Verslag Data Analysis

Darius Barsony, Ellen Bogaards, Joshua de Roos en Roos Vervelde

29 juni 2018

UNIVERSITEIT VAN AMSTERDAM

Vak

Data Analysis and Visualization

Docenten

Gosia Migut en Nick de Wolf

Inhoudsopgave

1 Inleiding

Er zijn allerlei factoren die de hoogte en verandering van voedselprijzen beïnvloeden (Abbott, Hurt, Tyner et al., 2009). Ook krantenberichten lijken dit beeld te bevestigen. Zo schrijft De Telegraaf (2018) dat de voedselprijzen in 2018 hard stijgen vanwege het weer en ziektes en stelt Business Insider (2018) dat ook brandstofkosten de voedselprijzen beïnvloeden. Daarnaast bericht ook United Nations (2017) dat voedselprijzen in Syrië, waar veel politieke onrust heerst en binnenlandse vluchtelingen zijn, verdubbeld zijn. Er zijn dus veel verschillende omstandigheden die invloed hebben op de hoogte van de voedselprijs.

Het is echter de vraag of deze verschillende factoren in elk land of gebied even sterk van toepassing zijn. In sommige landen bevinden zich misschien wel nauwelijks vluchtelingen maar stijgt de voedselprijs alsnog. Hoewel eerdere onderzoeken dus al verschillende factoren op voedselprijzen in kaart hebben gebracht (Abbott et al., 2009; Timmer, 2008), is het niet vanzelfsprekend dat voor alle landen elke factor even veel bijdraagt aan de hoogte van de voedselprijs.

Dit onderzoek richt zich daarom op de vraag: "Wat zijn de verbanden tussen voedselprijzen van verschillende producten met elkaar en met data over de populatie, vluchtelingen, BBP en regenval?". Het antwoord op deze vraag laat zien welke factoren het meeste bepalend zijn voor de voedselprijs in welke gebieden. Dat geeft dus inzicht in de verschillen tussen de genoemde gebieden maar geeft ook antwoord op de vraag of alle factoren een even grote rol spelen in alle gebieden. Dat laatste kan ook belangrijk zijn voor beleidsbepaling. Als een land namelijk maatregelen wil nemen voor hoge voedselprijzen, vereist dat wel dat de regering inzicht heeft in welke factoren voornamelijk bijdragen aan die hoge prijzen. Het antwoord op de hoofdvraag is dus zowel voor de wetenschap als de maatschappij relevant.

Om een zo compleet mogelijk antwoord te vinden op de hoofdvraag, is deze onderverdeeld in drie deelvragen:

- In hoeverre is er een correlatie tussen prijzen van verschillende goederen in hetzelfde land?
- In hoeverre vertonen landen in dezelfde regio overeenkomstige hoogte en correlaties op het gebied van voedselprijzen?
- In hoeverre hebben andere factoren invloed op de voedselprijzen in een land? Deze vraag is verder onderverdeeld in de vragen:
  - In hoeverre heeft het BBP van een land invloed op de voedselprijzen?
  - In hoeverre heeft de populatie van een land invloed op de voedselprijzen?

- In hoeverre heeft de jaarlijkse neerslag in een land invloed op de voedselprijzen?
- In hoeverre heeft het aantal opgevangen vluchtelingen in een land invloed op de voedselprijzen?

De eerste deelvraag richt zich dus vooral op verbanden tussen verschillende goederen in hetzelfde land. De hypothese hierbij is dat prijzen van verschillende goederen gecorreleerd zijn wanneer de goederen direct afhankelijk van elkaar zijn. Zo valt te verwachten dat de prijzen van goederen als melk en boter positief met elkaar correleren omdat boter wordt gemaakt van melk. Als de melk dus duurder wordt, is het aannemelijk dat boter ook duurder wordt.

Als er zulke correlaties zijn gevonden, is het ook interessant om te onderzoeken of die correlaties ook optreden in buurlanden. Landen in dezelfde regio hebben immers te maken met vergelijkbare omgevingsfactoren als klimaat of handelspartners. Daarom zal er bij de tweede deelvraag worden onderzocht of patronen die zich in een land voordoen ook te herkennen zijn in omliggende landen. Het doel hierbij is om zowel binnen regio's overeenkomsten te ontdekken als tussen regio's verschillen te laten zien. Hierbij is de hypothese dat landen uit dezelfde regio even hoge prijzen en dezelfde correlaties vertonen omdat omgevingsfactoren voor deze landen vergelijkbaar zijn. Het onderzoek zal zich bij deze deelvraag beperken tot de regio's West-Afrika, Oost-Afrika, het Midden-Oosten en Zuid-Azië omdat hier relatief veel data van is.

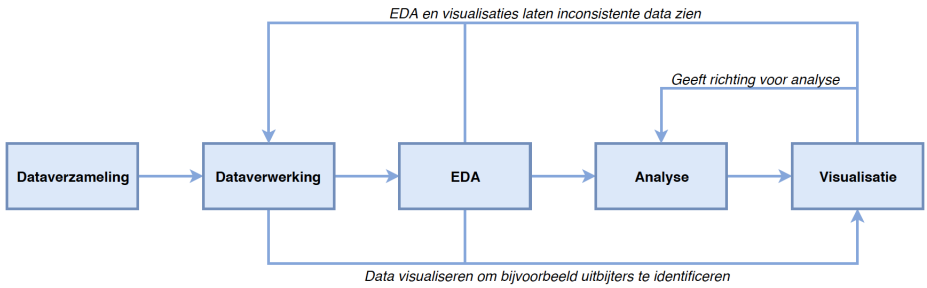
Een stille aanname van de vorige twee deelvragen is dat er omgevingsfactoren zijn die de voedselprijs beïnvloeden. De derde deelvraag onderzoekt of deze aanname waar is voor de factoren BBP, jaarlijkse neerslag, opgevangen vluchtelingen en populatie van een land. Deze deelvraag is dan ook onderverdeeld in vier aparte deelvragen om de resultaten overzichtelijk te houden. De antwoorden op deze vragen bieden dan mogelijke verklaringen voor uitkomsten uit de eerste en tweede deelvraag. De hypothese bij deze laatste deelvraag is dat de genoemde factoren een significante invloed hebben maar dat het per land verschilt welke van de factoren het meeste invloed heeft. Sommige landen hebben immers te kampen met grote getalen aan vluchtelingen en anderen niet.

De gebruikte data voor het onderzoeken van deze hypothesen is afkomstig van de Global Food Prices Database. Hoe deze data verder is verwerkt en welke analysemethoden er zijn gebruikt, komt in de volgende paragraaf aan bod. Daarna volgen de resultaten van de analyse voor de verschillende deelvragen. Het verslag eindigt met een conclusie die antwoord geeft op de hoofdvraag en kritisch reflecteert op de gebruikte methoden.

2 Methode

Voordat het mogelijk was om de resultaten te behalen die in de volgende paragraaf aan bod komen, is de data op verschillende manieren verwerkt, geanalyseerd en gevisualiseerd. Dit proces is grofweg in te delen in vijf stappen: dataverzameling, dataverwerking, Exploratory Data Analysis (EDA), analyse en visualisatie.

Hoewel het proces in deze paragraaf in deze volgorde wordt uitgelegd, moet benadrukt worden dat deze volgorde geen chronologische volgorde vertegenwoordigt. De genoemde stappen zijn iteratief doorlopen. Het daadwerkelijke proces is samengevat in figuur 1:



Figuur 1: Schematische visualisatie van het proces van het onderzoek

2.1 Dataverzameling

Om de Global Food Prices (WFP) te relateren aan externe factoren als populatie en Bruto Binnenlands Product (BBP) is eerst extra data verzameld. Uiteindelijk zijn naast de WFP dataset nog zes extra datasets uitgekozen om de analyse te versterken:

- Dataset met wisselkoersen voor alle valuta's van alle landen van 1992 tot 2017 (The World Bank, 2017b)
- Dataset met alle ISO-3 codes voor alle landen (Lukes, 2018)
- Dataset met maandelijkse neerslag voor alle landen van 1992 tot 2017 (Climate Change Knowledge Portal, z.d.)
- Dataset met jaarlijkse BBP van alle landen van 1992 tot 2017 (The World Bank, 2017a)
- Dataset met de populatie van alle landen van 1992 tot 2017 (The World Bank, 2017c)

- Dataset met aantal inkomende vluchtelingen voor veel landen van 1962 tot 2017 (The World Bank, 2017d)

De eerste dataset is gebruikt voor het omrekenen van alle prijzen in andere munteenheden naar prijzen in US dollar. Dat was nodig zodat in de analyse alle prijzen in dezelfde munteenheid stonden en de prijzen met elkaar vergeleken konden worden.

De tweede dataset is gebruikt om in de datasets waar landen alleen met hun code (zoals NED voor Nederland) werden aangegeven deze codes te converteren naar de namen van de landen. Dat maakte het latere proces van data uit al deze datasets matchen makkelijker.

Drie andere datasets zijn uitgekozen na een klein vooronderzoek in wetenschappelijke literatuur en nieuwsartikelen. Zo stelt Timmer (2008) dat droogte een grote invloed hebben op voedselprijzen. Naar aanleiding hiervan is ervoor gekozen om een dataset over de neerslag te zoeken. De dataset van het aantal inkomende vluchtelingen is gekozen omdat in Syrië voedselprijzen zijn verdubbeld sinds de oorlog is begonnen (United Nations, 2017). De dataset van het jaarlijks BBP van alle landen is uitgezocht omdat blijkt dat mensen uit rijkere landen meer geld uitgeven aan voedsel (Roser & Ritchie, 2018).

De laatste dataset met alle populaties van alle landen is ook in dit onderzoek meegenomen omdat het aannemelijk leek dat een grotere populatie leidt tot meer vraag naar voedsel en dus de prijzen omhoog drijft.

Nadat al deze datasets zijn verzameld, diende de data eerst opgeschoond, aangevuld en genormaliseerd te worden. De volgende subparagraaf geeft een samenvatting van dat proces.

2.2 Dataverwerking

Om uiteindelijk de data op een zinvolle manier te kunnen vergelijken en te kunnen matchen, is alle data zoveel mogelijk aangevuld en in hetzelfde format gezet. Deze paragraaf legt eerst uit hoe de verwerking van de WFP-dataset verliep en daarna hoe de andere datasets zijn verwerkt.

2.2.1 WFP dataset

De WFP dataset bestaat uit ruim 700.000 maandelijkse metingen van prijzen van verschillende producten van meer dan 1500 markten verdeeld over 76 landen. Al deze prijzen staan in de munteenheid van het land waar de meting is gedaan. Verder hebben vergelijkbare producten vaak niet dezelfde naam. Zo bestaat er een productcategorie genaamd 'Rice (low quality)' en een categorie genaamd 'Rice'. Al deze discrepanties zijn door middel van een iteratief proces genormaliseerd. Eerst is echter alle ontbrekende data aangevuld.

De enige ontbrekende data waren lege invullingen bij de regio en marktnaam van prijzen die het nationaal gemiddelde vertegenwoordigden. Daarom is op alle plekken waar de prijzen het nationaal gemiddelde vertegenwoordigden ook bij de regio en markt 'National Average' ingevuld.

Vervolgens zijn vergelijkbare productcategorieën met ongeveer dezelfde naam samengevoegd tot één productcategorie. Zo zijn alle verschillende rijstsoorten samengevoegd onder de productnaam 'Rice' en alle soorten vlees onder één noemer 'Meat' gebracht. Alle prijzen in het bestand waren na deze ingreep verdeeld over 40 categorieën.

