Universiteit van Amsterdam

Kunstmatige Intelligentie

Data analysis and visualization

De wondere wereld van voedselprijzen

Auteurs:	*Studentnummers:*
Chandni Bagchi	11824883
Stijn Hamerslag	11323795
Emiel Sanders	11836741
Luc Vink	10806113

1 Introductie

Er zijn talloze factoren die de globale voedselprijzen beïnvloeden. Om voedselcrises te voorkomen is het belangrijk om deze prijzen zo goed mogelijk te analyseren en om te onderzoeken hoe ze elkaar beïnvloeden zodat er ingegrepen kan worden, mocht dat nodig zijn.[1] Over het algemeen zijn voedselprijzen over de hele wereld consistent aan het stijgen.[2] In een onderzoek naar de reden van deze stijging concludeert de World Bank dat dit grotendeels wordt veroorzaakt door de opkomst van biobrandstoffen. Vanwege de hierdoor stijgende vraag naar voedsel en de minder snel groeiende productie van voedsel zou de prijs stijgen. Met deze conclusie schuift de World Bank de invloed van de wereldwijd stijgende temperatuur af als een factor die weinig invloed heeft op de prijs van voedsel.[3] Echter beweert Earth Policy Institute dat stijgende temperaturen als gevolg van klimaatverandering wel degelijk invloed hebben op voedselprijzen.[4] In dit onderzoek wordt onderzocht hoeveel invloed temperatuur werkelijk heeft op voedselprijzen.

[1] https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/1630.pdf

[2] http://www.fao.org/worldfoodsituation/foodpricesindex/en/

[3] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1233058

[4] http://www.earth-policy.org/images/uploads/book_images/Chapter8_Notes.pdf

Er is een dataset onderzocht met de voedselprijzen per markt in verschillende landen in Afrika, Azië, Noord- en Zuid-Amerika per maand sinds 1992. Aan de hand van deze dataset zijn er een aantal vragen opgesteld.
Zijn er voedselprijzen die een negatieve/positieve correlatie vertonen met elkaar en is deze correlatie aanwezig over de jaren heen of alleen in een bepaalde periode? In sommige gevallen kunnen veranderingen in de prijs van een bepaald product de prijs van een ander product beïnvloeden, doordat er verschuivingen plaatsvinden in vraag en aanbod. Het is interessant om te onderzoeken bij welke producten dit het geval is. Met deze informatie kan men aan de hand van productprijzen voorspellen of de vraag naar een ander product zal stijgen of dalen.
Een special soort correlatie is de correlatie tussen de ingrediënten en het eindproduct. Het lijkt logisch dat als de prijs van de ingrediënten omhoog gaat, dat de prijs van het eindproduct beïnvloed wordt. Een voorbeeld hiervan is brood en granen.
Is er bovendien een gelijkenis te vinden tussen de voedselprijzen van landen die bij elkaar in de buurt liggen? En zijn de verschillen te verklaren? Er zijn heel veel verschillende factoren die voedselprijzen kunnen beïnvloeden, zoals burgeroorlogen of politieke onrust en de hoeveelheid voedsel die een land importeert. Om de invloed van deze factoren vast te stellen, zouden de productprijzen van landen die dicht bij elkaar liggen, vergeleken kunnen worden. Als er weinig invloed is van deze factoren, dan zouden de prijzen dicht bij elkaar moeten liggen. Bovendien zouden de gelijkenissen mogelijkerwijs toe te schrijven zijn aan overeenkomsten in temperaturen in die regio's. Het weer is namelijk een grote factor in voedselproductie en dat beïnvloedt op zijn beurt weer de voedselprijs. Is er een correlatie te vinden tussen de gemiddelde temperaturen per maand in de landen en de voedselprijzen?

Om voorgenoemde vragen te beantwoorden is de behandelde dataset uitgebreid bestudeerd om alle relevante datapunten te isoleren en te analyseren. Wij veronderstellen dat de prijzen van vergelijkbare typen voedsel een correlatie vertonen, zoals brood en bepaalde soorten graan, aangezien dit het hoofdbestanddeel is. Daarnaast veronderstellen we dat de prijzen in bepaalde regio's, zoals Afrika, lager zullen liggen dan gemiddeld aangezien deze regio's merendeels uit ontwikkelingslanden bestaan.

