University of Amsterdam

Final Report

Analysis of global food prices

June 28, 2018

Students:
Jesper van Duuren
10780793

Roel Klein
11693924

Tessel Wisman
11050519

Matthijs Bes
10667598

Group:
Group 40

Lecturer:
Gosia Migut, Nick de Wolf

Course:
Data Analysis and Visualization

Course code:
5082DAAV6Y

1 Introduction

The food price of individual products have a direct influence on which products are bought [3]. The choices of products which are bought affects human lives in various ways. One of the main effects in human health. People could be eating too much food or die from a lack of it. Still in the twenty-first century are people starving from a lack of food. The examination of food prices gives probably some insight in food-related problems, and in subject worthy study.

The Global Food Price dataset [6] will be analysed in this report. This dataset includes food price data of various products from 76 countries, divided over more than 1500 markets, with some countries having recordings from as early as 1992. Specifically, there will be an estimation about the correlations between individual products and prices, a food price correlation between countries in similar regions, and a estimation of a correlations between the production of food and the price of food. The data of the food production is a dataset from the Food and Agriculture Organization of the United Nations (FAO) [2]. In this dataset records are provided of crop production per year.

1.1 expectations

Various products in the food price dataset are a combination of raw products like bread. One of the main ingredients of bread is wheat flour, so if the price of wheat flour increases the cost of creating a bread should also increase. This is why we expect to find a strong positive correlation between a product and its ingredients. Thereby exists an expectation of a strong positive correlation between food products and its direct related product like white and black beans.

On regional level prices are determined by a large number of factors. Climate, conflicts, culture are all examples of these factors which influence food prices. So we expect to find a positive correlations between food prices and their location.

The correlations between nearby countries are also expected to be mainly positive, since importing products from countries in the area is relatively cheap. Major differences between prices would lead to increased import, dropping the price in the more expensive country and forcing it to lower self-produced product prices to stay competitive. This kind of correlation is expected to have gotten stronger and to have it's reach increased over time, since the transport of goods has gotten cheaper, especially via airplanes. More short-term, sudden increases in price in a region could be explained by for example natural disasters or droughts. Furthermore, we expect to find negative correlations between the amount of production and the price of a product. This hypothesis is based on the law of supply and demand —citation needed—, where the price of a product decreases if the production increases, assuming the demand stays the same (which it obviously didn't, this hypothesis needs fine-tuning).

2 Method

2.1 Preprocessing WFP-dataset

Before performing the data analysis the WFP-data has been preprocessed in a number of ways. This section discusses steps that have been taken prior to analysis.

2.1.1 Non-foods

The dataset appeared to contain quite a few entries describing the prices of non-food products such as fuel, wage and exchange rates. Because the focus of this project lies on global food prices, these entries were deleted from the dataset. The full list of entries can be found in Appendix A.

2.1.2 Unit normalization

In different countries at different markets, product prices in the dataset were often measured at different units. For example, the price of bread may be measured in kg in Afghanistan, while Guatemala uses Pounds and Lesotho has pricedata per loaf. To be able to analyze product prices over different countries, these units were normalized where possible to standard metric units kg, L and unit. Some units, like 'Bunch', '100 Tubers', 'Marmite' and 'Package' that are not standardized have been kept. Countries that use these unconvertable units could not be used in analysis over broader regions.

2.1.3 Handling missing entries

In cases where certain month entries for a product were missing, these were interpolated between the previous and the following month. Only years that contained data for all twelve months after interpolation were taken into account in further analysis.

2.1.4 Taking averages

To answer the third subquestion the foodprices data should be compared to the production dataset. Because the production dataset only contains entries for each year, the food data had to be converted to a similar format in order to be able to compute correlations and regression models. Therefore, for each country, district, market, product and year the average was computed over all twelve months. Next, the country average for all districts and markets was taken for each product to produce a dataset of the same dimension as the production dataset.

2.2 Preprocessing crop production dataset

The production dataset did not require as much preprocessing as the price dataset. The production dataset also contained columns with yield, which would not be necessary and were removed. All entries where the production was missing were removed as well. The amount of entries removed was less than 1%, which is why there was no technique used to interpolate data.It was confirmed that the same unit was used in the entire dataset, so there was no need for unit conversion. The tricky part was to link products in the production dataset to the price dataset, since in many cases they used different product names for similar products. This is why the production dataset was first reduced to only include countries that were in the price dataset. After that, lists of products available were made for each dataset. By searching through these lists, the similar products were manually linked and written to a csv file, which was used to be able to connect both datasets.

2.3 Exploratory Data Analysis

After preprocessing the dataset, an Exploratory Data Analysis (EDA) has been made. An EDA provides general information about the data. It is used to detect mistakes, check if assumptions line up with the data and detect relationships in the data.

2.3.1 Univariate non-graphical EDA

The first step in the data analysis was to calculate the basic statistics. For these purposes, Python functions were written to calculate statistics about a subset of the data, which can be extracted by queries. The functions return mean prices, standard deviations, and the number of entries that correspond to the query.

In order to find out which countries have (un)stable foodprices, an analysis was made on the standard deviation per country per product. The standard deviation should provide information about the distribution of entries. Generally, a higher standard deviation means a larger spread of entries applies. However, since the product prices are given per county in their own currency, the standard deviation appeared higher for currencies that have a lower exchange rate. For example, in September 2003, 5 KG of wheat costed about 9 Indian Rupees, but about 150 Nigerian Naira. A standard deviation of 5 would then be much more significant for Rupees than it would be for Naira. This was solved by dividing the standard deviation by the mean price in the country.