Na het normaliseren van alle productcategorieën zijn ook alle prijzen omgerekend naar US dollar (USD). De dataset met alle historische wisselkoersen van alle landen diende als basis voor deze conversie. Aangezien de prijzen in de WFP database per maand gemeten zijn en in de dataset met wisselkoersen één wisselkoers voor het hele jaar staat, zijn alle prijzen van dat jaar, ongeacht de maand, met die wisselkoers omgerekend. Voor de zekerheid zijn de prijzen in de oude munteenheden ook nog bewaard omdat deze later misschien nodig waren.

Als laatste zijn ook alle eenheden genormaliseerd en de bijbehorende prijzen daarop aangepast. Met behulp van online conversietools en omrekenfactoren is voor elke productcategorie geprobeerd alle metingen om te rekenen naar dezelfde eenheid. Zo zijn bijvoorbeeld alle prijzen van suiker omgerekend naar de kiloprijs. Alleen van de producten tee, koffie, plantanen en yams was het niet mogelijk om alle prijzen om te rekenen naar één eenheid. Sommige metingen van deze producten stonden namelijk in eenheden die niet om te rekenen waren naar gewicht (bijvoorbeeld 'Unit' en 'Package'). Deze producten zijn daarom uitgesloten van de analyse.

Dit hele proces, dat nog is herhaald na het visualiseren van data, leverde uiteindelijk een complete dataset op waarin van 40 producten de prijs in dezelfde eenheid en in dollar stond gegeven. Deze dataset is vervolgens geëxporteerd naar een nieuw csv-bestand die gebruikt kon worden voor de analyse.

2.2.2 Overige datasets

Ook de overige datasets benodigden eerst enige bewerking voordat de data gebruikt kon worden in de analyse. Deze subparagraaf bespreekt kort welke bewerkingen zijn uitgevoerd.

Ten eerste zijn met behulp van de dataset met ISO-3 codes bij alle datasets waar de landen alleen met hun code stonden aangegeven deze codes geconverteerd naar de namen zoals die in de WFP dataset staan. Daarnaast zijn ook bij de datasets waar de namen van de landen afweken van de WFP dataset, deze afwijkingen gecorrigeerd. Hierna stonden in alle datasets dus de namen van de landen zoals die ook in de WFP dataset staan.

Ten tweede is bij alle datasets geprobeerd de missende data zoveel mogelijk op te vullen met

data uit andere internetbronnen. Zo ontbrak in de dataset van de BBP's alle waarden voor 2017. Deze zijn vervolgens aangevuld met data van andere bestanden die in het procesboek genoemd staan. Als er geen data beschikbaar was voor bepaalde jaren is waar mogelijk het gemiddelde genomen van het jaar ervoor en daarna. Als ook dat niet mogelijk was, is de onbekende waarde niet aangevuld. Vooral voor de dataset van inkomende vluchtelingen was het lastig om aanvullende data te vinden. Daarom ontbraken bij de analyse in deze dataset voor een aantal landen veel waarden. Daarnaast is ook de dataset van alle wisselkoersen niet compleet omdat van sommige munteenheden niet genoeg historische data was. Daarom zijn de landen Mauritanië en Somalië uitgesloten van de analyse. Toch zijn de meeste datasets compleet aangevuld met data uit andere bronnen.

Ten derde zijn in het kader van efficiëntie alle landen die niet voorkwamen in de WFP dataset verwijderd uit de overige datasets.

Uiteindelijk zijn alle datasets geëxporteerd naar nieuwe csv-bestanden. Deze bestanden zijn vervolgens gebruikt bij de Exploratory Data Analysis die in de volgende paragraaf aan bod komt.

2.3 Exploratory Data Analysis

Een Exploratory Data Analysis (EDA) dient voornamelijk om de eerste patronen in data te ontdekken om zo te weten wat het beste is om op te focussen in de diepere analyse. Er zijn vier manieren om deze EDA uit te voeren: grafisch en niet-grafisch voor zowel een enkele variabele (univariaat) als meerdere variabelen (multivariaat). In dit onderzoek is alleen de niet-grafische, univariate EDA nauwelijks aan bod gekomen omdat aangenomen is dat prijzen altijd relatief zijn en dus geïnterpreteerd moeten worden ten opzichte van andere prijzen. De andere drie manieren van analyse zijn wel uitgebreid uitgevoerd en worden in deze paragraaf kort samengevat.

Ten eerste zijn er voor veel landen boxplots gemaakt met alle metingen van een bepaald product in die landen. Deze univariate, grafische analyse laat goed de spreiding van de data zien en maakt het ook mogelijk om de prijzen van verschillende landen visueel met elkaar te vergelijken. Aan de hand van deze boxplots is vervolgens gekeken welke landen en regio's overeenkomstige prijzen vertoonden.

Daarnaast is ook voor veel combinaties van goederen en landen de Pearson correlatiecoëfficiënt uitgerekend met behulp van de Numpy library van Python. De Pearson correlatiecoëfficiënt is een gestandaardiseerde coëfficiënt tussen -1 en 1 die uitdrukt in hoeverre twee series metingen gecorreleerd zijn. De coëfficiënt wordt als volgt berekend waarbij $cov(x, y)$ de covariantie van x en y is en σ_x en σ_y de standaarddeviaties van beide dataseries:

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Wanneer de coëfficiënt de -1 benadert, is de correlatie sterk negatief. Bij een waarde van 0 is er geen correlatie en bij een waarde van 1 is er een sterk positieve correlatie.

Aangezien in deze fase van het onderzoek nog zo min mogelijk aannames gedaan konden worden, is dit voor zoveel mogelijk combinaties gedaan. Daarna zijn alle correlatiecoëfficiënten van meer dan 0.75 of minder dan -0.75 en meer dan 40 datapunten naar nieuwe csv-bestanden geschreven om later verder gebruikt te kunnen worden. Uiteindelijk zijn voor de volgende correlaties uitgerekend:

Tabel 1: Overzicht van alle uitgerekende Pearson correlaties

Correlaties van alle combinaties van twee goederen per land
Correlaties van alle combinaties van twee goederen van alle landen bij elkaar
Correlaties van het BBP per jaar en rijstprijzen per jaar van alle landen
Correlaties van de jaarlijkse neerslag per land met alle individuele goederen in dat land
Correlaties van de bevolking per jaar met alle individuele goederen in een bepaald land
Correlaties van inkomende vluchtelingen per jaar met graanprijzen per jaar van alle landen

Een aantal van deze correlaties zijn vervolgens gevisualiseerd in een scattermatrix waarin van één land van vier goederen scatterplots, de bijbehorende correlatiecoëfficiënt en distributie van alle metingen zijn weergegeven. Deze matrix is daarmee een combinatie van grafische en niet-grafische analyse van twee variabelen en een grafische, univariate analyse. In het gedeelte van de matrix boven de diagonaal zijn namelijk de scatterplots weergegeven, wat valt onder grafische multivariate analyse. Op de diagonaal is de distributie van alle metingen weergegeven, een grafische univariate analyse. Op de onderste helft van de scattermatrix staan alle correlatiecoëfficiënten, een niet-grafische multivariate analyse.

Al deze verschillende correlaties en visualisaties gaven richtingen voor de diepgaandere analyse die in de volgende subparagraaf beschreven.

2.4 Analyse

Om uiteindelijk de juist resultaten te presenteren en te visualiseren is de data die volgens de correlaties van de EDA veelbelovend was, verder geanalyseerd. Daarnaast is ook de dataset als geheel dieper geanalyseerd om overkoepelende verbanden te kunnen ontdekken. Dit is gedaan met classificatie volgens het KMeans algoritme en lineaire regressie. Beide methoden en de toepassing ervan in dit onderzoek worden in deze paragraaf uitgelegd.

2.4.1 Classificatie

Classificatie is een manier om in een bepaalde dataset clusters te ontdekken of te voorspellen of een waarde bij een bepaald cluster hoort. Classificatie is daarmee dus een methode om te kijken of de beschikbare data op een bepaalde manier te categoriseren is.

Dit onderzoek heeft gebruik gemaakt van het KMeans algoritme uit de Python scikit-learn (SKLearn) library om te kijken of er in de beschikbare data bepaalde clustering te zien was. Dit algoritme gaat op zoek naar de best mogelijke manier om de data in N aantal clusters te verdelen. Het aantal clusters waar naar gezocht moet worden, moet de gebruiker van tevoren meegeven. Als dat gedaan is, vindt het algoritme in drie stappen de beste manier om de data in dit N aantal clusters in te delen.

Eerst worden vrij willekeurig N centroïdes in de data geplaatst. Dit zijn 'verzonnen' datapunten die het midden vormen van een cluster. Vervolgens koppelt het algoritme elk datapunt aan de dichtstbijzijnde centroïde. Als laatste berekent het algoritme van alle punten die aan een centroïde zijn gekoppeld het gemiddelde en maakt van dat gemiddelde de nieuwe centroïde. Die laatste twee stappen worden herhaald totdat de centroïdes niet meer significant verschuiven.

Na het doorlopen van het algoritme behoort elk datapunt tot een bepaald centroïde en daarmee tot een bepaald cluster. Een metriek genaamd de 'inertie' drukt vervolgens uit hoe goed de clustering is. Deze inertie wordt berekend door de afstanden van alle datapunten tot de centroïde te kwadrateren en bij elkaar op te tellen. Hier hoort de volgende formule bij:

$$\sum_{i=0}^{n} \underset{\mu_j \epsilon C}{min}(||x_j - \mu_i||^2)$$

Er geldt dat hoe kleiner de inertie, hoe coherenter de datapunten en hoe beter de clustering.

Dit onderzoek heeft dit KMeans algoritme toegepast op een dataframe met daarin alle beschikbare voedselprijzen, BBP-gegevens, neerslagcijfers en populatiedata. Per meting in de WFP data is gekeken wat het BBP, de neerslag en de populatie in die maand of jaar was en dat is vervolgens aan die meting gekoppeld. De vluchtelingendata zijn hierbij niet meegenomen omdat er voor veel landen weinig cijfers hierover bekend zijn en veel datapunten dan dus onbruikbaar geweest zouden zijn.

Op al deze data is vervolgens het KMeans algoritme losgelaten om te kijken of er in de data een clustering te ontdekken was. Deze clustering is vervolgens gevisualiseerd met TSNE van de SKLearn library. Dit is een manier om hoogdimensionale data weer te geven in twee dimensies. Naast KMeans is ook lineaire regressie toegepast voor de diepgaandere analyse. Dit wordt in de volgende subparagraaf besproken.

2.4.2 Regressie

Naast clustering is ook lineaire regressie een manier om data diepgaander te analyseren. Bij lineaire regressie berekent het algoritme op basis van bestaande numerieke datapunten de lijn die het beste bij de lijn past. Deze lijn vangt de datapunten dus in een soort functie om zo meetpunten te kunnen voorspellen.

De Mean Sqaure Error (MSE) is, net als de inertie bij KMeans clustering, vervolgens een maat van hoe goed de lijn past bij de datapunten. Net als de inertie wordt de MSE ook berekend door alle afstanden van de punten tot de lijn te kwadrateren, bij elkaar op te tellen en vervolgens te delen door het aantal punten.