2 Methode

2.1 Cleaning

De dataset bevat de prijzen van 331 verschillende producten uit 76 verschillende landen per maand vanaf 1992. Dit zijn landen gelegen in Noord- en Zuid-Amerika, Azië en Afrika. De dataset wordt elke maand opnieuw bijgewerkt. Niet elk land heeft data verstrekt vanaf hetzelfde jaar, daarom is er niet van elk land evenveel data. Ook verkopen niet alle landen dezelfde producten. Per land staan er de prijzen van de verschillende markten en de staat waarin de prijs geldt. De prijzen gaan over ongeveer 1500 verschillende markten. Daarnaast

staan ook de valuta, de meeteenheid en de hoeveelheid product waar de prijs op is gebaseerd weergegeven. Verder staat er voor elk gegeven in de dataset een ID-nummer en een bron.

Om te onderzoeken of de temperatuur invloed heeft op de voedselprijzen is er ook data over de temperatuur verzameld. Er is een dataset gevonden met de afwijkingen van de langjarige gemiddelde temperatuur per maand per werelddeel. [5] Het is niet gelukt om deze data per land te vinden, want alle preciezere *datasets* op internet hielden op rond 2013 en gaven alleen de jaarlijkse waarden weer. Er is dus gekozen voor een dataset met vijf verschillende regio's: Afrika, Azië, Noord-Amerika, Zuid-Amerika en Oceanië. De temperatuur-data is in tabellen uitgezet tegen de voedselprijzen om eventuele correlaties tussen deze data te kunnen analyseren.

Er waren geen *missing values* in de originele dataset met de voedselprijzen of waardes die heel erg buitengewoon waren. De dataset is opgesplitst in een aantal kleinere *datasets*, bestaande uit data over de zeven meest in de dataset voorkomende producten en brood. Brood is speciaal gekozen omdat veel van de meestvoorkomende producten ingrediënten kunnen zijn voor brood. Producten die hetzelfde zijn, maar een klein beetje variëren, zijn in dezelfde dataset gezet. *Rice* (*low quality*) en *Rice* (*imported*) zitten bijvoorbeeld in dezelfde dataset. Alle *ID's* en bronnen zijn verwijderd uit de dataset, want deze waren niet relevant voor het onderzoek.

Veel landen uit de *datasets* gebruiken verschillende meeteenheden voor hun producten. Zo stonden sommige producten er per megaton in, of juist per 0,7 kg. Dit hebben we allemaal genormaliseerd tot 1 KG. Dit is gedaan met een *python-dictionary* waar, voor iedere meeteenheid, in staat hoeveel kilogram het voorstelt. Deze *dictionary* is zelf ontwikkeld door op te zoeken hoe de meeteenheden zich verhouden tot een kilogram, waarna die waarden in de *dictionary* zijn gezet. Hierna is er een algoritme geschreven dat alle waarden in de datasetkolom omzette naar kilogram met behulp van deze *dictionary*.

De prijzen van verschillende markten per land staan in de dataset, echter zijn alleen de nationale gemiddelden belangrijk. Hiervoor is een algoritme geschreven waarbij de prijzen van de verschillende markten per land per datum bij elkaar zijn opgeteld en zijn gedeeld door het aantal markten. Hierdoor is de dataset kleiner geworden, omdat de markten bij elkaar gevoegd zijn. Dit was wenselijk, omdat de dataset zeer omvangrijk was en het veel rekentijd kostte om bepaalde operaties uit te voeren. Op deze manier werd de dataset overzichtelijk.