Furthermore, an analysis was made on the amount of available data for different products and countries by feeding the dataset different queries and checking for the amount of records the queries produced.

Lastly, a function was coded that detected outliers for a certain threshold. This threshold was a multiple of the standard deviation. So, if the threshold is 2, then all data points that are outside of the range $(mean - 2 * stdev, mean + 2 * stdev)$ are classified as outliers.

2.3.2 Univariate graphical EDA

The next step in the EDA was to visualize the data. To do this, functions were written that plot price over time, as well as production over time. The plotting functions again work with a query that subtracts a subset of data. An issue that was faced in plotting price data, was that a query might result in data for multiple currencies. As explained in section 2.2.1, different currencies have different absolute price ranges. This is problematic when plotting the price in different currencies in one plot. The y-axis should have a different scale for different currencies. That's why the logic of this plotting function was extended, so that it detects the amount of different products and currencies that are extracted by the query it received. Based on this, it displays the plots in different ways. Figure 16 and figure 17 illustrate this. In figure 2, a correlation is visible between the products "beans (red)" and "beans (silk red)". Spotting relationships like this one is the purpose of the graphical EDA. In the in-depth analysis, such relationships are explored more deeply.

2.3.3 Multivariate graphical EDA

Now that data can be plotted for both price- and production history seperately, the next . For the multivariate graphical EDA, a function has been written that plots the production numbers and price in the same plot, using different scales for the y-axis. With this, we can easily spot potential correlations between production numbers and price of the corresponding products. In figure 1, the example of Wheat in India is shown.

Figure 1: Example plot: Price vs production for wheat in India

2.4 Detecting correlation

In order to answer the subquestions, the next step has been to detect possible correlations between product prices in regions, correlations between the prices of different products and correlations between food prices and food production. To analyze product prices in regions, a visual analysis of the mean changes in foodprices has been implemented. Correlation between prices of different products and correlation between food production and price have been computed using Spearman rank correlation.

2.4.1 Correlation amongst related food groups

The search for correlations in product prices was automated by checking every single possible product pair. After data cleaning, there were 288 total different products, giving $\binom{288}{2} = 41328$ possible product pairs. The first step was to calculate Spearman rank correlations for all these product pairs, to get a quick overview of potentialy related products.

To eliminate the effect a difference in economies of different countries might have on the calculation of the Spearman's correlation, the calculations were only done on price data from the same country. To clarify by example: When calculating the correlation in price change between wheat and bread, no Spearman's correlation was calculated on price data from wheat in China and bread in the USA. Rather, two Spearman's correlations were calculated: the correlation of wheat in China and bread in China, and the correlation of wheat in the USA and bread in the USA. The mean value of these correlations were then taken as the final result.

At first, this algorithm was done on the percentile change of yearly average prices, and without any specified criteria on which calculations of Spearman's correlations would be

considered trustworthy. This resulted in 1077 product pairs having a Spearman's correlation of 1 (which is the maximum result). This result did not seem reliable, especially since most of these product pairs seemed unrelated to each other. (e.g. "Meat (Pork)" and "Mangoes" had a Spearman's correlation of 1)

It was clear that this result was not right. Criteria had to be defined in order to eliminate the calculations that were not accurate. The decision was made to only count Spearman's correlations as valid if the p-value was below 0.05, which makes such a correlation statistically relevant. This still resulted in 425 product pairs with a Spearman's correlation of 1, and again most did not seem to make sense. Thus, the restrictions for trustworthy calculation had to be sharpened even more. The next restriction that was added was to eliminate the markets for which a product pair had less than 5 years of overlapping data. This now resulted in 292 pairs for which 21 pairs had a Spearman's correlation of 1.

Still, it seemed odd that there were 21 pairs with the maximum Spearman's correlation. Looking into these pairs, it was noted that most of them were sold at different markets within countries. However, the data used for the calculation was the change of the yearly average price per country (so differences in months and markets were lost). In order to get an even greater precision, it was decided that the monthly data per market had to be used, calculating a Spearman's correlation for each market that had at least 5 years of overlapping data. Running the same algorithm again on this dataset resulted in a more reliable list of related products. However, when looking into this list, it was noted that the correlations were mostly from product pairs that only had overlap in one single country. It was the intention to also find correlations that were present in multiple countries. For this reason, the algorithm was extended once again with 2 parameters: *minYears* and *minCountries*, specifying the minimum amount of years of data a market should have in order for it to be valid, and the minimum amount of different countries in which there should be valid markets in order for a product pair's Spearman's correlation to be valid. The return value of the function was also extended. It now also returned the list of countries from which data was used to calculate the correlation, as well as the number of countries in that list. This was done to generate a CSV file that contained the correlations for product pairs, in which it was possible to filter on the amount of countries used. The final version of this function is shown in Appendix C.

2.4.2 Correlation in geographical regions

To detect correlation between food prices in geographical regions, the mean changes in food prices as computed in the EDA have been plotted onto a map plot, visualizing the changes at all markets. These plots have been analyzed manually looking at similar price changes at markets that lie near to eachother. Results of this analysis can be found in section 3.2.

