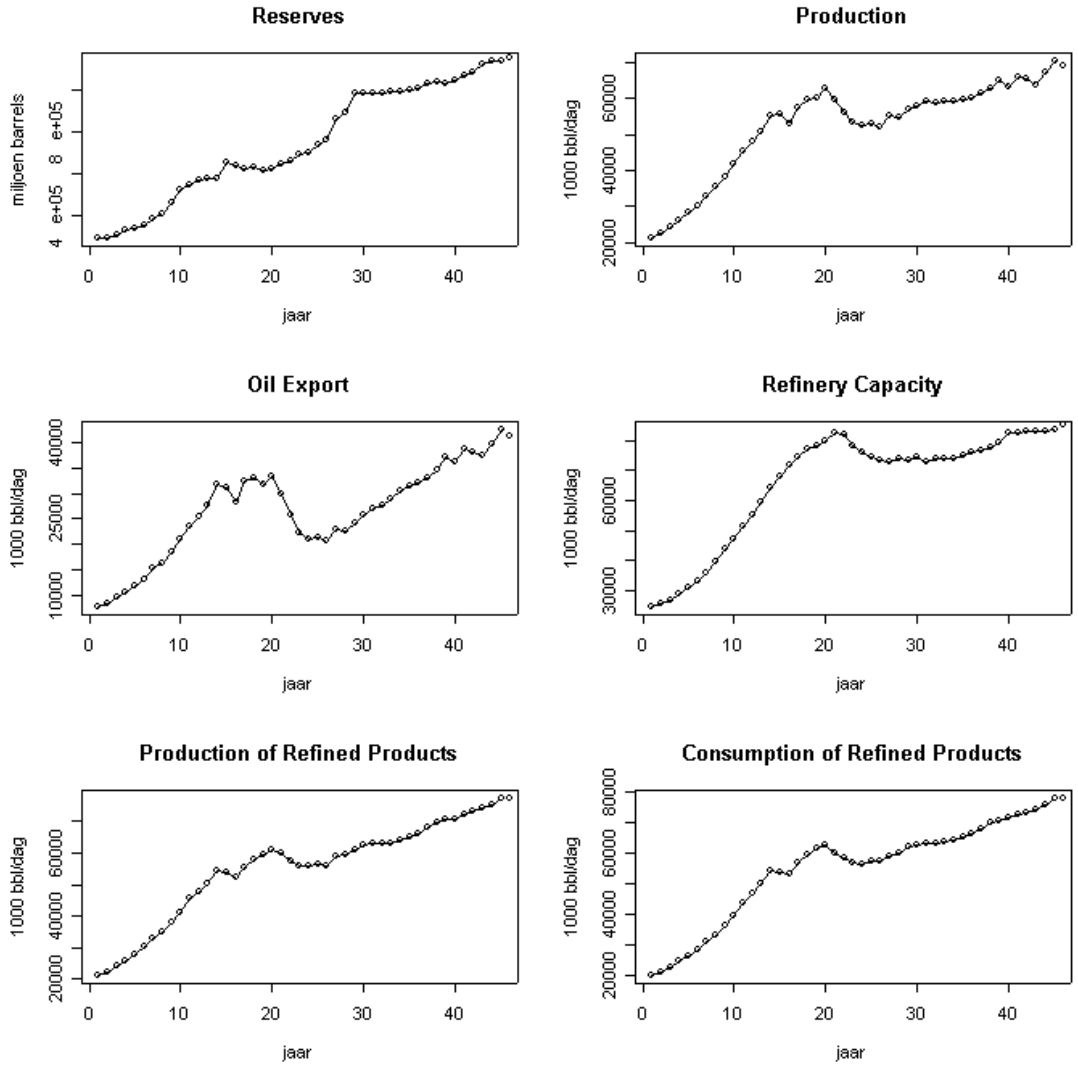


Figuur 11

Voor het voorspellen van alle andere variabelen is op te merken dat de waarden hele kleine verschillen vertonen tussen het voorspellen met alleen de eerste principale component en het voorspellen met de eerste twee principale componenten.

Theoretisch gezien zou het betrekken van de eerste twee componenten betere voorspellingen geven. Het is echter niet het geval bij dit onderzoek. Dit is te wijten aan de lage voorspelde score van de tweede principale component, die vrijwel nihil is. Zo is het voorspellen met beide componenten is zo goed als het voorspellen met alleen de eerste component.

In Figuur 12 zijn alle variabelen geplot met de voorspelde waarden (het losse punt voor het laatste jaar). Deze voorspellingen zijn gedaan met de eerste twee principale componenten.