Dit onderzoek heeft gebruik gemaakt van het lineaire regressie algoritme van de SKLearn library van Python om te kijken in hoeverre er significante relaties zijn tussen voedselprijzen en het BBP, de populatie en neerslagcijfers. Deze multivariate regressie gaf vervolgens voor elk van de factoren aan hoeveel de voedselprijs verandert als die bepaalde factor verandert.

Naast lineaire regressie op de complete dataset, is er ook lineaire regressie gebruikt voor de eerder gevonden correlaties tussen verschillende goederen. Dit is gedaan door in alle scatterplots ook de regressielijn te plotten en de daarbij berekende MSE te printen. Hierdoor ontstond er een goed beeld van hoe sterk het verband daadwerkelijk is en of het verband werkelijk lineair is.

Na deze diepgaandere analyse op zowel globaal als lokaal niveau zijn de resultaten zo goed mogelijk gevisualiseerd. Hoe dit is gedaan, komt in de volgende paragraaf aan bod.

2.5 Visualisatie

Bij het visualiseren van de data en resultaten is gezocht naar een balans tussen overzicht en diepgang. Aan de ene kant moet de data namelijk zo gevisualiseerd worden dat snel te zien is waar het over gaat en wat de patronen zijn. Maar aan de andere kant moet de data niet te simpel worden weergegeven en moeten fgiuren genoeg informatie bevatten om een compleet beeld te geven van de datasets. Met dit in het achterhoofd zijn verschillende visualisaties gemaakt (zie figuur 2).

Dit is gedaan met de Bokeh library van Python. Deze library stelt de onderzoeker in staat interactieve visualisaties te maken en deze naar een html-bestand te exporteren om te gebruiken op bijvoorbeeld een website.

Tabel 2: Gemaakte visualisaties met Bokeh

Choropleth map
Boxplot
Scattermatrix
Barchart
Bubble chart
Scatterplot

Deze visualisaties komen in de volgende paragraaf naar voren en zijn ook gebruikt op een website die gemaakt is voor dit onderzoek [1].

2.6 Software en Hardware

Voor het onderzoek is een project aangemaakt op GitHub. Vervolgens is alle code geschreven in Microsoft Visual Studio Code. Deze codes zijn gerund op verschillende computers met diverse specificaties. De resultaten hiervan zijn te zien in de volgende paragraaf.

[1] https://joshroos.github.io/DataAnalysis-groepje17/

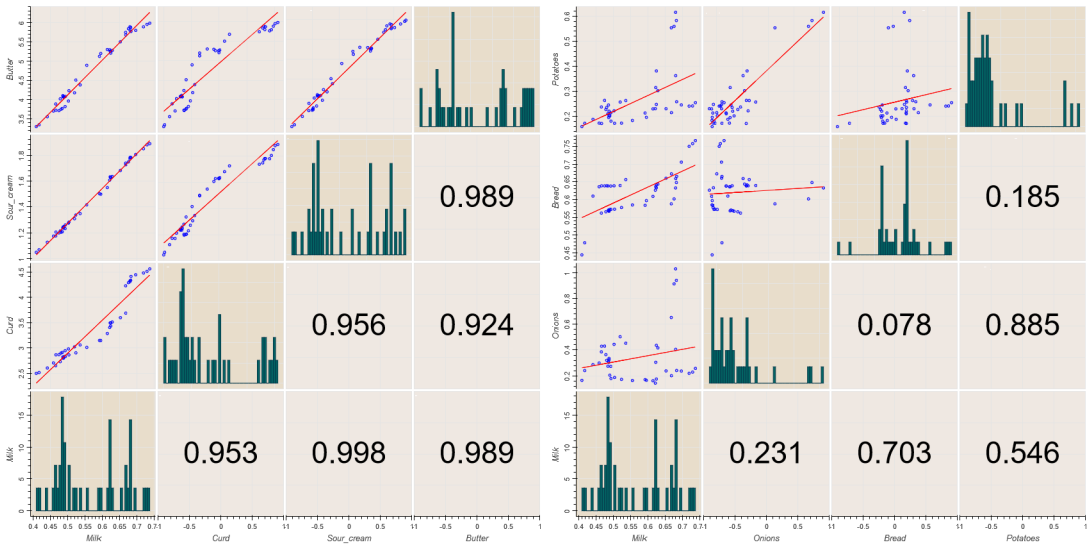
3 Resultaten

Om de hoofd- en deelvragen te beantwoorden zijn verschillende visualisaties gemaakt, afhankelijk van het te onderzoeken verband. Deze visualisaties en overige resultaten worden in deze paragraaf gepresenteerd. Dit zal gedaan worden door per deelvraag de resultaten en visualisaties te bespreken.

3.1 Correlaties tussen prijzen van goederen

Veel verschillende soorten goederen hebben op een bepaald vlak met elkaar te maken. Ze hebben bijvoorbeeld een vergelijkbare productie, zijn gemaakt uit dezelfde grondstoffen, of dienen hetzelfde doel. Hierdoor is het aannemelijk dat de prijzen van deze goederen correleren. Om te kijken of dit inderdaad geldt voor bepaalde goederen, is de Pearson-correlatie uitgerekend voor combinaties van goederen wereldwijd voor de jaren 1992 tot 2017. Hieruit blijkt dat de gemiddelde prijzen van veel producten in de wereld een hoge correlatie laten zien. Dit zijn voornamelijk positieve correlaties. Van de 110 combinaties van goederen met een relevante correlatie ($0.75 <$ correlatiecoëfficiënt $<$ -0.75), zijn er 26 negatief. Een aantal interessante correlaties zijn te vinden tussen aubergine en courgette (0.9671), bloem en mais (0.8164), vee en yoghurt (-0.8519) en vee en kaas (-0.8818).

In plaats van te kijken naar de correlatie tussen goederen van alle landen gedurende alle jaren, kan er ook gekeken worden naar de correlatie tussen goederen binnen één land. Een opmerkelijk voorbeeld hiervan is te vinden in Oekraïne. Vooral de correlaties tussen diverse melkproducten zijn opvallend en zijn zichtbaar in de eerste grafiek in figuur 2. De andere grafiek, laat zien dat er ook correlaties zijn tussen andere goederen, maar dat deze niet zo hoog is als tussen de melkproducten. De MSE's van deze scattermatrices zijn te vinden in tabel 3. Ook hier zijn de MSE's vrij laag.

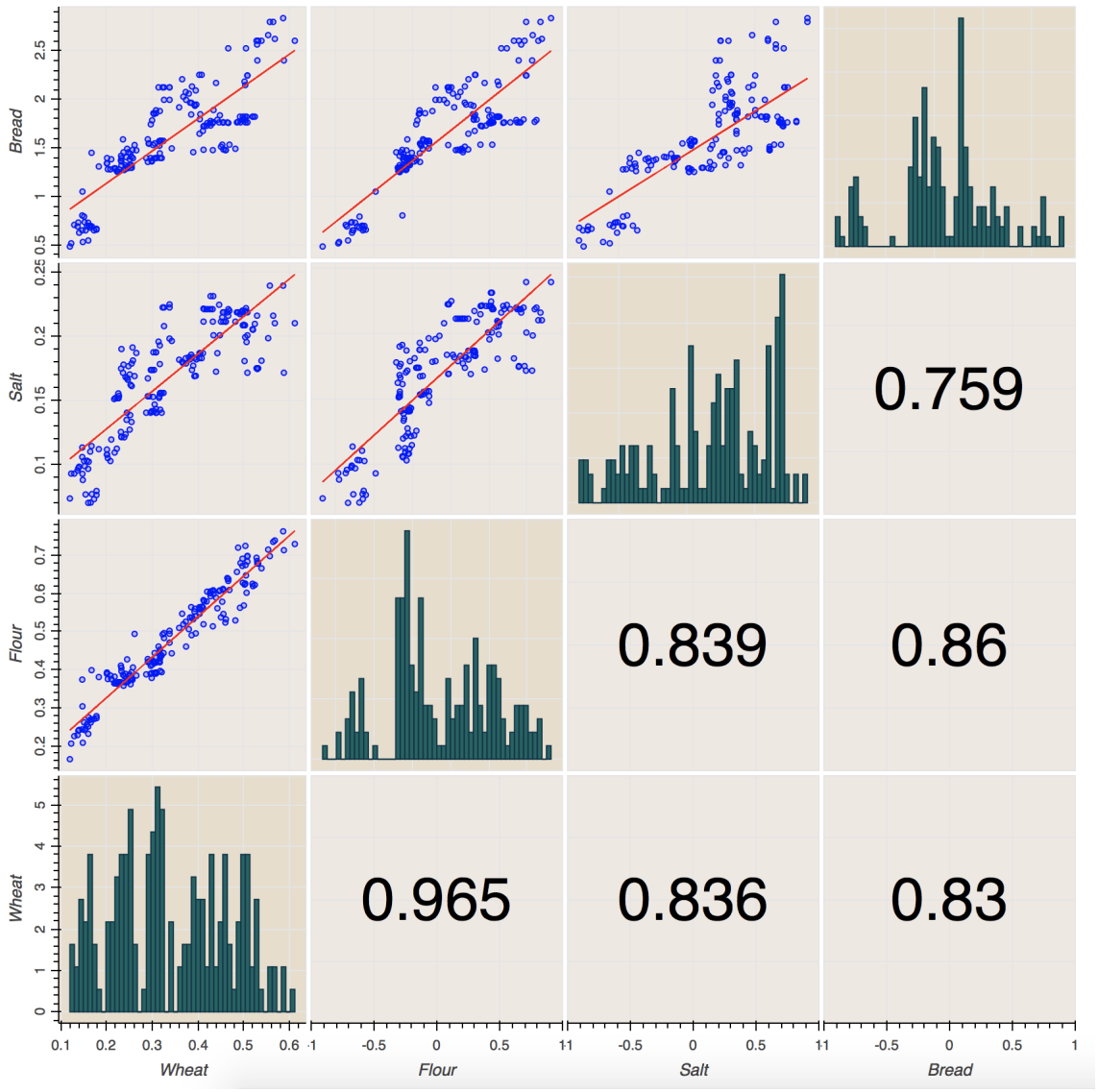


Figuur 2: Scattermatrix van melk, curd, zure room en boter (links) en de scattermatrix van melk, uien, brood en aardappelen in Oekraïne (rechts)

Tabel 3: Mean square error scattermatrix Oekraïne

Goederen combinatie	**MSE**
Melk & Boter	0.172
Curd & Boter	0.110
Zure room & Boter	0.017
Melk & Zure room	0.000
Curd & Zure room	0.006
Melk & Curd	0.038
Melk & Aardappelen	0.009
Uien & Aardappelen	0.003
Brood & Aardappelen	0.012
Melk & Brood	0.002
Uien & Brood	0.004
Melk & Uien	0.39

Ook in Tajikistan is te zien dat er hoge correlaties zijn tussen producten die gerelateerd zijn aan elkaar (zie figuur 3 en tabel 4)