2.4.3 Correlation between production and price

The third subquestion revolves around the correlation between the production values of a certain product and its price. To detect these correlations, Spearman rank correlation was computed between the change in production value and the change in price for each product in the food dataset. This has been performed for both production-price correlations in individual countries, as well as in broader regions.

First, equal (or very similar) products in the foodprice- and production dataset were detected manually. For each product, every year entry that was available in the production dataset was then aligned with a corresponding entry in the food dataset, to make it possible to directly relate production in a specific year to price. Correlation was then computed between the price values and the production values if 10 or more overlapping entries were available.

After detecting production-price correlations in specific countries, an attempt was also made to detect correlations between production and price in broader regions. These correlations had to be based on price change and production change data, for absolute values could not be used because of the different currencies available. All countries in the dataset were manually divided in five regions: Africa, the Middle-Eastern and Central Asian countries, Asia, Europe countries and South-America. The list of countries for each region can be found in Appendix B. For these regions, Spearman correlation was computed between aligned production entries and food price entries for each available product.

2.5 Linear Regression

In instances where significant correlations were found, a linear regression model was computed to describe and predict the food prices from their production values. This has been performed for both significant correlations in specific countries as well as correlations in regions, using the sklearn LinearRegression() module. A linear model was computed from a randomized training set that consisted of 80% of the data, and was tested against the 20% remaining testing data. For each prediction, mean squared error and variance was computed. Because the limited amount of available data the linear models were suspected to be quite unstable, mean squared error and variance results are presented as the mean values out of 5000 randomized model computations.

2.5.1 Clustering

To automatically detect groups of countries that show similar product-price relations, as well as to extract correlated food groups, k-means clustering has been considered. To examine the potential of this method, production change data was plotted against price change data. This was done for every specific year for every specific product, so that potential clusters would indicate groups of countries that showed the same production-price change. However, no such cluster forming was visible. This is partly due to the limited amount of available data. Because this line of research did not seem fruitful, further implementation of clustering algorithms has been omitted from the analysis.

3 Results

3.1 Correlation between products

To find correlations between products, the function *calc_product_correlation()* was used. (see Appendix C and section 2.3.1) The function was run for all product pairs, and for different sets of parameters. The top ten negative- and positive correlations for the different parameter configurations are listed in the tables of Appendix E.

3.1.1 Configuration 1: minYears=5, minCountries=1

The results of this configuration are listed in Appendix E.1 It stands out that for most of the top ten correlations, there was only 1 country that had overlapping data for at least 5 years. A possible explanation for this might be that the most extreme values are more likely to be found for data pairs that are present in only a small number of countries. After all, the mean Spearman's correlation values are taken from all available markets. If the number of markets is low, then the average value has been taken from a smaller set of data, making it more likely to be extreme then if the mean was taken from a large set of data. Another possible explanation may be that there are simply not many product pairs for which there are multiple countries that have at least 5 years of overlapping data.

Another thing that stands out, is that a lot of these highly correlated product pairs are found from markets in Congo. To explore what is going on, a plot has been made

of all price data from Congo in Figure 2. Looking at the figure, it is immediately clear that these product correlations have much more to do with developments in Congo, than actual correlations between the product prices. From December 2013 to January 2014, all prices drop immensely. On December 30, 2013, Congolese youth attacked several targets in the capital, Kinshasha, including the airport, a military barracks and state television headquarters. [5] Perhaps this uproar is related to the price drops in the country. In may 2016, there was also an enormous price drop, after which the prices shot right back up. In this month, there were protests again. [4] Whether this tumultuous political climate actually had a direct influence on these rapid price changes, is out of scope for this research.

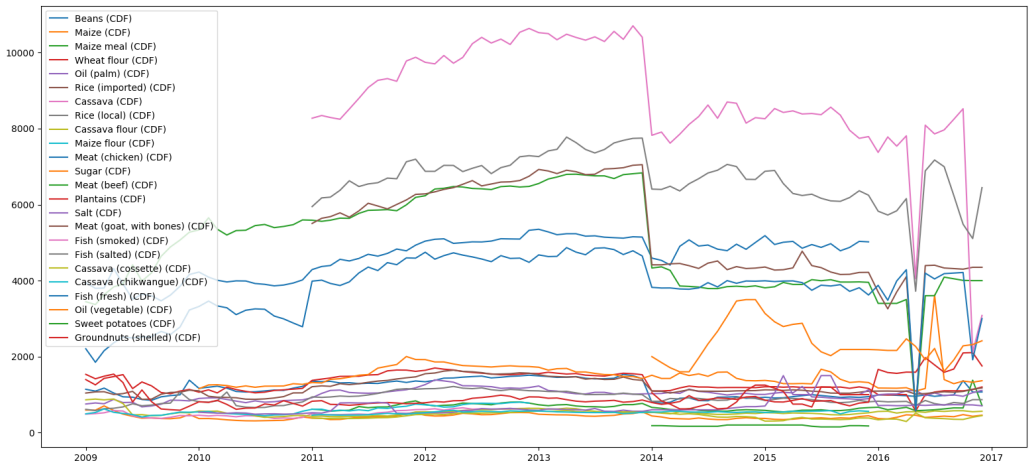


Figure 2: Plot for all data of Congo

One correlation is found between a product and its ingredient: Maize/Cornstarch in Ivory Coast. Most other positive correlations are between products that are very similar to each other (e.g. "Groundnuts (large, shelled)"/"Groundnuts (small, shelled)"

It could also be noted that for the negative correlations, there are some products/country pairs that appear multiple times in the top 10. Those are cassava (chikwangue) in Congo, peas in Colombia, and spinach in Rwanda. It is suspected that these products showed a price development that was diverging from the norm in their corresponding countries, making them appear in multiple spots in the top ten.