Figuur 12

De reserves en de raffinaderijencapaciteit geven een kleine stijging aan. De productie en de export geven een kleine daling weer. Zowel de productie als de consumptie van de geraffineerde producten blijven vrijwel gelijk.

Er is een logisch verband te vinden tussen de variabelen. Wanneer de productie laag gaat, zal de export ook lager gaan bijvoorbeeld. En wanneer de consumptie van de geraffineerde producten gelijk blijft dan zal de productie daarvan ook gelijk blijven. Dit verband benadrukt wat er al uit de loadings is te halen in het vorige hoofdstuk.

Na de resultaten te hebben besproken, zullen de conclusies aan bod komen waarmee het onderzoek zal worden afgesloten.

5.2 Conclusies

Bij het onderzoek naar de olieprijzen en de samenhangende factoren en de analyses op de data met R zijn er resultaten gekwantificeerd. Aan de hand van deze resultaten en de analyses is het een en het ander te concluderen

- De voorspelling van de olieprijzen is niet vlekkeloos verlopen. Het is bekend dat de olieprijzen erg hoog waren in het jaar 2005 (het voorspelde jaar) en wel de 60 $/bbl hebben bereikt. De voorspelde waarde van 26 $/bbl ziet er wel redelijk uit omdat gemiddeld gezien de olieprijzen ook in het echt zich zo gedragen. Dit benadrukt de grote invloed die de onverwachte calamiteiten kunnen hebben. De langdurige verwarrende toestand in Irak en de natuurrampen in Amerika in het jaar 2005 zijn daar een voorbeeld van. Deze onverwachte gebeurtenissen zijn niet te voorspellen en moeilijk te kwantificeren en kunnen ervoor zorgen dat de jaargemiddelden fors omhoog gaan.

- Voor betere en nauwkeurigere voorspellingen zou er voor nader onderzoek:
    1. Een uitgebreider onderzoek worden gedaan naar de economische aspecten en het verband tussen de olieprijzen en de andere factoren. Zo kan men meer factoren in het onderzoek betrekken en betere inzicht maken in de samenhang tussen de variabelen.
    2. Een beter tijdreeksmodel bouwen voor de scores. Hiermee kunnen betere voorspellingen worden gedaan met kleinere effect van de ruizen.

- Het voorspellen van de variabelen uit de data had ook op andere manieren gekund. Alle stappen bij de toepassing voor het uitvoeren van de PCA kunnen worden behouden. Het omgaan met de nieuw verkregen scores kan anders.

  Bijvoorbeeld:
    1. door de data op te splitsen in subdelen. Als het eerste gedeelte uit de eerste 20 jaar bestaat, dan zou de PCA op het eerste gedeelte worden uitgevoerd en de nieuwe scores (op de eerste en de tweede principale component) de

21ste scores zou worden. Deze nieuw scores zijn nu het begin van de analyses. Hiermee wordt er een punt van historische simulatie uitgestipt. Met de nieuw verkregen waarden voor de originele data en de rest van de data wordt de PCA uitgevoerd en voorspellingen gedaan. Zo zijn de eerste 20 jaren samengevat en de nadruk is gelegd op de latere jaren die meer veranderingen meemaakten.

2. door de eerste subgedeelte te bekijken (zeg 10 of 15 jaar). Daar wordt de PCA op toegepast. De nieuwe score voorspelt de waarden voor het 16de jaar. Met het nieuwe subgedeelte (11 of 16 jaar) wordt weer de PCA uitgerekend. Met de nieuwe scores voorspellen we weer het nieuwe jaar enzovoorts. We gaan net zo lang stapsgewijs door tot we bij het jaar 2004 aankomen. De nieuwe waarden kunnen we vergelijken met de wereklijke cijfers. Uit zo'n analyse kan men meer zeggen over de nauwkeurigheid van de voorspelling en aan de hand daarvan betere uitspraken maken over het jaar 2005.
3. door bij de voorspelling van de scores naar de boven- en ondergrenzen (betrouwbaarheidsinterval) te kijken. Men kan kijken wat er met de voorspelling gebeurt wanneer de bovengrenzen worden meegenomen om de nieuwe waarden te voorspellen en zo ook met de ondergrenzen.

- Naast het gebruiken van de PCA en de Time Seris Analysis, kan Simulatie ook van toepassing zijn. Simulatie kan worden gebruikt bij het analyseren van de scores als univariate data. Na het opstellen van het model voor de scores op de principale component, kunnen er jaarlijks voor de ruis random trekkingen uit de normale verdeling worden gedaan. Het simuleren (met Crystal Ball) moet uiteraard met 1000 of 10000 trekkingen worden gedaan. Simulatie zal voor de nieuwe score een betere schatting geven van de betrouwbaarheidsinterval en de verwachting om de voorspelde scores te behalen.