Daarnaast stonden de prijzen in verschillende valuta's. Er is voor gekozen deze waarden om te zetten naar de waarde van de US dollar zoals die was in juni 2018. Dit is wederom met een eigenhandig ontwikkelde *python-dictionary* gedaan, op gelijksoortige wijze als de normalisatie van de meeteenheden. Er is voor de US dollar gekozen omdat deze wordt gebruikt door een van de grootste economieën ter wereld en een relatief stabiele munteenheid is. Er is expliciet niet gekozen voor gebruik van de Euro als valuta omdat de data deels nog van voor de invoering van de euro is. Er is vanuit gegaan dat de *exchange rate* over

[5] `http://berkeleyearth.org/data/`

de jaren heen niet is veranderd, alhoewel dat in werkelijkheid wel het geval is geweest. Hier is vanwege de omvang van het onderzoek voor gekozen, vanwege de arbeidsintensiviteit van het verzamelen van deze *exchange rate* data.

Tot slot stond de data nog niet in zijn geheel in chronologische volgorde, dit is rechtgezet met een algoritme.

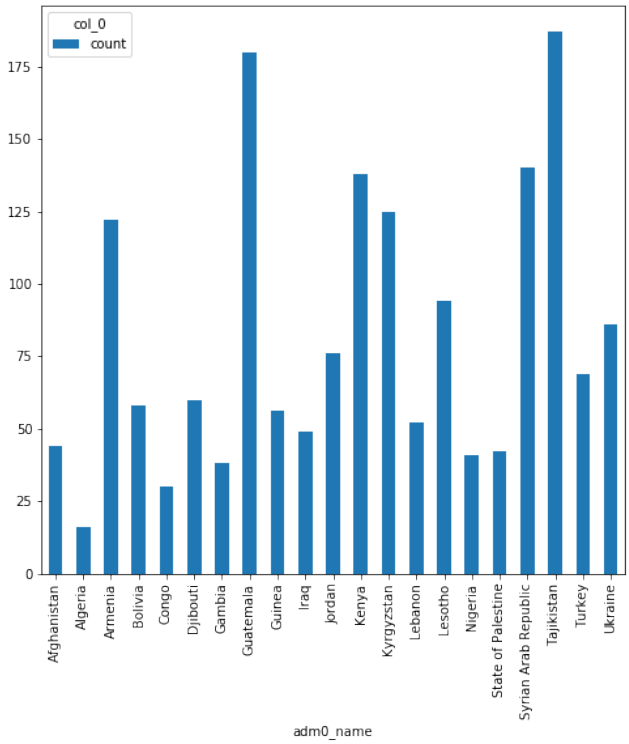
2.2 Exploratory Data Analysis

Voor de *univariate non-graphical* analyse zijn er tabellen opgesteld met frequenties van landen en jaren. Hierdoor werd meer overzicht gecreëerd over de inhoud van de dataset, aangezien het een zeer omvangrijke dataset is. Daarnaast zijn niet alle landen op hetzelfde moment gestart met data leveren, waardoor het nuttig was om de frequentie van elk land en elk jaar te meten, om vast te stellen waar de meeste data geconcentreerd is.

Voor de *multivariate non-graphical* analyse zijn kruistabellen gemaakt met meerdere variabelen, zoals jaren uitgezet tegen landen en productprijzen uitgezet tegen landen, om eventuele correlaties bloot te leggen en trends te ontdekken. Dit is gedaan om duidelijkheid te krijgen over het verband tussen de productprijzen en de tijd, en om een vergelijking te kunnen maken met andere landen en producten.