Figuur 3: Scattermatrix van tarwe, bloem, zout en brood in Tajikistan

Tabel 4: Mean square error scattermatrix Tajikistan

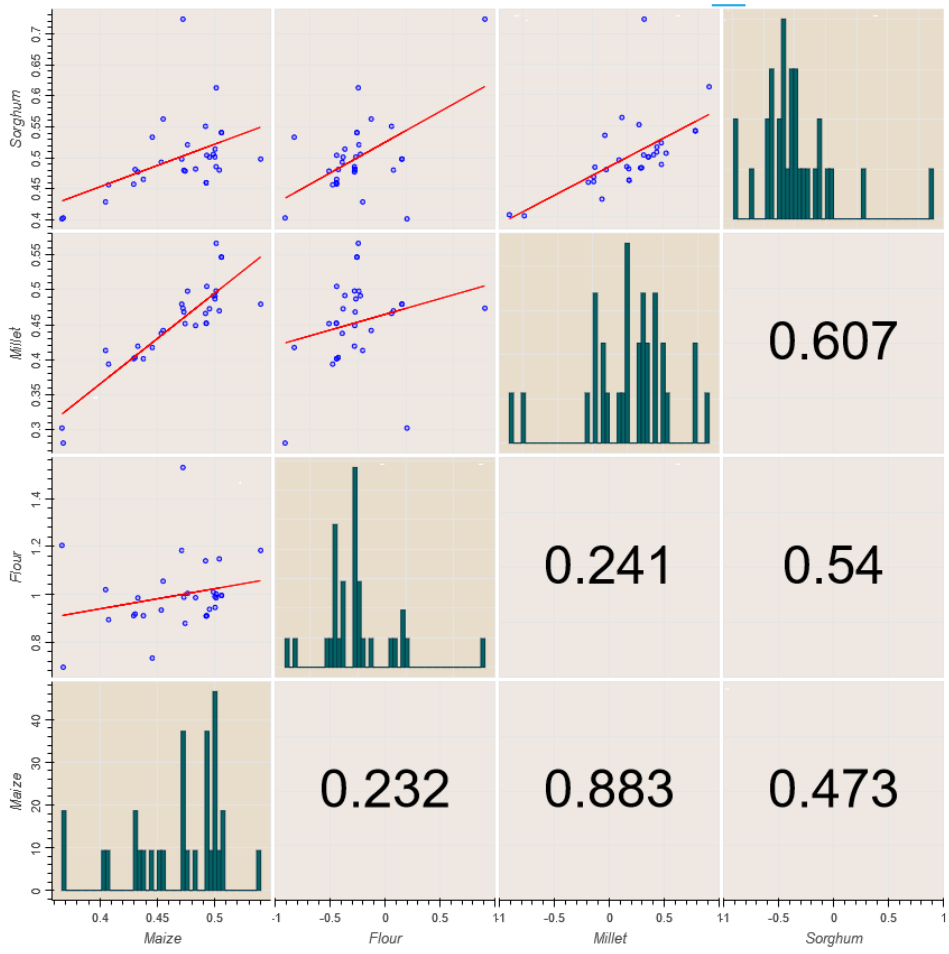
Goederen combinatie	**MSE**
Tarwe & Brood	0.077
Bloem & Brood	0.064
Zout & Brood	0.105
Tarwe & Zout	0.001
Bloem & Zout	0.001
Bloem & Tarwe	0.001

Natuurlijk is het krachtiger als correlaties tussen goederen in meerdere landen optreden. Daarom is ook gekeken naar het aantal landen met relevante correlaties tussen twee bepaalde goederen en de hoeveelheid landen die deze twee goederen in zijn data set heeft. Zo kan vergeleken worden welke combinatie van goederen in relatief de meeste landen een hoge correlatie hebben. Een overzicht van de producten met de hoogste correlatie in relatief de meeste landen is te vinden in tabel 5.

Tabel 5: Goederen met een relevante correlatie in een relatief groot aantal landen

Producten	**Aantal landen met deze producten**	**Aantal landen met significante correlatie tussen de producten**	**Percentage landen met correlatie t.o.v. totaal landen met deze producten**
Brood & Bloem	13	5	38.462
Graan & Olie	10	4	40
Graan & Bloem	8	3	37.5
Brandstof (diesel) & Brandstof (benzine)	20	7	35
Brandstof (diesel) & Bloem	20	7	43.75
Pasta & Bloem	11	4	36.364
Pasta & Zout	5	2	40
Sorghum & Maize	20	11	55
Sorghum & gierst	16	9	56.25
Maize & gierst	14	9	64.288

Wanneer een aantal goederen uit deze tabel met elkaar vergelijken worden voor een specifiek land komen er niet altijd relevante correlaties uit. Om hiervan een voorbeeld te geven, geeft figuur 4 een scattermatrix weer van Gambia, waarin de scatterplots tussen verschillende goederen in dat land, de distributie van data en de correlaties zijn weergegeven. Hierin is te zien dat niet alle combinaties van goederen hier een relevante correlatie tonen. Bloem en maïs, en bloem en millet (gierst) geven een relatief lage correlatie rond 0.24. Ook de correlatie tussen maïs en sorghum, en bloem en sorghum is niet erg hoog. Toch vertonen de andere goederen wel een hoge correlatie. De mean square errors (MSE) van deze scattermatrix zijn te vinden in tabel 6 en zijn relatief laag.



Figuur 4: Scattermatrix van gierst, maïs, sorghum en bloem in Gambia

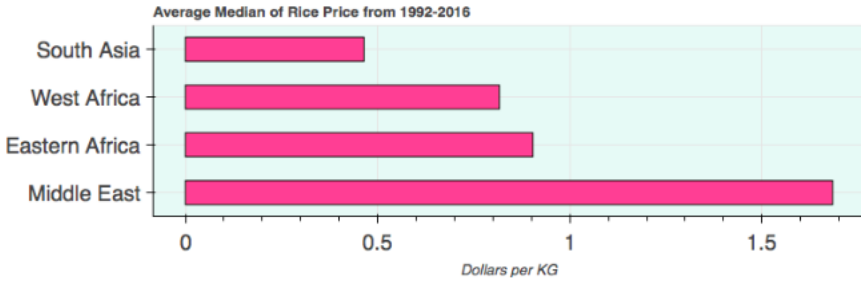
Tabel 6: Mean square error scattermatrix Gambia

Goederen Combinatie	**MSE**
Maïs & Sorghum	0.003
Bloem & Sorghum	0.002
gierst & Sorghum	0.002
Maïs & gierst	0.001
Bloem & gierst	0.003
Maïs & Bloem	0.021

Nog interessanter wordt het, wanneer wordt gekeken of bepaalde combinaties goederen vooral in bepaalde gebieden correleren zoals in de volgende paragraaf wordt besproken aan de hand van de goederen combinatie van rijst en olie.

3.2 Patronen van voedselprijzen tussen regio's

Deze paragraaf toont de resultaten op de vraag of binnen regio's dezelfde patronen in voedselprijzen zijn te zien en of deze patronen tussen regio's verschillend zijn. Hierbij zijn vier regio's gekozen: het Midden-Oosten, Oost-Afrika, West-Afrika en Zuid-Azië. Eerst zal een kort algemeen beeld geschetst worden van de rijstprijzen in deze regio's. Daarna volgen per regio scatterplots, tabellen en boxplots van de prijzen van rijst in combinatie met olie.



Figuur 5: Staafdiagram prijs van rijst voor verschillende regio's

In figuur 5 is te zien dat de gemiddelde mediaan van de prijs van rijst door de jaren heen per regio relatief veel lijkt te verschillen. De verschillende regio's bestaan uit dezelfde landen als de landen die gebruikt zijn voor de figuren 7-13. De aangrenzende regio's Zuid-Azië en het Midden-Oosten vertonen een groot verschil in prijs. De gemiddelde mediaan van de prijs van rijst is in het Midden-Oosten 1.68 US dollar per KG, dit gemiddelde is meer dan een US dollar hoger dan

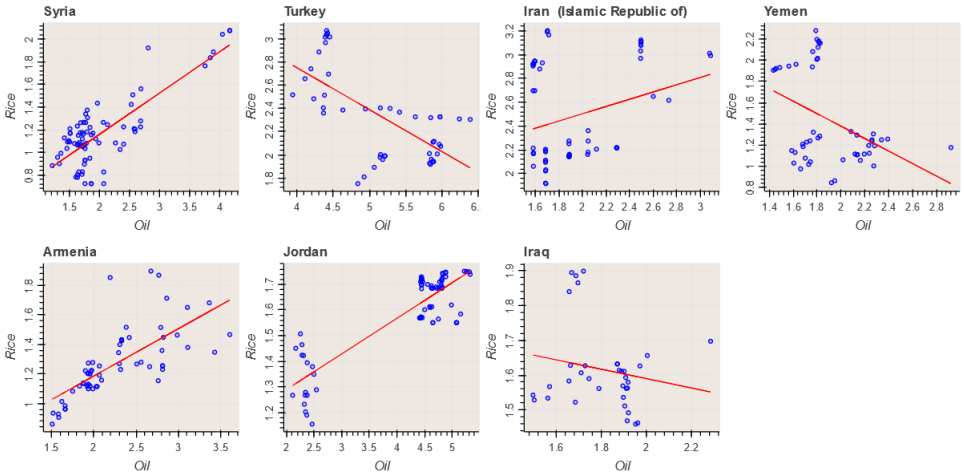
in Zuid-Azië, waar de gemiddelde mediaan op 0.46 US dollar per KG ligt. In het Midden-Oosten is rijst dus ongeveer 3.7 keer zo duur dan in Zuid-Azië. Voor de aangrenzende regio's Oost- en West-Afrika is dit verschil een stuk minder groot. De gemiddelde mediaan van de prijs van rijst is in Westelijk Afrika 0.82 US dollar per KG, in Oost-Afrika is dit 0.90 US dollar per KG. Dit is een verhouding van bijna 1 op 1. Om de verschillen echter nog beter uit te lichten, is het noodzakelijk meer in te zoomen op de regio's.

3.2.1 Midden-Oosten

De correlaties van olie en rijst, van verschillende landen in het Midden-Oosten zijn te zien in figuur 6. De Pearson correlatie, hoeveelheid meetpunten, MSE en gemiddelde prijzen van rijst en olie in deze landen is te vinden in tabel 7.

Opvallend aan de data van het Midden-Oosten, is dat dit het enige werelddeel is, waar negatieve correlaties te vinden zijn tussen de twee genoemde producten. Dit geldt voor drie van de zes gegeven landen: Turkije, Yemen en Iraq. De laatste twee hebben een Pearson correlatie tussen -0.5 en 0.5 en zijn daarmee relatief minder significant. Turkije daarentegen heeft een correlatie van -0.656 en is daarmee wel relatief significant.

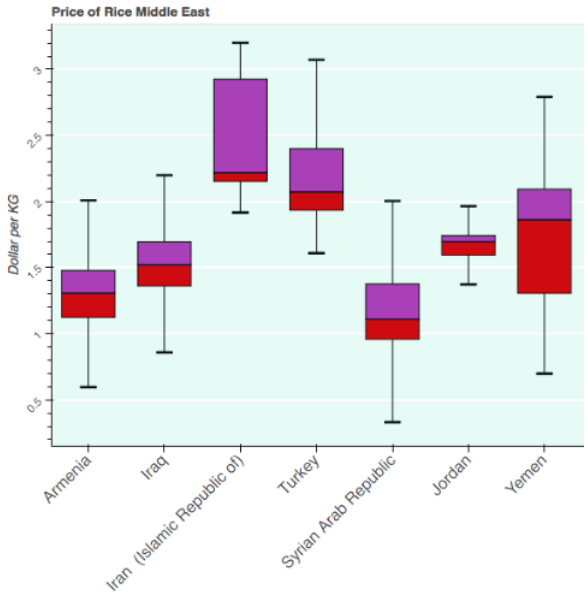
In figuur 7 is de prijsverhouding van rijst en olie van landen in het Midden-Oosten te zien. Uit de prijsverhouding van deze landen blijkt dat er tussen landen in deze regio geen grote afwijkende verschillen zijn. Aan de boxplots is te zien dat de meeste medianen tussen de 1 en 2 US dollar per KG liggen. In Turkije en Iran ligt de mediaan tussen de 2-2.5 US dollar per KG, hier heeft de prijs van rijst door de jaren heen hoger gelegen dan de rest van het Midden-Oosten.