3.1.2 Configuration 2: minYears=5, minCountries=3

Since setting the minCountries parameter to 1 mainly gave correlations within single countries, and at most in two countries, it was decided to set the minCountries parameter to 3 in order to find new correlations that appear in more countries. In Appendix E.2 the resulting top tens can be found.

In the top ten of positive correlations, most product pairs seem to make sense. The odd ones out are probably Milk/Oil (vegetable) and Meat (chicken)/Plantains. These two pairs have in common that their data comes from multiple continents.

There are two correlations with significantly more countries than the others. These are the pairs Sorghum/Millet (12 countries) and Sorghum/Maize (9 countries). Sorghum, millet and maize are all grain type crops, and the data of the correlations all originate from countries in Africa. In Figure 3, the price history of these 3 grains are shown for the currencies of the 12 African countries.

In the top ten of negative correlations, it is noted that all found correlations have $\rho > -0.5$, which means the connections are less than moderately strong. It can be concluded that the negative correlations are less strong than the positive ones.

Figure 3: Sorghum, Millet and Maize for African currencies

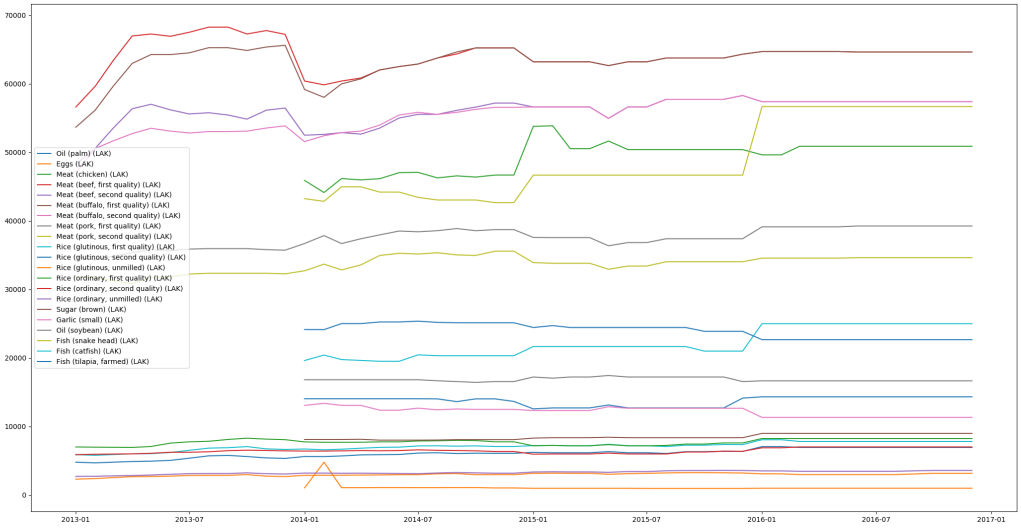
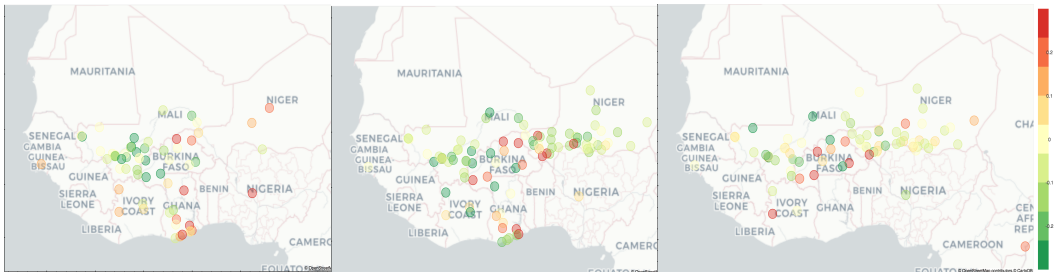
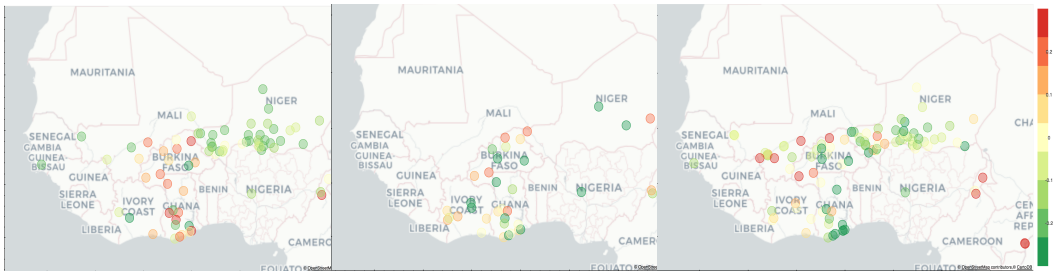


Figure 4: Plot for all data of Laos



(a) 2008 (b) 2009 (c) 2010



(a) 2011 (b) 2012 (c) 2013

3.1.3 Configuration 3: minYears=3, minCountries=1

For configuration 1, a lot of correlations came from a single country, most of them from Congo. For configuration 3, almost all correlations come from data from Laos. To again see what was going on, a plot was made from all data in Laos. it is shown in Figure 4. In this case, the figure shows exactly three years of data for most of the products. In contrast to Congo, however, the prices are quite stable. Combining the results of configuration 1 and configuration 3, one could say that if no minimum amount of countries is specified for which a correlation should be present, then the best correlations come from some single country. It thus seems that product prices depend on the political-economical situation of a country rather than on the price of related products.