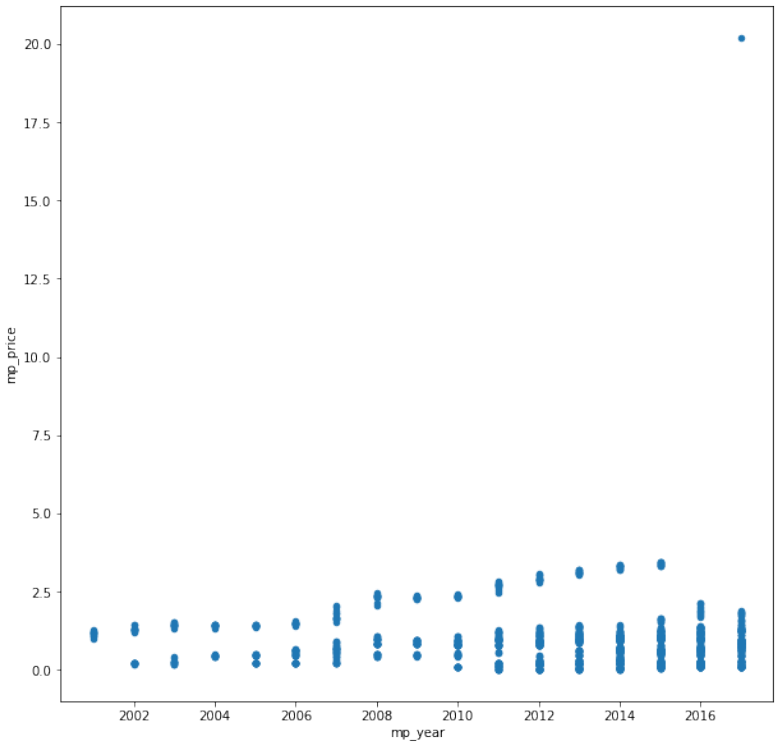
Voor de *univariate graphical* analyse zijn histogrammen, *density* plots en *bar* plots gemaakt om de frequentie van bepaalde gegevens in de dataset te visualiseren. Hierdoor was het makkelijker om in één opslag te zien hoe de frequenties zich verhielden tot elkaar. In de histogram hieronder is te zien hoe vaak elk land voorkwam per jaar in de dataset over brood. Hieruit kon afgeleid worden welk land de meeste data heeft geleverd over de verkoop van brood en welke landen er überhaupt brood verkochten.

Figure 1: Histogram frequentie van landen in brood dataset



Voor de *multivariate graphical* analyse zijn er box plots, *scatter* plots en *stacked bar* plots gemaakt. Hierdoor waren relaties tussen variabelen, of de afwezigheid hiervan, makkelijker in te zien. In de *scatter plot* hieronder is te zien hoe de prijzen van brood per land zich verhouden tot elkaar. Er is in een aantal landen een duidelijke stijging te zien in de prijs.

Figure 2: Prijzen van brood per land per jaar



Het idee was eerst om een plot te creëren met de prijzen van bepaalde voedselproducten in alle landen. Echter bleek dit lastig, omdat ook de datum van de prijs een factor is. Er is besloten om de prijzen per land te plotten, zodat de prijs op de y-as gezet kon worden en de datum op de x-as.

2.3 In depth analysis

Een methode van data analyse is regressie, echter kwamen er voor onze dataset geen relevante bevindingen uit. Bijna al onze plots vertoonden een duidelijke stijging. Daarnaast bevat de dataset slechts een enkele variabele, de prijs, wat de uitkomst van regressie minder nuttig maakt. *Clustering* was lastig aangezien er vooral non-numerieke data voorhanden was terwijl voor *clustering* numerieke data een vereiste is. Slechts de productprijzen zijn numeriek in deze datasets. Een manier om *clustering* toch toe te passen is om de coördinaten van de (hoofdsteden) van de landen te gebruiken, maar dit bracht teveel complicaties met zich mee met oog op de omvang van het onderzoek. Een andere manier is om *one hot encoding* te gebruiken. Hiermee word de categorische data omgezet naar numerieke data, zonder de data te vertekenen. Hiervoor is een algoritme geschreven dat alle landen tot het juiste continent toeschrijft. De vijf continenten die in de

dataset zitten werden met *one hot encoding* omgezet tot numerieke data waar *multivariate clustering* toegepast is. Er is besloten om dit per continent te doen en niet per land, vanwege het hoge aantal landen.