- Er zijn geen schrikbarende resultaten terug te vinden. Dit duidt erop dat de methoden die in dit onderzoek zijn gebruikt, wel degelijk zijn voor zulke toepassingen.

- Het loslaten van de tijdreeksanalyse op de olieprijzen om zo voorspellingen te doen is mogelijk. Echter in dit onderzoek is het de bedoeling geweest om voorspellingen te doen rekening houdend met de samenhang met de externe factoren.

- Er bestaat geen beste manier van het doen van voorspellingen. Kijken in de toekomst blijft kijken in een “Crystal ball”

Literatuurlijst

1. C.Chatfield & A.J. Collins
   'Introduction to Multivariate Analysis'
   School of mathematics
   Bath University, 1980

2. David C. Lay
   'Linear Algebra and its Applications'
   University of Maryland, Seceond Eddition

3. M.C.M. de Gunst
   'Statistical Models'
   Vrije Universiteit, najaar 1997

4. Prof. Dr. A.W. van der Vaart
   'Handleiding S-Plus'
   College Statistische Data Analyse
   Vrije Universiteit, najaar 2000

5. Oil and Gas Primer
   'Introduction to the Oil and Gas Business'
   14 May, 2000

6. Integrated Oil/E&P/R&M
   Integrated Oil, E&P, and R&M investment issues
   December 14, 2004, issue 75
   Goldman Sachs

7. Reassessing long-term oil prices
   Finding a new equilibrium
   August 17, 2005
   Goldman Sachs

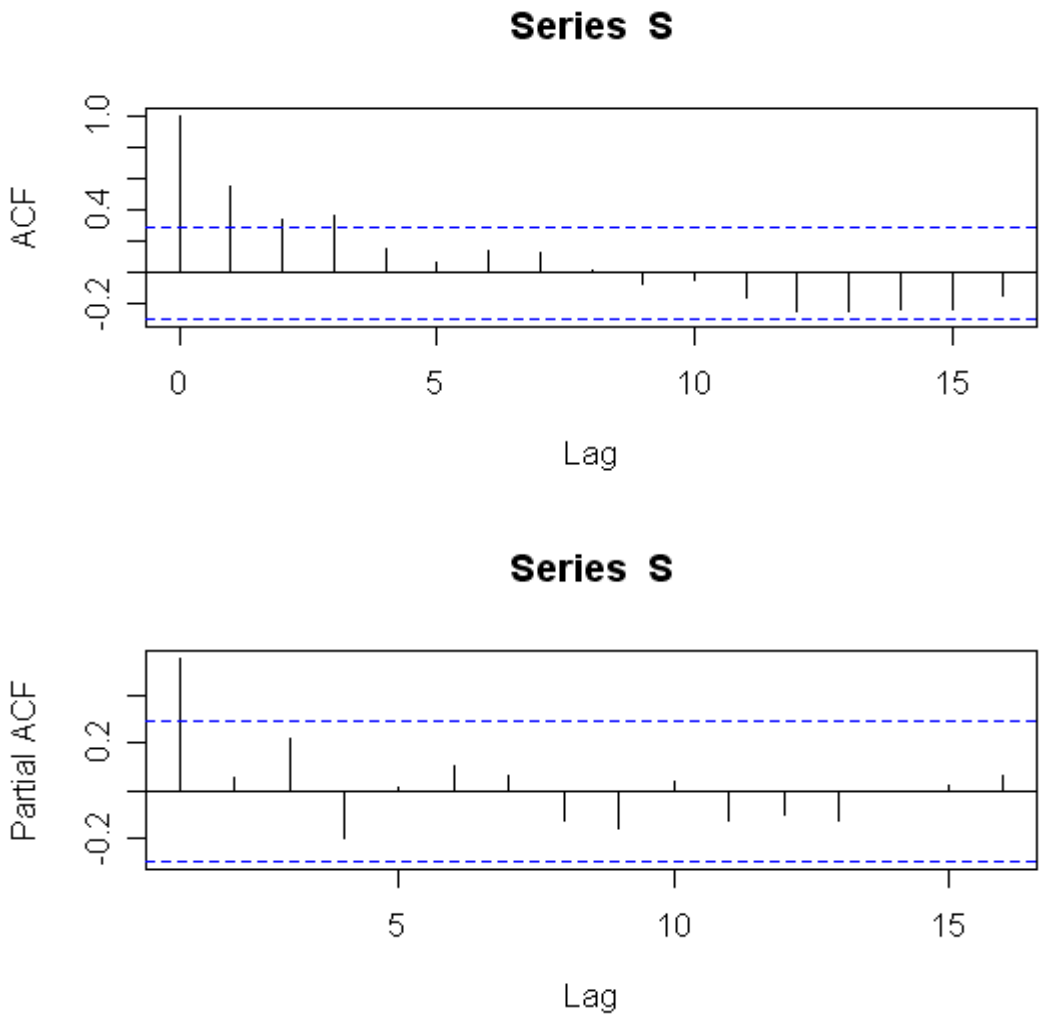
Appendices

I Het kiezen van het geschikte AR model.