Figuur 6: Prijs distributie van rijst en olie, Midden-Oosten

Tabel 7: Correlaties Rijst en Olie, Midden-Oosten

Land	Correlatie	N	MSE	Gem. rijstprijs (USD)	Gem. olieprijs (USD)
Yemen	-0.405	50	0.158	1.436	1.903
Jordan	0.867	76	0.007	1.593	4.180
Syria	0.802	77	0.033	1.169	2.023
Turkey	-0.656	47	0.085	2.335	5.149
Iran	0.827	64	0.155	2.483	1.933
Armenia	0.702	61	0.027	1.272	2.270
Iraq	-0.183	35	0.014	1.616	1.811



Figuur 7: Prijs distributie van rijst van landen in het Midden-Oosten 1992-2016

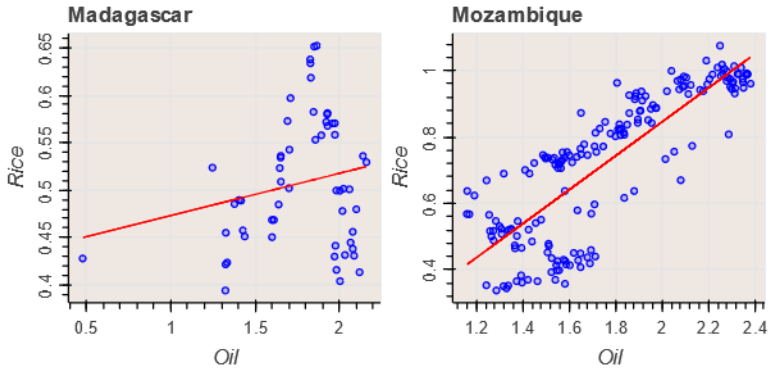
3.2.2 Oost-Afrika

Van de landen in Oost-Afrika zijn er slechts twee landen met data van olie en rijst. Dit zijn Madagascar en Mozambique. De verdelingen zijn te vinden in figuur 8. De corresponderende correlaties, hoeveelheid meetpunten, MSE en gemiddelde prijzen zijn te vinden in tabel 8.

Zoals te zien is in tabel 8, is de correlatie tussen olie en rijst niet significant in Madagascar. In Mozambique daarentegen is de correlatie relatief hoog met een groot aantal meetpunten.

In Oost-Afrika ligt de prijs van rijst door de jaren heen een stuk lager dan in het Midden-Oosten. In figuur 9 is te zien dat de meeste waardes van de prijs van rijst tussen de 0.5 en 1.2

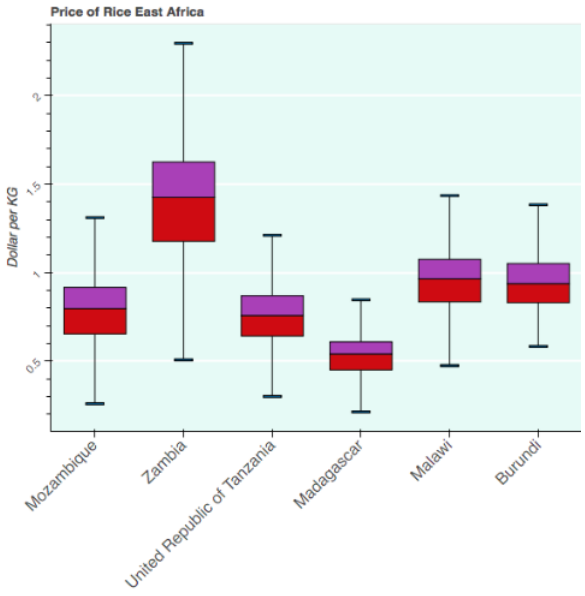
US dollar per KG ligt. Zambia is het enige Oost-Afrikaanse land dat hier lichtelijk vanaf wijkt, Zambia heeft grote verschillen in de prijs van rijst gehad, met een spreiding van meer dan 2.5 US dollar per KG.



Figuur 8: Prijs distributie van rijst en olie, Oost-Afrika

Tabel 8: Prijs distributie van rijst en olie, Oost-Afrika

Land	**Correlatie**	**N**	**MSE**	**Gem. rijstprijs (USD)**	**Gem. olieprijs (USD)**
Madagascar	0.147	16	0.004	0.508	1.774
Mozambique	0.822	177	0.015	0.724	1.763



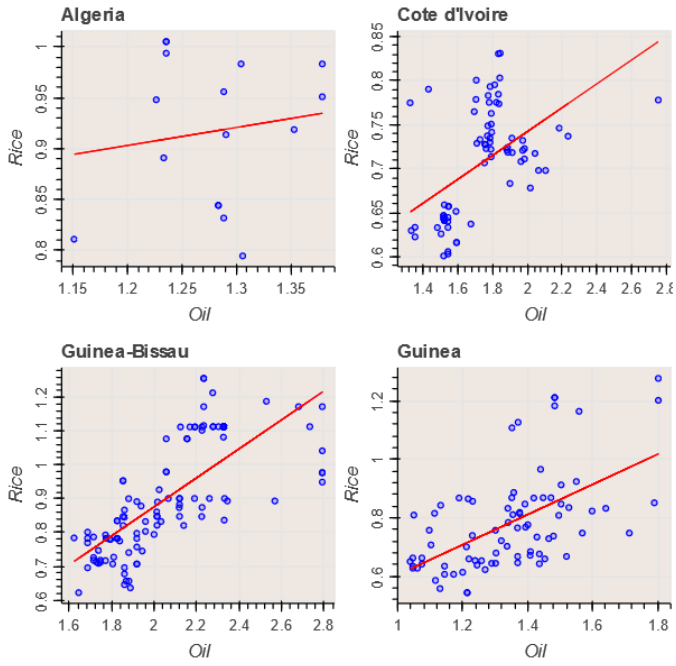
Figuur 9: Prijs distributie van rijst van landen in Oost-Afrika 1992-2016

3.2.3 West-Afrika

West-Afrika heeft een aantal landen met een relatief hoge correlatie (zie tabel 9). Wie echter kijkt naar het corresponderende figuur 10, ziet dat de distributie van Algerije te willekeurig is, en dat de lineaire regressie lijn hier erg afwijkt van alle meetpunten. Voor 16 meetpunten is de MSE erg groot. Hier lijken de punten dus relatief willekeurig verdeeld te zijn en geen lineair verband te vertonen.

Wat verder opvallend is, is dat de meeste datapunten van de andere drie landen ook ver van de lineaire regressie lijn staan. In verhouding met het aantal data punten is de MSE groot.

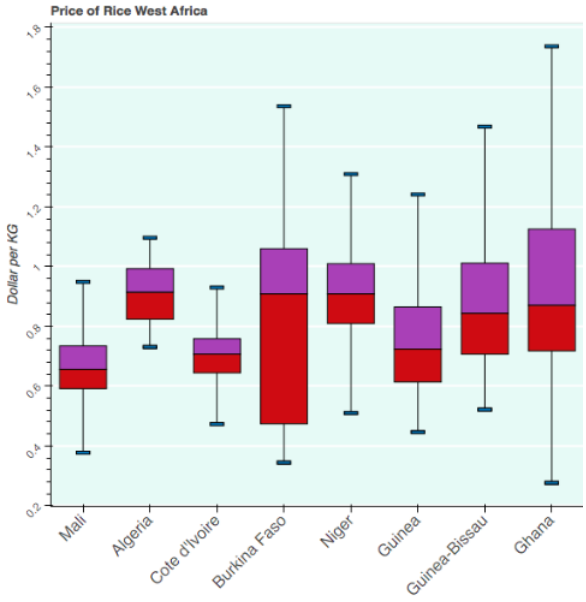
Wat betreft rijstprijzen zijn de waardes in West-Afrika vergelijkbaar met de aangrenzende regio Oost-Afrika. In figuur 11 is te zien dat de meeste waardes van de prijs van rijst tussen de 0.4 - 1 US dollar per KG ligt. Aan de boxplot is te zien dat de verschillende landen op gelijke plekken in de grafiek liggen. Dit betekent dat alle landen in West-Afrika vergelijkbare prijzen voor rijst hebben gehad.



Figuur 10: Prijs distributie van rijst en olie, West-Afrika

Tabel 9: Correlaties Rijst en Olie, West-Afrika

Land	Correlatie	N	MSE	Gem. rijstprijs (USD)	Gem. olieprijs (USD)
Algeria	0.147	16	0.005	0.917	1.280
Cote d'Ivoire	0.522	74	0.003	0.707	1.739
Guinea-Bissau	0.738	114	0.012	0.904	2.065
Guinea	0.564	80	0.019	0.780	1.340



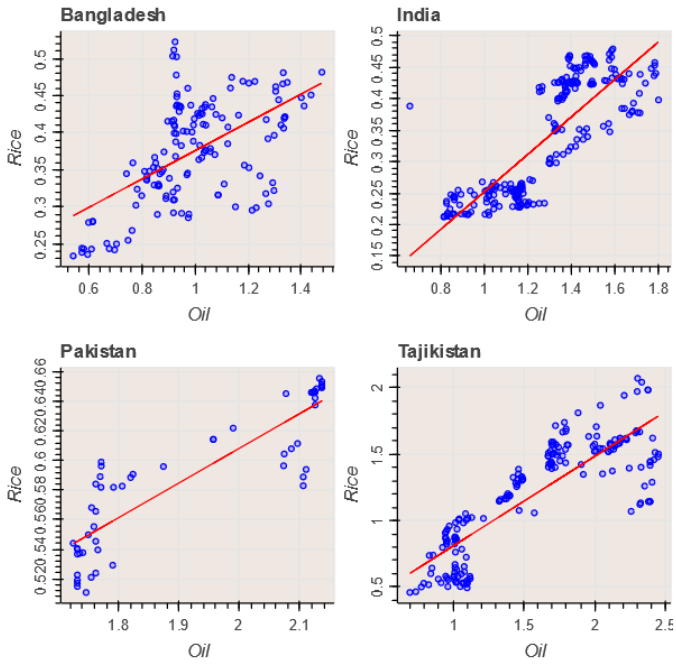
Figuur 11: Prijs distributie van rijst van landen in West-Afrika 1992-2016

3.2.4 Zuid-Azië

Van alle beschikbare landen in Zuid-Azië zijn er drie landen met datapunten van olie en rijst (zie figuur 12). In Zuid-Azië is het opvallend dat er veel datapunten beschikbaar zijn in Bangladesh, India en Tajikistan en dat deze ook een hoge correlatie vertonen (zie tabel 10). Vooral in India is de correlatie hoog en de MSE in verhouding vrij laag. Ook Tajikistan en Pakistan hebben een hoge correlatie tussen olie en rijst. De lineaire regressie geeft hier een betere indicatie van de data dan bij de landen in West-Afrika. Aangezien de MSE afhankelijk is van de hoeveelheid datapunten, is deze ook extreem laag als je het vergelijkt met de landen in West-Afrika. India en Cote d'Ivoire hebben bijvoorbeeld een even grote MSE, maar hebben respectievelijk 228 en 74 data punten.