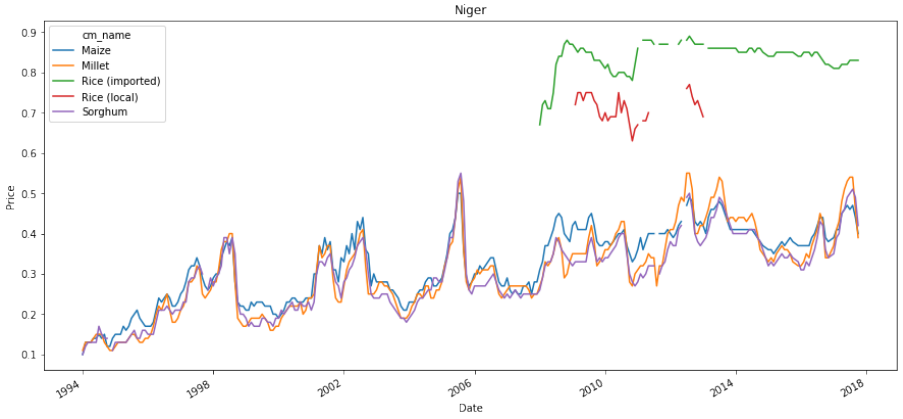
3 Resultaten

Wat direct opvalt uit de resultaten is dat de voedselprijzen bij een groot deel van de landen stijgen. Van de 76 geanalyseerde landen vertonen 52 landen een duidelijke stijging in voedselprijzen over tijd. Bij 22 landen is er sprake van een lichte daling of stijging. Bij twee landen is een duidelijke daling geconstateerd. Bij het beantwoorden van de volgende vragen wordt er dieper op de analyse ingegaan.

3.1 Zijn er voedselprijzen die een negatieve/positieve correlatie vertonen met elkaar en is deze correlatie aanwezig over de jaren heen of alleen in een bepaalde periode?

Bij de landen die de producten *millet*, *maize* en *sorghum* verbouwen en verkopen kwam duidelijk naar voren dat de schommelingen in de prijs overeenkomen. In Niger, Senegal en Mali zijn de prijzen zelfs nagenoeg gelijk (zie figuur drie).

Figure 3: Voedselprijzen van Niger



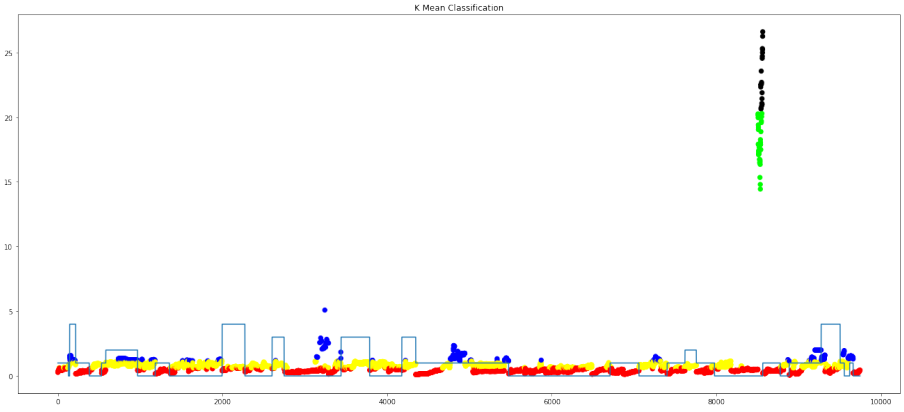
3.2 Is er een relatie te vinden tussen de prijzen van producten en de prijzen van de ingrediënten daarvan?

Er is in het bijzonder gekeken naar een mogelijke correlatie tussen brood en *wheat*, *sorghum* en *millet*, aangezien de laatste drie ingrediënten kunnen zijn van brood. Echter bleek er geen land in onze dataset aanwezig dat zowel data van brood als van de voorgenoemde ingrediënten had, waardoor geen conclusie getrokken kon worden over deze relatie. Zoals bij de vorige vraag te zien is, houden *wheat*, *sorghum* en *millet* onderling wel verband met elkaar, maar niet met het eindproduct.

3.3 Is er een gelijkenis te vinden tussen de voedselprijzen van landen die bij elkaar in de buurt liggen? En zijn de verschillen te verklaren?