Zoals besproken, is de keuze van het AR model aan de gebruiker om het meest geschikte model te kiezen. Er zijn wel ondersteunende methoden die de analyst helpen de juiste keus te maken. Zo kan de autocrrelation functie (ACF) worden getest met de correllogram. Wanneer de correlogram een duidelijk 'cut off' laat zien bij een bepaalde 'lag', dan is het keizen voor een MA(q) process het meest voor de hand liggend.

We kunnen ook de partiele autocorrelation functie bepalen met de partiele correllogram (PACF). Als de partiele correllogram duidelijke 'cut-off ' laat zien op een bepaalde 'lag', is het AR(p) model het meest van toepassing.

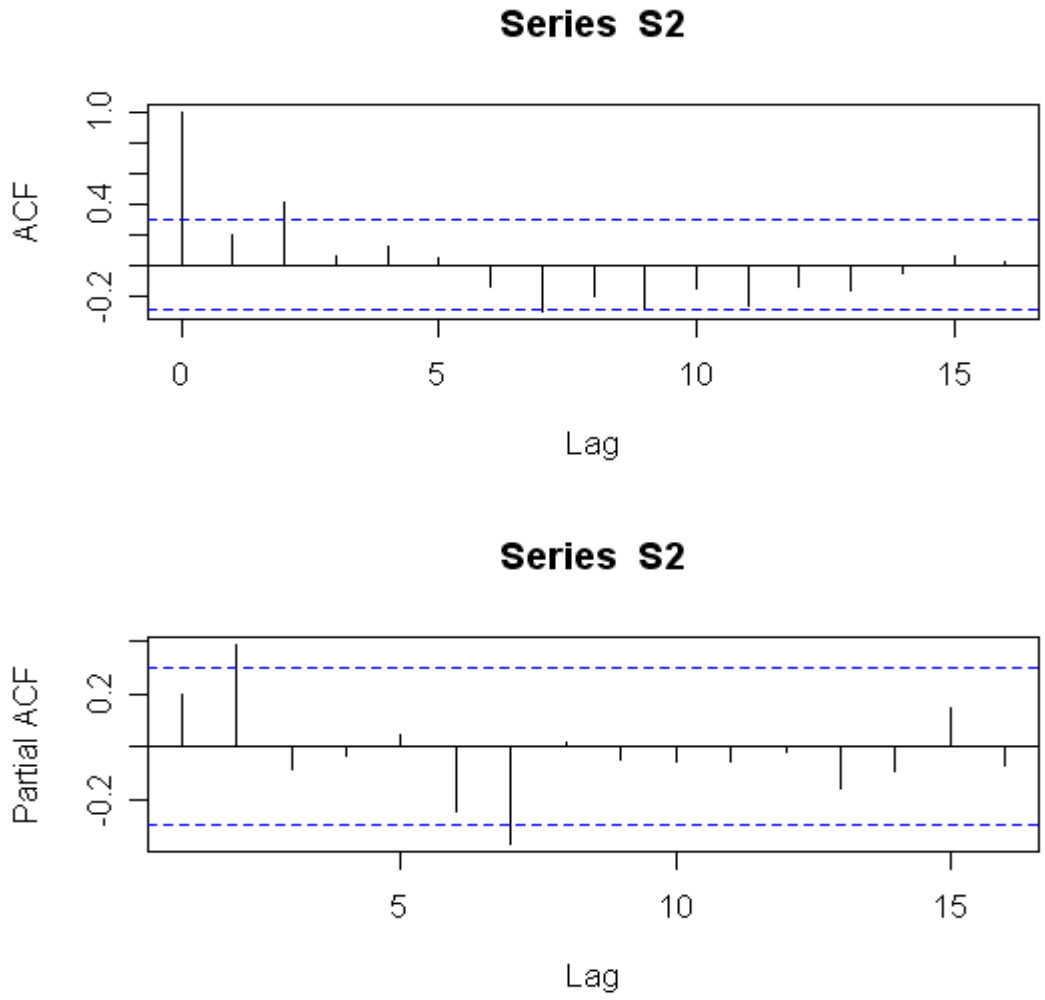
Voor het modelleren van de scores op de eerste principale component zijn de ACF en de PACF hieronder geplot (Figuur 13)



Figuur 13

De logische beredenering om een AR model te kiezen voor onze data, wordt hier nog ondersteunt de PACF. De partiele correllogram laat duidelijk zien dat er een duidelijk cut-off is bij lag 1 (1 staafje steekt boven het grensgebied). Zo is de keus voor een AR(1) model niet zo gek (zie voor de toepassing 4.3).