In figuur 13 is te zien dat Zuid-Azië de goedkoopste rijstprijzen door de jaren heen heeft gehad. De prijs voor rijst ligt hier tussen de 0.2-0.8 US dollar per KG. Aan het figuur is te zien dat India,

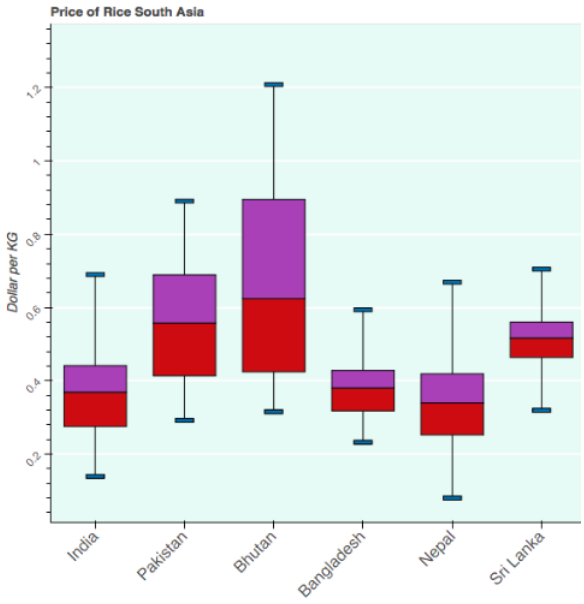
Bangladesh en Nepal zeer vergelijkbare boxplots vertonen. Dit suggereert dat de prijs van rijst door de jaren heen zeer vergelijkbaar is geweest.



Figuur 12: Prijs distributie van rijst en olie, Zuid-Azië

Tabel 10: Correlaties Rijst en Olie, Zuid-Azië

Land	**Correlatie**	**N**	**MSE**	**Gem. rijstprijs (USD)**	**Gem. olieprijs (USD)**
Bangladesh	0.569	143	0.003	0.374	0.993
India	0.822	228	0.003	0.336	1.280
Pakistan	0.874	55	0.001	0.588	1.915
Tajikistan	0.842	187	0.051	1.205	1.589



Figuur 13: Prijs distributie van rijst van landen in Zuid-Azië 1992-2016

3.3 Invloed van andere factoren

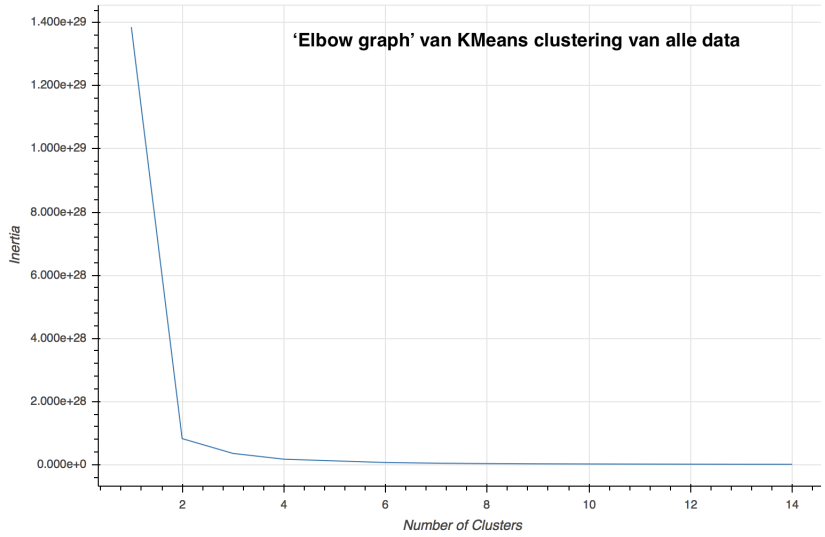
Naast correlaties en verbanden van goederen onderling, zijn ook de verbanden en correlaties van voedselprijzen met externe factoren als BBP, populatie, vluchtelingen en neerslag onderzocht. In deze paragraaf volgen de resultaten daarvan. Eerst worden de resultaten van het KMeans algoritme en lineaire regressie deze datasets als geheel besproken. Daarna wordt per factor dieper ingegaan op individuele regio's.

3.3.1 Globaal beeld

Door mondialisering raken de handelsstromen in de wereld steeds verder met elkaar vervlochten en kunnen gebeurtenissen in het ene deel van de wereld invloed hebben op het andere deel van de wereld (Lévy, 2007). Daarom zijn de voedselprijzen van bepaalde producten niet alleen geïsoleerd geanalyseerd maar ook op een globaal niveau onderzocht met het KMEans algoritme en lineaire regressie. Deze subparagraaf wijdt uit over deze analyse.

Ten eerste is gekeken of alle meetpunten op een bepaalde manier te clusteren zijn door het KMeans algoritme toe te passen op een combinatie van bijna alle datasets [2]. Aangezien van te voren niet bekend was hoeveel clusters er mogelijk in de data zaten, is het KMeans algoritme toegepast op een bereik van één tot tien clusters.

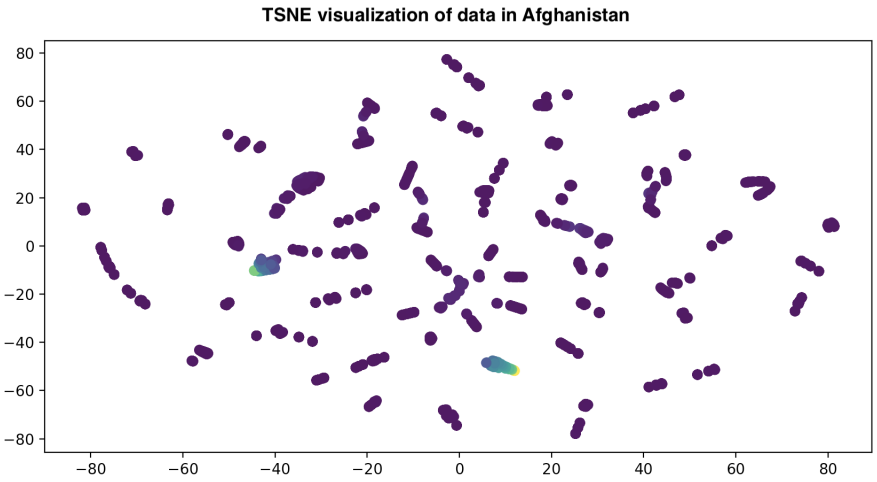
[2]Zoals al eerder gezegd is de dataset van de vluchtelingen niet meegenomen omdat er te veel waarden ontbraken



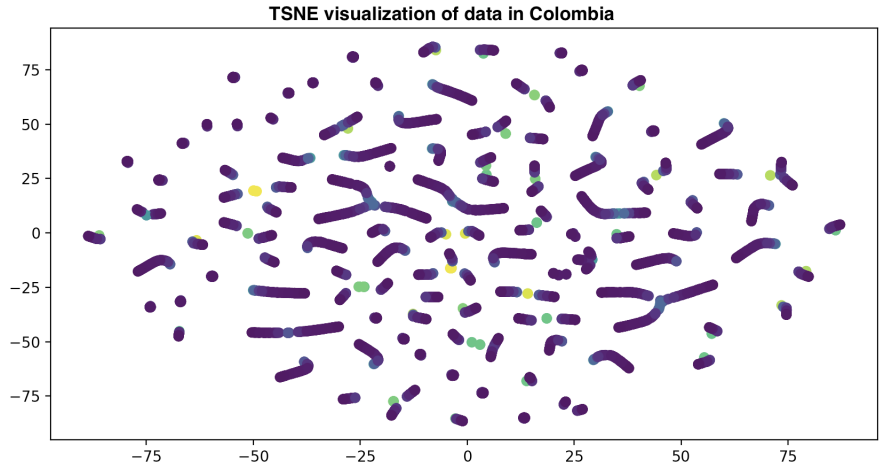
Figuur 14: 'Elbow graph' van de KMeans clustering met de inertie uitgezet tegen het aantal clusters (N = 556188)

In figuur 14 is te zien dat twee het optimale aantal clusters is. Hier bevindt zich namelijk de zogenoemde elleboog. Tegelijkertijd lijkt de inertie bij twee clusters nog steeds relatief hoog te zijn. Bij nadere analyse blijkt dat de inertie bij twee clusters ongeveer $8,21 * 10^{27}$ bedraagt. Hoewel het veel datapunten betreft, is dit nog steeds een vrij hoge inertie. Dit suggereert dat de data relatief willekeurig verdeeld is over de hoog-dimensionale ruimte.

Een steekproefgewijze TSNE-visualisatie van de data van Afghanistan en Gambia laten dit patroon ook zien (zie figuur 15 en 16).



Figuur 15: TSNE visualisatie van alle datapunten in Afghanistan (N = 4900)



Figuur 16: TSNE visualisatie van alle datapunten in Colombia (N = 8648)

Naast classificatie is ook geprobeerd om op de data als geheel lineaire regressie uit te voeren. In tabel 11 is te zien dat de geschatte coëfficiënten van de regressielijnen allemaal nul benaderen. Ook is te zien dat de Pearson-coëfficiënten zeer laag zijn en dat er dus nauwelijks correlatie is tussen deze factoren en voedselprijzen op globale schaal.

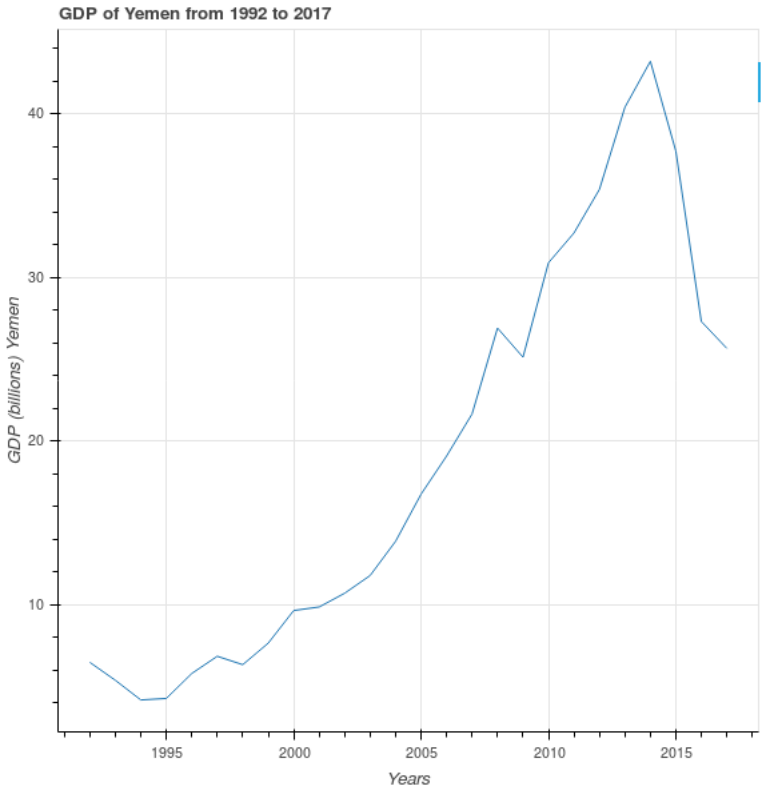
Tabel 11: Geschatte coëffciënten van de regressielijnen van de alle data van een bepaalde factor tegenover de voedselprijzen (N = 556188)

Factor	**Geschatte regressie coëfficiënt**	**Correlatiecoëfficiënt**
Populatie	$3.36 * 10^{-3}$	-0.029
BBP	$-9.52 * 10^{-9}$	-0.030
Neerslag	$1.33 * 10^{-6}$	0.013

3.3.2 BBP

In tabel 12 is te zien wat de correlatie is van het BBP met de prijzen van verschillende goederen in verschillende landen in het Midden-Oosten, door de jaren heen. Wat opvallend is, is dat de correlaties van de verschillende goederen tussen de verschillende landen sterk wisselt. Een voorbeeld hiervan is dat het BBP in Jordanië een hoge correlatiecoëfficiënt heeft met de rijstprijs terwijl deze correlatie in het buurland Irak negatief is. Een ander opvallend punt is dat zowel de rijst- als suikerprijs in Yemen negatief correleert met het BBP. Uit de data van het BBP van Yemen blijkt dat deze laatste van 2014 tot 2017 gedaald is met ongeveer 17 miljard zoals geïllustreerd in 17.