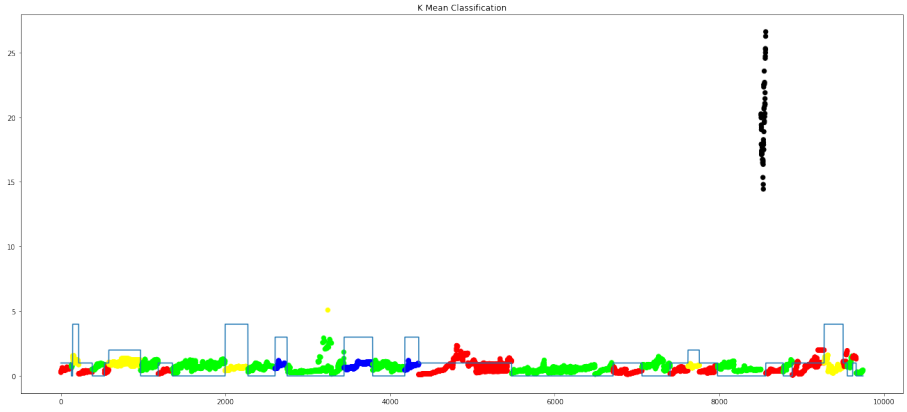
Het *k-means clustering* algoritme, getraind op alleen de prijzen kwam uit op de clusters zoals te zien in figuur vier.

Figure 4: Clustering zonder *one hot encoding*



Hierbij staat de index van ieder datapunt op de x-as en de prijs op de y-as. De blauwe lijn door de grafiek heen geeft aan op welk continent de prijs op die index ligt waarbij geldt: 0 = Afrika, 1 = Azië 2 = Noord-Amerika, 3 = Zuid-Amerika en 4 = Europa. *k-means clustering* is niet bedoeld voor *training sets* met een enkele variabele dus het resultaat viel te verwachten. Om meer variabelen te introduceren is de dataset bewerkt met *one hot encoding*. Zo kreeg ieder continent een cijfer toegewezen. De hierop getrainde dataset wist bijna ieder data punt aan het juiste continent toe te wijzen, zoals in figuur vijf te zien is.

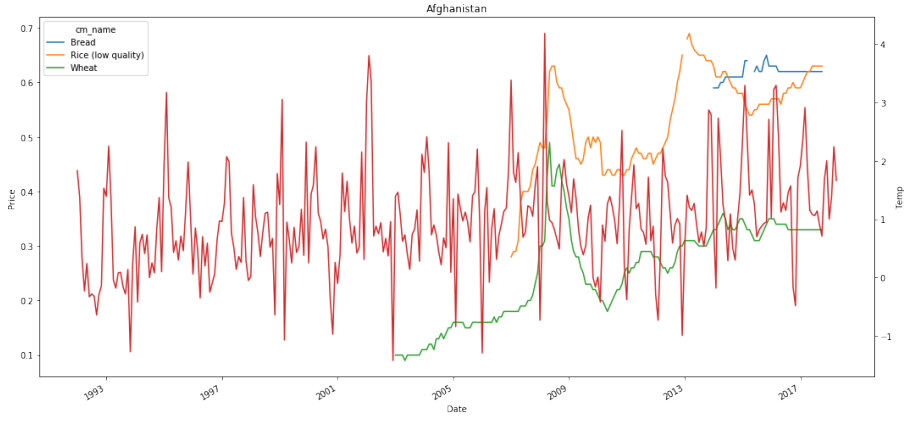
Figure 5: Clustering met *one hot encoding*



3.4 Is er een correlatie te vinden tussen de gemiddelde temperaturen per maand in de landen en de voedselprijzen?

Er is geen correlatie gevonden tussen temperatuur en voedselprijzen. De temperaturen per maand zijn uitgezet tegen de prijzen van bepaalde producten in landen. De temperatuur schommelde erg in alle landen, zoals verwacht, maar er was geen gelijkenis met de grafiek van de voedselprijzen. Hierom is er geconcludeerd dat er geen directe correlatie bestaat tussen temperatuur en voedselprijzen. Figuur zes illustreert dit goed; de temperatuur (rode lijn) fluctueert sterk terwijl dit niet correleert met de prijzen.

Figure 6: Temperatuur in Afghanistan



4 Discussie

Prijzen van graansoorten blijken een correlatie met elkaar te vertonen, wat in de lijn der verwachting ligt. In het bijzonder lijken de prijzen van *millet*, *sorghum* en *maize* erg dichtbij elkaar te liggen over de jaren heen. Aangezien de graansoorten groeien onder dezelfde weersomstandigheden, is dit te verwachten. Deze correlatie is te zien in de landen Mali, Niger en Senegal. Deze landen liggen dicht bij elkaar, waardoor ook geconcludeerd kan worden dat er een gelijkenis te vinden is tussen de voedselprijzen van landen die zich dicht bij elkaar bevinden. Verschillen in de prijzen van naburige landen zouden verklaard kunnen worden aan de hand van externe factoren zoals politieke onrust.