Op dezelfde manier kijken we voor de score op de tweede component naar de ACF en de PACF. Zie Figuur 14.



Figuur 14

Met dezelfde beredenering gaan we hier te werk. Het MA(q) proces kan wel worden gekozen. Echter nader onderzoek heeft getoond dat bij het keizen van een AR model de ruis een kleinere variantie heeft en zo de benadering van de scores betrouwbaarder kan zijn (zie 4.3).

II R codes

In dit hoofdstuk zal ik de meest interessante codes met de bijbehorende uitvoer in het programma R weergeven die ik tijdens het onderzoek heb gebruikt. Sommige codes zal ik achterwege laten zoals, het aanmaken van vectoren of het differenceren van vectoren, etc. Dit om de lange rijen van getallen te voorkomen en het lezen van het verslag leuk te houden. Voor het gemak zal ik in grote lijnen dezelfde stappen nemen als bij het hoofdstuk

1. Na de 7 vectoren opgesteld te hebben en de differencing (voor iedere vector is er daarom een D als beginletter gebruikt) te hebben uitgevoerd is onze matrix M in het leven geroepen met:

```
>M <- cbind(DOilPrices, DResevers, DProduction, DRefineryCapacity, DOilExports,
DConsumRefinedProducts, DProRefinedProducts)
```

2. Vervolgens het standaardiseren van de vectoren en het uitvoeren van de PCA met een plot van de loadings van data op de eerste en de tweede component (Figuur 6).

```
>pcM <- princomp(M, cor = TRUE)
>biplot(pcM)
```

3. Hier wordt de data benaderd met de eerste principale component. Allereerst worden de scores gemodelleerd om daarna een voorspelling te maken van de nieuwe scores. Uiteindelijk wordt er teruggerekend naar de originele waarden van de data.

```
>S <- pcM$scores[,1]
>par(mfrow =c(2,1))    # Figuur 13
>acf(S)
>pacf(S)
>fit <- ar(S, orde.max= 1)          # het fitten met een AR(1) model
>fit
Call:
ar(x = S, orde.max = 1)
Coefficients:
  1
0.5581
```

```
Order selected 1  sigma^2 estimated as  2.810

>s45 <- predict(fit, S, n.ahead = 1, se.fit = TRUE)       # de nieuwe score voorspellen
>s45
$pred
Time Series:
Start = 45
End = 45
Frequency = 1
[1] -0.9612759
$se
Time Series:
Start = 45
End = 45
Frequency = 1
[1] 1.676195
comp1 <- (loadings(pcM)[,1])
>y45 <- s45$pred[1]*comp1    # het terugrekenen om de voorgespelde waarde van de
                             olieprijzen uit te rekenen
>yOilPrices <- (sd(DOilPrices)*y45[1]) + mean(DOilPrices)
>geschatteOilPrijs <- yOilPrices+oilPrices[44]
>geschatteOilPrijs
>DOilPrices
 25.80030
```

En zo worden alle voorspellingen van de andere variabelen gedaan.

4. Hier wordt de data benaderd met de eerste en de tweede principale component. Ik zal de uitvoer achterwege laten.

```
>S2 <- pcM$scores[,2]
> par(mfrow =c(2,1))    # Figuur 15
>acf(S2)
>pacf(S2)
>fit2 <- ar(S2, orde.max= 2)        #het fitten met een AR(2) model

>s45.2 <- predict(fit2, S2, n.ahead = 1, se.fit = TRUE)
> comp2 <- (loadings(pcM)[,2])    # het terugrekenen om de voorgespelde waarde van de
                                  olieprijzen uit te rekenen
```

```
>y45.2 <- s45$pred[1]*comp1 + s45.2$pred[1]*comp2
>yOilPrices2 <- (sd(DOilPrices)*y45.2[1]) + mean(DOilPrices)
>geschatteOilPrices2 <- yOilPrices2 + oilPrices[44]
>geschatteOilPrices2
>DOilPrices
 26.01093
```

Alle anderen voorspelde waarden worden op dezelfde wijze uitgerekend.