3.2 Correlation in regions

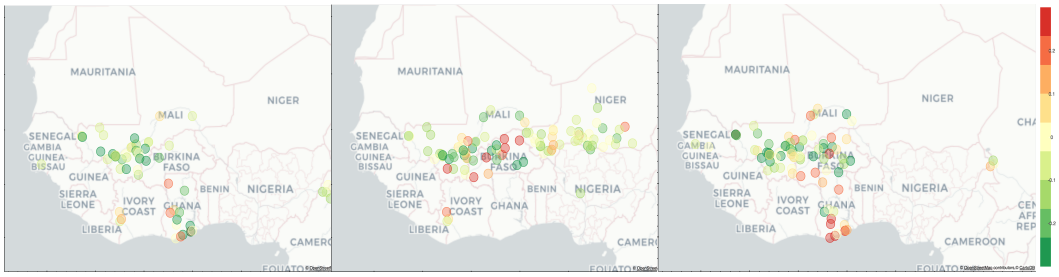
This section discusses some interesting phenomena that were discovered in the region-correlation analysis. While many more visualized data has been made available, the topics discussed below have been selected on significance.

3.2.1 Price change clustering in West-Africa

Figure 7 shows the price changes of Maize in the West-African region. In this instance, red dots indicate an increase of the mean food prices on a certain market, while green dots indicate a price decrease.
Over the years 2008-2016, some clustering can be seen in the price fluctuations. For example in Burkina Faso, price increase in 2011 (which can be related to the Burkinabé protests [1]) appears in all markets in the country and corresponds to price fluctuations in Ghana. However, the significance of the perceived correlations is doubtful, mainly due to the fact that not all markets provide consistent data over all years.

3.2.2 Price change clustering observations

For analysing price change over regions, a web application was made to provide geographical data. Only the main crops like sugar and rice have been made due to technical issues.



(a) 2014 (b) 2015 (c) 2016

Figure 7: Yearly food price changes in the West-African region

Normally food price change per year lies within a limit of -0.3% and 0.3%. What follows is a list of observations about these plots.

- Maize prices in 2006 at the southeast part of Africa shows a smooth transition in price from decreasing in Malawi to increasing in Tanzania. There is a clear cluster of food price changes going on.
- Maize prices in the western part of Africa shows a clear border between Burkina Faso and Mali.
- Maize prices in the central part of Africa shows a clear transition between location where prices increases and decreases.
- Wheat prices in 2009 show a clear cluster of food prices which are incresing in Ethiopia, and surrounding places where foodprices are dropping.
- Wheat prices in 2014 in India shows a transition from coast where prices are increasing to mountains where prices are just slihtly incresing.
- Bread prices in 2013 in Kyrgyztan shows a clear increase.
- Rice prices in 2009 in Nepal show a interesting pattern. On the south side of Nepal the prices are dropping but on the northern part the prices are increasing. This may caused by the presence of the Hymalayas.
- Sugar prices in 2011 show a very nice contrast between the central part of Africa and India.
- Sugar in 2014 in Madagascar shows a little disturbing pattern. Some markets have increased their prices while most dropped them.

3.2.3 Regional Price Changes Conclusion

Do foodprices show a regional pattern was te question to investigate. Therefore a visual representation of foodprices has been made. There has not been any computational procress been executed, but there is a list of observations which have been made. This list gives a clearer indication if foodprices show a regional patter. There has to be noted that only the price changes are observed and not the actual prices. The observation list shows a significant amount of observations which shows clustering and transitions between areas. This both support the hypothesis that foodprices show a regional pattern. A couple of observations note however a disturbance. This disturbance could have many explanations, but it is not in line with the hypothesis.

3.3 Correlation between production and price

This section contains the results of the correlation analysis between food production and food prices, computed as described in section 2.2.3.

3.3.1 Discovered correlations in specific countries

Listed in figure 8 below are the countries for which a significant correlation was found between the production of a certain product and its retail price. A correlation has been regarded significant if it $\rho > 0.5$ (which indicates a moderately strong correlation) or $\rho < -0.5$ and its p-value lies below 0.05 (statistically significant).

Country	Product	Correlation	p-value
Niger	Sorghum	0.81	8.958357547402335e-06
Burkina Faso	Maize	0.73	0.00319061631898815
Nepal	Wheat	0.84	0.0001597377591745287
Tajikistan	Cabbage	0.89	1.998540422478126e-05
Tajikistan	Carrots	0.9	9.558659401055547e-06
Tajikistan	Maize	0.84	0.00018639737783200425
Tajikistan	Potatoes	0.92	2.981543161456545e-06
Tajikistan	Wheat	0.67	0.00932299909684176
Guatemala	Maize (white)	0.84	0.0003802080801851202
Guatemala	Maize (yellow)	0.85	0.00026601205575225274
Mali	Maize	0.65	0.015348516732105476
Chad	Maize (white)	0.86	0.0003316683391269209
Kenya	Beans (dry)	0.86	0.0006116938500931706
Kenya	Maize (white)	0.68	0.020842853610945056
Kenya	Sorghum	0.74	0.009759535959916903
Peru	Potatoes	0.84	0.0013331850799508562
Tajikistan	Onions	0.85	0.0010451821458586869
Zambia	Maize (white)	0.8	0.0031104283103858483
Ethiopia	Maize (white)	0.82	0.0038149200825507135
Ethiopia	Wheat	0.9	0.00034361219776328223
Indonesia	Chili (green)	0.88	0.0008138621117322101
Peru	Maize (local)	0.84	0.0022200312259168407