Uit het *clusteren* blijkt dat de prijzen in Afrika over het algemeen enigszins lager liggen dan op de andere continenten. Dit blijkt uit de blauwe lijn in figuur vier en vijf. Waar de lijn op de x-as ligt, zijn de in Afrika gelegen landen. Behalve dit gegeven valt er verder nauwelijks onderscheid te maken tussen verschillende continenten op basis van alleen de prijs. Dit betekent dat de voedselprijzen van de verschillende continenten niet te categoriseren zijn omdat ze allemaal dicht bij elkaar liggen.

Aangezien inflatie buiten beschouwing is gelaten, is het geen verrassing om de prijzen te zien stijgen. Voor preciezere conclusies die uit de data getrokken kunnen worden, had er rekening gehouden moeten worden met de veranderingen in de wisselkoers tussen de valuta van de landen en de US dollar. Als dat gebeurd was zouden de wisselingen in de prijs een accurater beeld schetsen van de ontwikkelingen op de markt. Het is nu namelijk niet vast te stellen of er daadwerkelijk een prijsstijging is geweest, waardoor er nu ook niet geconcludeerd kan worden of de stijgende temperatuur een factor hierin is geweest.

Over het algemeen is gebleken dat de voedselprijzen minder schommelen dan voorheen. Extreme pieken in de voedselprijzen van bepaalde landen zijn te verklaren aan de hand van gebeurtenissen in dat land, zo is bijvoorbeeld de voedselcrisis in Nigeria en de destabilisatie van delen van het land dankzij Boko Haram in de afgelopen jaren duidelijk terug te zien. Toch is deze data wellicht niet wenselijk bij een onderzoek naar de invloed van temperatuur op de voedselprijzen, dus deze data had gefilterd kunnen worden bij het vergelijken van de data over de temperatuur.

Daarnaast zijn enkele *outliers* niet weggefilterd, deze hebben mogelijk de *k-means clustering* beïnvloed. Om dit in een vervolgonderzoek te voorkomen verdient het aanbeveling om bij aanvang van het onderzoek een uitgebreidere *Exploratory Data Analysis* te verrichten. Daarnaast had meer aandacht besteed moeten worden aan regressie. Er zouden betere conclusies getrokken kunnen worden als de *mean squared error* beter was bestudeerd.

In dit onderzoek is gekeken naar een verband tussen temperatuur en voedselprijzen. Een onderzoek met een uitgebreidere database zou nog interessanter zijn. Zo is er in dit onderzoek geen rekening gehouden met neerslag in de landen van de dataset. Dit had tot relevantere conclusies kunnen leiden omdat neerslag, naast temperatuur, een belangrijke factor kan zijn in de totstandkoming van voedselprijzen. Daarnaast zaten in deze database vrijwel alleen maar on-

twikkelingslanden. Veel wereldeconomieën zoals de Verenigde Staten, China en Duitsland ontbraken. Wellicht kan de data van deze landen een beter inzicht geven in de verschillende aspecten die de prijzen van voedsel beïnvloeden.

5 Referenties

Geraadpleegd op 17-06-2018
http://berkeleyearth.org/data

Geraadpleegd op 19-06-2018
http://www.fao.org/worldfoodsituation/foodpricesindex/en/

Geraadpleegd op 19-06-2018
https://ourworldindata.org/food-prices

Geraadpleegd op 21-06-2018
https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/1630.pdf

Geraadpleegd op 25-06-2018
http://www.earth-policy.org/images/uploads/book_images/Chapter8_Notes.pdf

Geraadpleegd op 26-06-2018
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1233058

Geraadpleegd op 26-06-2018:
https://reliefweb.int/sites/reliefweb.int/files/resources/Nigeria%20-%20Food%20Security%20Outlook_%20Tue,%202014-04-01%20to%20Tue,%202014-09-30.pdf