Een andere opvallende bevinding is dat het BBP sterk positief correleert met de prijs van benzine en petrol-gasoline. Het BBP heeft een correlatie coëfficiënt van 0.94 met benzine (17 datapunten) en ook 0.94 (18 datapunten) met petrol-gasoline. Aangezien deze voor respectievelijk 17 en 18 jaren zijn gemeten, lijkt dit resultaat vrij betrouwbaar.



Figuur 17: BBP Yemen 1992 tot 2017

Tabel 12: Correlaties BBP met verschillende goederen prijzen in het midden-oosten door de jaren heen

Land	**Rijst**	**Olie**	**Suiker**
Turkije	0.81	-0.81	0.97
Iran	0.13	0.83	0.91
Yemen	-0.95	0.39	-0.83
Armenië	-0.41	0.01	0.15
Jordan	0.95	0.87	0.60
Irak	-0.60	0.72	0.91
Afghanisten	0.21	Onbekend	Onbekend

Voor de compleetheid is het de moeite waarde om te vermelden dat er in de andere regio's geen opvallend resultaten waren voor de correlatie tussen BBP en prijzen van goederen. Het BBP speelde vooral in het Midden-Oosten een grote rol bij de fluctuatie in voedselprijzen.

3.3.3 Populatie

In tabel 13 zijn de Pearson correlaties te zien tussen de populatiegrootte van een land en de productprijzen. Dit is uitgevoerd voor alle landen en goederen in het databestand. Hieruit is naar voren gekomen dat er bijna geen correlaties zijn voor deze combinatie. De enige correlaties die van significant belang zijn, zijn de correlatie tussen brood en de populatiegrootte van Algerije en de correlatie tussen brood en de populatiegrootte van Afghanistan. Deze zijn -0.99 en -1.00, dit betekent dat de prijs van brood in deze twee landen daalt als de populatiegrootte stijgt. Deze waarden zijn echter niet zeer verrassend omdat in beide landen respectievelijk 2 en 1 datapunten zijn gebruikt. Met zo weinig datapunten zegt een hoge correlatiecoëfficient weinig.

Tabel 13: Correlaties tussen producten en populatiegrootte

Land	**Product**	**Correlatie**
Afghanistan	Bread	-1.00
Algeria	Bread	-0.99

3.3.4 Regenval

De regenval per jaar heeft een hoge positieve correlatie met de prijzen van brood, maar niet met andere producten. De Pearson correlatie is ook toegepast op de prijs van producten van een jaar later met de regenval per jaar. De aanname hierbij was dat een hoge of lage regenval pas een jaar later effect heeft. Ook dit toonde echter geen correlaties tussen andere producten dan brood. De resultaten hiervan zijn te in tabel 14. Daaruit is op te maken dat de correlaties in alle gevallen relatief sterk positief zijn.

Tabel 14: Correlaties tussen brood en regenval in hetzelfde jaar voor alle landen met N > 5

Land	**N**	**Correlatiecoëfficiënt**
Guatemala	14	0.80
Kenya	9	0.76
Kyrgyzstan	10	0.67
Tajikistan	13	0.53

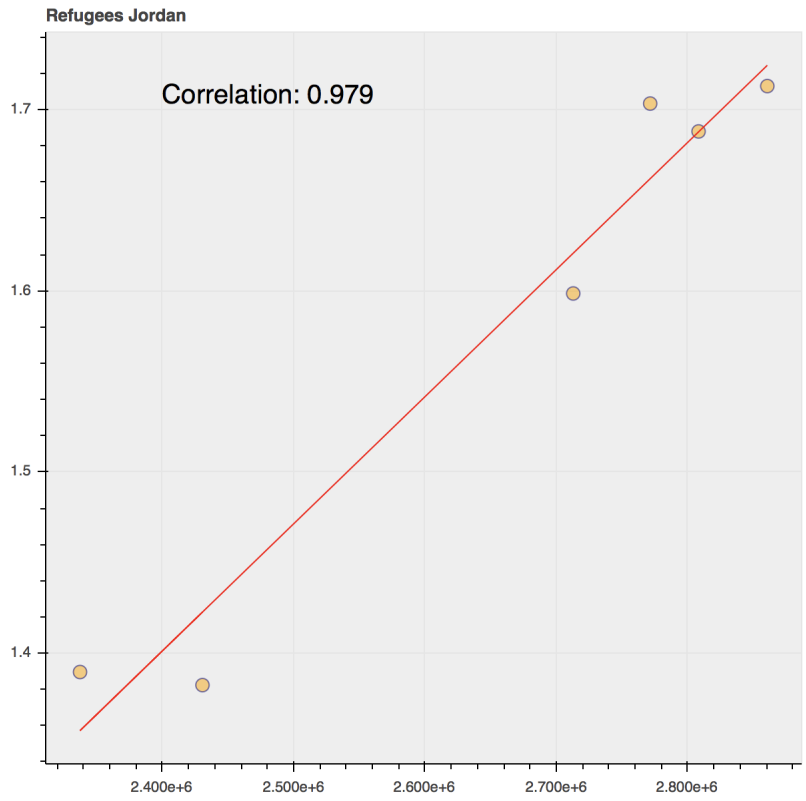
3.3.5 Vluchtelingenstromen

In tabel 15 staan de 10 landen die de meeste vluchtelingen hebben opgenomen ten opzichte van de populatiegrootte van het land (Guterres & Peter, 2012). Somalië heeft met de score 0.995 de

hoogste correlatie, Jordan en Irak volgen met een score van 0.979 (zie figuur 18) en 0.881. Bij deze landen stijgt de graanprijzen dus als de vluchtelingenpopulatie groter wordt. Turkije heeft een opvallend negatieve correlatie, met een score van -0.603. Dit betekent dat de prijs van graan daalt als de vluchtelingenpopulatie in Turkije groter wordt.

Tabel 15: Correlatie tussen aantal vluchtelingen in een land en graanprijzen

Land	**Relevante correlatie graanprijzen en opgevangen vluchtelingen**
Lebanon	0.403
Pakistan	0.684
Ethiopia	0.379
Chad	0.826
Jordan	0.979
Uganda	0.421
Burundi	0.792
Somalia	0.995
Iraq	0.881
Mali	0.443
Turkey	-0.603
Cote d'Ivoire	-0.693



Figuur 18: Vluchtelingen uitgezet tegen graanprijzen in Jordanië (N = 6)

4 Conclusie en discussie

Uit de resultaten blijkt dat een algemeen antwoord op de hoofdvraag lastig is. Zoals in de globale analyse met KMeans en lineaire regressie bleek zijn er namelijk geen factoren die op globale schaal, dat wil zeggen in alle landen, invloed hebben op voedselprijzen. De verbanden zijn per land specifiek en veel genuanceerder.

Wel zijn er een aantal interessante globale correlaties te vinden tussen bepaalde producten die de hypothese op de eerste deelvraag lijken te ondersteunen. Voorbeelden hiervan zijn aubergine en courgette (0.9671), bloem en mais (0.8164), vee en yoghurt (-0.8519), en vee en kaas (-0.8818). De eerste combinatie is te verklaren doordat de goederen vergelijkbaar zijn, beiden groenten. De tweede combinatie is ook logisch omdat mais vaak als ingrediënt dient voor bloem. Een mogelijke verlaring voor de derde en vierde combinatie is duurder vee meer melk geeft waardoor de kaas- en yoghurtprijzen omlaag gaan. Deze correlaties zijn echter uitgerekend met alle landen voor alle

jaren. Hierdoor kunnen correlaties gevonden worden die in geen enkel land voorkomen, maar wel gelden voor alle landen samen. Een voorbeeld hiervan is de combinatie aardappelen en vis, die wereldwijd een correlatie hebben van -0.8221, terwijl er geen enkel land is die een relevante correlatie heeft tussen deze twee goederen. Aangezien de visprijzen van het ene land waarschijnlijk enkel een toevallige correlatie hebben met de aardappelprijzen van het andere land, zijn deze correlaties niet veelzeggend. Als oplossing hiervoor is er toen gekeken naar her percentage landen die een hoge correlatie heeft voor een bepaalde productcombinatie. Door deze te tellen, is in tabel 5 te zien dat sommige producten in een groot deel van de landen een hoge correlatie hebben met elkaar. Dit suggereert dat deze producten over het algemeen gecorreleerd zijn in prijs.

Sommige producten in deze lijst zijn ingrediënten van elkaar. Voorbeelden hiervan zijn brood en bloem, pasta en bloem, en pasta en zout. Andere landen zijn gecorreleerd omdat ze qua productie op elkaar lijken en hetzelfde doel dienen. Voorbeelden hiervan zijn brandstof (diesel) en brandstof (benzine), sorghum en maïs, sorghum en gierst, en gierst en maïs. De overige producten, graan en olie, en brandstof en bloem, hebben een minder opvallende oorzaak voor hun correlaties. Wellicht is er sprake van toevalligheid, of wordt het veroorzaakt door het feit dat deze producten veel worden gebruikt en daarom op dezelfde manier mee schalen met de economie van een land. Over het algemeen ondersteunt dit de hypothese van de eerste deelvraag dat goederen die op een bepaalde manier aan elkaar gekoppeld zijn ook in prijs aan elkaar gekoppeld zijn.

Ook de scattermatrices in figuur 2 ondersteunt deze conclusie. Hierin is de correlatie tussen zuivelproducten aan de ene kant en de correlatie van een aantal willekeurige producten aan de andere kant van Oekraïne weergegeven. Hierin is heel duidelijk te zien dat de melkproducten een veel hogere correlatie hebben met elkaar, dan de andere producten. Melkproducten hebben natuurlijk een vergelijkbaar maakproces en dezelfde ingrediënten. Ondanks dat uien en brood veel worden geconsumeerd, hebben deze niet een relevante correlatie. Uien en aardappelen daarentegen hebben wel een relevante correlatie (0.885), maar zijn geen ingrediënten van elkaar. Deze correlatie is niet zo hoog als bij de melkproducten. Aardappelen en uien hebben een vergelijkbaar oogstproces en worden vaak naast elkaar gegeten. Wellicht is dat de oorzaak voor hun hoge correlatie. Dit neemt niks weg van het feit dat hier geldt dat producten die ingrediënten van elkaar zijn een hoge correlatie tonen. Dit is ook weer te zien in figuur 3.