Figure 8: Correlation between production and price

What immediately catches the eye is that correlation between production and price for these countries are exclusively positive: a higher production corresponds with a higher product price. This contradicts the posed hypothesis that production increase leads to lower food prices. Another notable fact is that of these countries that show significant correlation, four out of thirteen show this correlation for more than one food group. Tajikistan even shows production price correlation in five different food groups, which might be considered significant given the overall limited amount of correlations found.

3.3.2 Discovered correlations in regions

Correlation between food production and food price has also been computed over broader regions. Significant results are listed in figure 9.

Region	Product	Correlation	p-value
Europe	Apples	0.94	0.0050975405648614755
Asia	Cabbages and other brassicas	-0.84	0.0360575728451592
Asia	Chick peas	-0.85	0.016197127467871632
Asia	Lentils	-0.66	0.00424028798199035

Figure 9: Significant correlations in bigger regions

In contrary to the correlations found in countries, three out of four regional correlations appear to be negative: that is, the food price drops as production increases. Possible explanations for this will be further reviewed in the discussion.

3.4 Linear Regression

Products that showed a significant correlation between production value and food price were submitted to a linear regression algorithm that attempts to predict food prices based on production value. The predictions that were made by the regression algorithm are visualized below for the three countries for which most data was available in figure 12. Additionally, figure 10 shows the mean squared error and variance for each prediction. Because the limited datapoints to which the regression model could be fitted, these statistics have been computed by taking the mean over 5000 trials, using randomized train- and testsets.

Country	Product	Mean Squared Error	Variance
Tajikistan	Onions	0.2243447492698076	-2.5566134074214997
Tajikistan	Wheat	0.5301440225518627	-7.6160706493757955
Tajikistan	Carrots	0.12060605756343437	-8.711328083021032
Peru	Potatoes	0.07407684666224529	-3.333685582481241
BurkinaFaso	Maize	412.06192756717064	-7.662080065394659
Kenya	Sorghum	102.36317341350814	-4.028441745114724
Ethiopia	Maize (white)	1.3425864859366448	-9.827434117016175
Niger	Sorghum	1371.6183758933291	-0.28048950535495143
Guatemala	Maize (yellow)	0.15197373954580212	-0.24100778916492005
Indonesia	Chili (green)	21338148.230479505	-23.74596549257457
Nepal	Wheat	30.331512721505884	-0.3088900519237505
Ethiopia	Wheat	2.119972871577983	-87.64708380183816
Peru	Maize (local)	0.14105371847614662	-9500.115330657189
Kenya	Maize (white)	134.37085211066622	-5.579201103740688
Mali	Maize	569.8293767746771	-11.971001003026506
Tajikistan	Potatoes	0.11161551086356312	-1.0482580698875357
Guatemala	Maize (white)	0.15676806303134266	-3.8257401139367633
Chad	Maize (white)	1101.4858478721144	-2.1009200857271666
Zambia	Maize (white)	0.0847067279258514	-0.9704336876530637
Kenya	Beans (dry)	175.25783702750314	-2.0182644982306805
India	Wheat	5.020952463316244	0.6360269864731807
Tajikistan	Cabbage	0.3235316713136579	-3.2212557949821377
Tajikistan	Maize	0.31215450973253	-12.060406300975695

Figure 10: Mean Squared Error and Variance for production-price correlations in countries

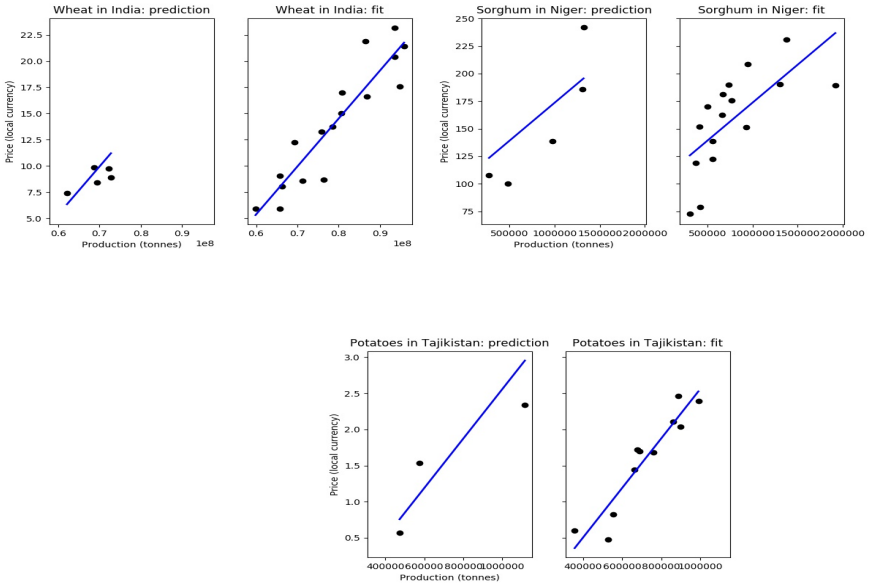
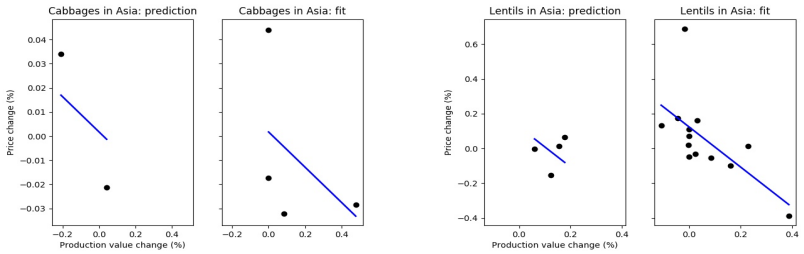


Figure 12: Linear Regression plots for most stable correlations in countries

In these example figures we can see that linear regression succeeds in detecting trends in foodprices based on the production data. However, these predictions are based on very limited data which makes them very subsceptible to certain changes and outliers.This is is further reviewed in the discussion. Another notable thing is that mean squared error comes in very different orders of magnitude for different countries. This can be explained by the fact that the absolute price value lies higher in certain countries that use certain currencies.

Region	**Product**	**Mean Squared Error**	**Variance**
Asia	Lentils	0.017669184688202156	-6.6560768232473695
Europe	Apples	0.02603550192197396	-29154.05498147092
Asia	Chickpeas	0.025472802042346875	-180.1937979101613
Asia	Cabbages	0.10251216681128178	-28.398571406399178

Figure 13: Mean Squared Error and Variance for production-price correlations in regions



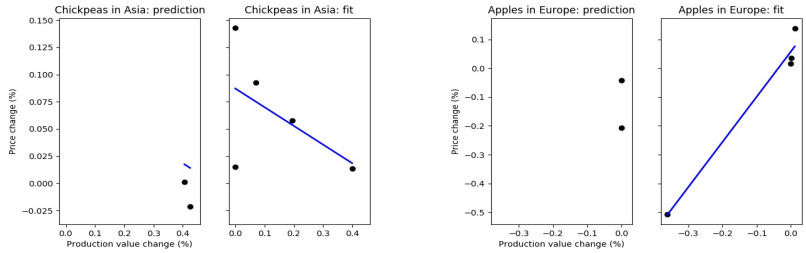


Figure 15: Linear Regression plots for regions

The regression models in figure 15 are based on even fewer data points than the regression models for individual countries. The fact that significant correlation was only detected for these small samples indicates an unstable relationship between production and prices over regions. The performance of the regression model is limited as well due to this lack of training data.

4 Discussion

4.1 Correlation between products

The hypothesis was that product prices show correlations with the prices of their ingredients (e.g. bread and wheat). One was found: the correlation between maize and cornstarch in Ivory Coast. However, most of the strongest correlations were between products that are similar to each other, rather than ingredients of each other. This can be seen in pairs like "Beans (red)"/"Beans (silk red)", "Wheat flour (first grade)"/"Wheat flour (high grade)" etc.

Furthermore, it was noted that if no minimum amount of countries was specified in which the correlations were present, most of the strongest correlations came from a single country. In the case of a minimum of 3 years of data this was Laos, and in the case of a minimum of 5 years of data, this was Congo. Both countries showed very different patterns in their price history, Laos being very stable and Congo being very unstable. What the countries had in common, however, was that prices of different products within the country behaved similar to one another. From this, one may sense that the economical situation of a country may be a better indicator for price of a product in that country than the price of some similar product. In other words: perhaps the connection between economic situation and product prices may be stronger than the connection between prices of different products. For future research, it could be interesting to look into this.

A last conclusion that could be drawn is that there seem to be more strong positive correlations between product prices, than there are negative correlations. A reason for this could be that all food prices are influenced by some common factors like inflation, economy, policies for import and export, and trade agreements between countries.

4.2 Production price correlation

There appeared to be 22 country-product combinations where a significant correlation exists between the food production and the food price. However, other than hypothesized, this correlation is mainly positive instead of negative, which does not correspond to the theory of supply and demand.

There are two possible explainations for this phenomenon. First, production in a country does not necessarily influence price directly because a significant proportion of the produced goods might be exported. These exported goods have no direct influence on the product

price in the production country. Second, an increase in production can also indicate economy growth. The WFP dataset contains a large number of developement countries, which have experienced economy growth and wealth increase in the past decade. The wealthier the nation, the more people have to spend which might influence retail prices, while economy growth will be related to increased production. To confirm or discard these speculations, more research regarding the economies of these countries in relation to these correlations is necessary.

Correlation over regions appeared to be even less common than correlation in independent countries. This might be also be explained by import and export: when production in a country increases it might export more goods to neighbouring countries, which will lead to a price increase or decrease in another country that can not be related back to the production. However, in contrary to the correlations in countries, three out of four region-correlations are negative, which does relate to the supply-demand hypothesis. In the plots in figure 15 can be seen that the correlations that were found significant are again based on few datapoints. To be able to draw robust conclusions, more data should be made available to analyze. The dimensions of the food prices dataset had to be heavily reduced because the production dataset only contained entries for each year. Furthermore, the amount of missing entries in the food prices dataset was very large: few countries and product provided information over the full span of 24 years. Research regarding the production-price correlation should be repeated preferably using a production dataset that contains monthly entries or a food prices dataset containing more entry years. This will also increase the performance and the significance of the linear regression analysis.