Deze algemene conclusie dat gekoppelde goederen ook in prijs gekoppeld zijn moet echter ook genuanceerd worden. Om de correlatie tussen deze goederen iets beter zichtbaar te maken, is nanelijk een scattermatrix gemaakt tussen de goederen sorghum, maïs, gierst en bloem in Gambia (figuur 4). Hierin is vooral te zien dat niet alle goederen in dit land met elkaar correleren. Dit laat dus zien dat de correlaties tussen de goederen uit tabel 5 niet altijd gelden, en ook niet in dezelfde

landen hoeven te zijn. Dit is iets om rekening mee te houden, aangezien bepaalde correlaties dus wel worden gevonden, maar dat het onzeker is of dit geldt voor landen verspreid over verschillende gebieden.

In toekomstig onderzoek is het dus van belang om een beter beeld te hebben van de verschillende correlaties per regio, en rekening te houden met het feit dat wereldwijde correlaties niet altijd iets zeggen over de producten, omdat deze ook toevallig een hoge correlatie kunnen hebben. Daarnaast zou de dataset aangevuld kunnen worden met meer goederenprijzen, zodat scattermatrices, zoals van Gambia, voor meerdere landen gemaakt kunnen worden en een beter algemeen beeld kunnen schetsen van de correlaties.

Na de analyse van voedsel over de hele wereld, is er meer gekeken naar correlaties tussen verschillende regio's in de wereld. Ook hier wordt de gestelde hypothese deels ondersteund door de resultaten. Aan de ene kant bleek dat de rijsprijzen binnen een regio goed overeenkwamen maar tussen regio's sterk konden verschillen. Zo waren de prijzen in Azië veel lager dan in Afrika en het Midden-Oosten. Een mogelijke verklaring hiervoor is dat ten opzichte van veel landen in Azië, de rijstconsumptie in Afrika de afgelopen jaren is toegenomen. 40 Procent van de geconsumeerde rijst wordt geïmporteerd, wat de hogere prijs kan verklaren. Verder is de algemene rijstprijs in het Midden-Oosten ook aanzienlijk hoger dan in de andere regio's, waar zo verder op in wordt gegaan.

Verder zijn er echter ook overeenkomsten tussen regio's. Zo was in veel landen olie en rijst gecorreleerd, al lagen de prijzen uit elkaar. Het figuur 6 en de bijbehorende tabel 7 laten de correlaties tussen rijst en olie zien voor bepaalde landen uit het Midden-Oosten. Opvallend hieraan is, dat Turkije het enige land is dat een relevante negatieve correlatie heeft tussen de twee producten. Turkije is een land met weinig rijstproductie, en veel import van rijst (CGIAR, Research Program on Rice, 2017b). Ook wordt er relatief weinig rijst geconsumeerd per maaltijd, wanneer je dit vergelijkt met de rijstconsumptie per maaltijd van bijvoorbeeld India (CGIAR, Research Program on Rice, 2017a). Bij een lagere vraag is de prijs vaak ook hoger. Azië voorziet zo'n 90 procent van de consumptie en productie van rijst (Abdullah, Ito & Adhana, 2006), wat wellicht een oorzaak is voor de lage prijs van het product en de grote hoeveelheid datapunten (te zien in tabel 10).

Olie daarentegen, is een bloeiende economie in Turkije, en zelfs belangrijk voor de economische groei van het land (Aktaş & Yılmaz, 2008). Een mogelijke verklaring voor de negatieve correlatie is daarom dat brandstofprijzen al jaren hoger worden en rijstimport dus duurder wordt terwijl olieproductie beter loopt en de olieprijs lager wordt. In ieder geval is er geen directe aanleiding om te vermoeden dat olie en rijst in Turkije negatief gekoppeld zijn, zeker gezien de positieve correlaties in andere landen.

Vervolgens zijn in figuur 8 de correlatie tussen rijst en olie te zien van Madagascar en Mozambique. In Madagascar was de correlatie niet relevant genoeg, maar in Mozambique was er juist een hoge correlatie tussen rijst en olie.

Concluderend is uit de figuren 7-11 gebleken dat landen in dezelfde regio vergelijkbare distributies van de prijs van rijst hebben. Ook is er in figuur 5 te zien dat de prijs van rijst per regio significant verschilt. Dit resultaat is aannemelijk omdat de prijzen van landen in dezelfde regio waarschijnlijk te maken hebben met dezelfde factoren.

Toch is de hypothese van de tweede deelvraag niet compleet juist omdat verschillende regio's wel overeenkomstige correlaties vertonen.

De resultaten voor de derde deelvraag zijn zeer wisselend. Uit de diepere analyse was geen globale correlatie te zien tussen een van de factoren en voedselprijzen. Wel zijn er op kleinere schaal relevante bevindingen gedaan.

Zo blijkt uit 15 dat zowel landen in Oost-Afrika, als de meeste landen in het Midden-Oosten positieve correlaties tonen tussen graanprijzen en de vluchtelingenstroom. Dit kan een verklaring geven waarom vergelijkbare gebieden soortgelijke prijs veranderingen hebben. Dit komt overeen met Guterres en Peter (2012) waarin wordt gesteld dat de vluchtelingenstroom in een land gekoppeld kan worden aan de prijzen van voedsel.

Vooralsnog kan deze correlatie niet direct gelinkt worden aan de resultaten van de figuren 7-13 en figuur 5. Bij deze figuren is namelijk enkel gekeken naar de prijs van rijst, terwijl bij de correlaties de prijs van zowel rijst als van maïs, gierst en sorghum is bekeken. Rijst is gekozen voor deze analyse omdat dit het meest voorkomende product is in de gegeven data set. Een verbeterpunt van dit onderzoek is om naast rijst de prijsverandering van alle graanprijzen te analyseren, dit zal de analyse vollediger maken.

Verder is uit tabel 13 naar voren gekomen dat de factor populatiegrootte nauwelijks van belang is bij het analyseren van voedselprijzen. De verschillende voedselprijzen in landen kunnen dus niet gekoppeld worden aan het verschil in populatiegrootte in een land. Wat in plaats van populatiegrootte misschien wel gelinkt is aan de prijs voor voedsel, is de populatiedichtheid. In een vervolgonderzoek kan het nuttig zijn om deze correlatie te berekenen.

Uit de correlaties tussen regenval en productprijs is naar voren gekomen dat de regenval per jaar geen hoge correlatie heeft met de prijs van een product. Een verklaring had kunnen zijn dat het effect van regenval pas een jaar later zichtbaar is in de prijs van een product. Deze correlaties zijn daarom ook geanalyseerd, maar ook dit geeft geen hoge correlaties. Er zou in vervolgonderzoek

nog gekeken kunnen worden naar de regenval per maand of per periode. De schommeling van de prijs van een product per jaar zou dan mogelijk verklaard kunnen worden door de regenval. Hierbij moeten dan echter wel meer datapunten aanwezig zijn om een relevante correlatie vast te stellen.

Ook wat betreft het BBP is er geen eenduidig beeld ontstaan. Sommige landen vertoonden een positieve correlatie met bepaalde goederen en het BBP. Andere landen vertoonden weer een negatieve correlatie met bepaalde goederen en het BBP. Aangezien de correlaties hier zeer wisselend waren, is het niet mogelijk hier een algemene conclusie uit te trekken.

Dat betekent dat voor de derde deelvraag het antwoord grotendeels is dat de externe factoren over het algemeen geen invloed hebben. Alleen in een paar specifieke gevallen zoals vluchtelingen in Jordanië en graanprijzen is de correlatie significant. Deze conclusie voor de derde deelvraag is tegen de verwachtingen in. Mogelijk kan in de toekomst worden gezocht welke factoren een grotere rol spelen bij voedselprijzen. Hierbij kan gedacht worden aan de politieke situatie van een land.

Uit dit onderzoek is vooralsnog gebleken dat gekoppelde goederen vaak ook gekoppeld zijn in prijs. Verder is een conclusie dat er binnen regio's overeenkomstige prijzen heersen en deze verschillen van regio's waar andere omstandigheden heersen. Welke omstandigheden uiteindelijk de meeste invloed hebben op prijzen, blijft echter ook na dit onderzoek nog de vraag.

Referenties

Abbott, P. C., Hurt, C., Tyner, W. E. et al. (2009). *What's driving food prices? march 2009 update* (Rapport). Farm Foundation. Verkregen van `https://ideas.repec.org/p/ags/ffispa/48495.html`

Abdullah, A. B., Ito, S. & Adhana, K. (2006). Estimate of rice consumption in asian countries and the world towards 2050. In *Proceedings for workshop and conference on rice in the world at stake* (Dl. 2, pp. 28–43).

Aktaş, C. & Yılmaz, V. (2008). Causal relationship between oil consumption and economic growth in turkey.

Business Insider. (2018, Jun). *7 foods you eat every day that will raise your grocery bill this year.* Business Insider. Verkregen van `http://www.businessinsider.com/foods-you-eat-everyday-raise-grocery-bill-2018-expensive-groceries-2018-6?international=true&r=US&IR=T`

CGIAR, Research Program on Rice. (2017a). *Rice consumption India.* Ricepedia. Verkregen van `http://ricepedia.org/india`

CGIAR, Research Program on Rice. (2017b). *Rice consumption Turkey.* Ricepedia. Verkregen van http://ricepedia.org/turkey

Climate Change Knowledge Portal. (z.d.). *Download Data [Dataset].* World Bank Group. Verkregen van http://sdwebx.worldbank.org/climateportal/index.cfm?page=downscaled_data_download&menu=historical

De Telegraaf. (2018, May). *Voedselprijzen hard omhoog.* Telegraaf. Verkregen van https://www.telegraaf.nl/financieel/2104426/voedselprijzen-hard-omhoog

Guterres, A. & Peter, S. (2012). The state of the world's refugees: Adapting health responses to urban environments. *JAMA*, *308*(7), 673-674. Verkregen van http://dx.doi.org/10.1001/2012.jama.10161

Lévy, B. (2007). The interface between globalization, trade and development: Theoretical issues for international business studies. *International Business Review*, *16*(5), 594–612.

Lukes. (2018, Apr). *lukes/ISO-3166-Countries-with-Regional-Codes [Dataset].* Verkregen van https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv

Roser, M. & Ritchie, H. (2018). *Food prices.* Verkregen van https://ourworldindata.org/food-prices

The World Bank. (2017a). *GDP (current US$)[Dataset].* The World Bank. Verkregen van https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

The World Bank. (2017b). *Official exchange rate (LCU per US$, period average) [Dataset].* The World Bank. Verkregen van https://data.worldbank.org/indicator/PA.NUS.FCRF

The World Bank. (2017c). *Population, total [Dataset].* The World Bank. Verkregen van https://data.worldbank.org/indicator/SP.POP.TOTL

The World Bank. (2017d). *Refugee population by country or territory of asylum [Dataset].* The World Bank. Verkregen van https://data.worldbank.org/indicator/SM.POP.REFG

Timmer, C. P. (2008). *Causes of high food prices* (Rapport). ADB Economics Working Paper Series.

United Nations. (2017, Oct). *Syria: Un agencies deliver critical food aid, medicines to families trapped in rural damascus — un news.* United Nations. Verkregen van https://news.un.org/en/story/2017/10/569722-syria-un-agencies-deliver-critical-food-aid-medicines-families-trapped-rural