References

[1] Bettina Engels. Different means of protest, same causes: popular struggles in burkina faso. *Review of African Political Economy*, 42(143):92–106, 2015.

[2] Food and Agriculture Organization of the United Nations (FAO). Crops production dataset.

[3] Simone A. French. Pricing effects on food choices. *The Journal of Nutrition*, 133(3):841S–843S, 2003.

[4] Al Jazeera. Dr congo protests against joseph kabila turn deadly.

[5] Reuters UK. Update 5-congo's army repels attacks in kinshasa, dozens killed.

[6] World Food Programme (WFP). Global food prices database.

Appendix A Deleted entries

Charcoal	Fuel (kerosine)
Cotton (kerosene)	Fuel (petrol-gasoline)
Exchange rate	Transport
Exchange rate (unoffical)	Wage (non-qualified labour)
Electricity	Wage (non-qualified labour, agricultural)
Fuel (diesel)	Wage (non-qualified labour, non-agricultural)
Fuel (gas)	Wage (qualified labour)

Appendix B Regions

Europe : Armenia, Georgia, Turkey, Ukraine
Middle-East and Central Asia: Afghanistan, Azerbaijan, Lebanon, Iran (Islamic Republic of), Iraq, Jordan, Syrian Arab Republic, Yemen, State of Palestine, Kyrgyzstan, Tajikistan
Asia: Bangladesh, Cambodia, India, Indonesia, Lao People's Democratic Republic, Myanmar, Nepal, Pakistan, Philippines, Sri Lanka, Timor-Leste
Africa: Benin, Central African Republic, Chad, Congo, Djibouti, Cameroon, Burkina Faso, Cape Verde, Cote d'Ivoire, Democratic Republic of the Congo, Ethiopia, Gambia, Ghana, Guinea-Bissau, Guinea, Kenya, Madagascar, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Rwanda, Senegal, Somalia, Swaziland, Uganda, United Republic of Tanzania, Zambia, Zimbabwe, Sudan, Egypt, South Sudan, Burundi, Liberia, Lesotho
South America: Bolivia, Colombia, Costa Rica, El Salvador, Guatemala, Haiti, Honduras, Panama, Peru

Appendix C Spearman correlation's algorithm

```
def calc_product_correlation(df, prod1, prod2, minYears, minCountries):
    spearmans = []
    prod1_df = get_data_selection(df, products=[prod1])
    prod2_df = get_data_selection(df, products=[prod2])
    markets = overlap_in_markets(prod1_df, prod2_df)
    countries = overlap_in_countries(prod1_df, prod2_df)

    #Only return a spearman's correlation if enough countries have data on it
    if(len(countries) >= minCountries):
        for market in markets:
            market_prod1_df = get_data_selection(prod1_df, markets=[market])
            market_prod2_df = get_data_selection(prod2_df, markets=[market])
            years = overlap_in_years(market_prod1_df, market_prod2_df)

            #Only use markets that have enough years of data
            if(len(years) >= minYears):
                year_prod1_df = get_data_selection(market_prod1_df, years=years)
                year_prod2_df = get_data_selection(market_prod2_df, years=years)
                spearman, p_value = spearmanr(year_prod1_df['price_change'],
                    year_prod2_df['price_change'])

                #Only count spearman value if p-value is <=0.5
                if(p_value <= 0.05 and not(isnan(spearman))):
                    spearmans.append(spearman)
        return (np.mean(spearmans), countries, len(countries))
    else:
        return (np.nan, [], 0)
```

Appendix D Example EDA plots

D.1 Multiple currencies, 2 products

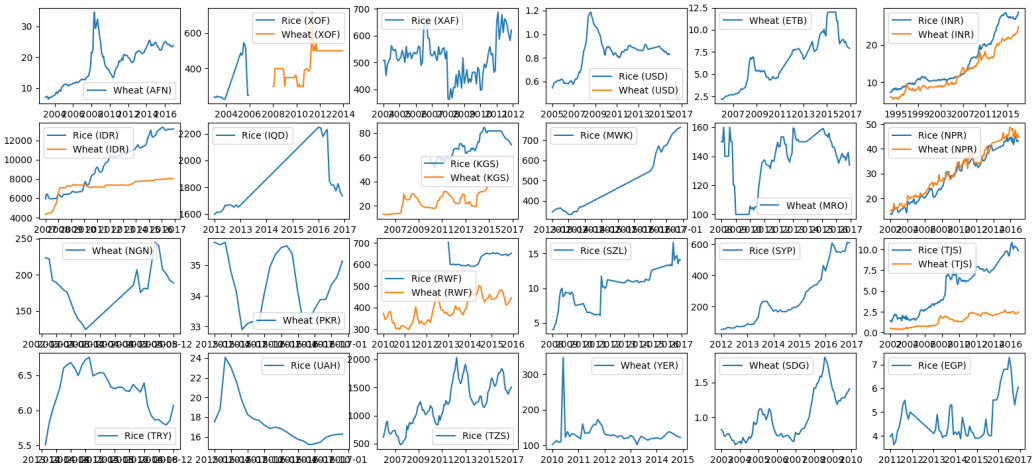


Figure 16: Example plot (query: product=="Rice" OR product=="Wheat")

D.2 Single currency, all products

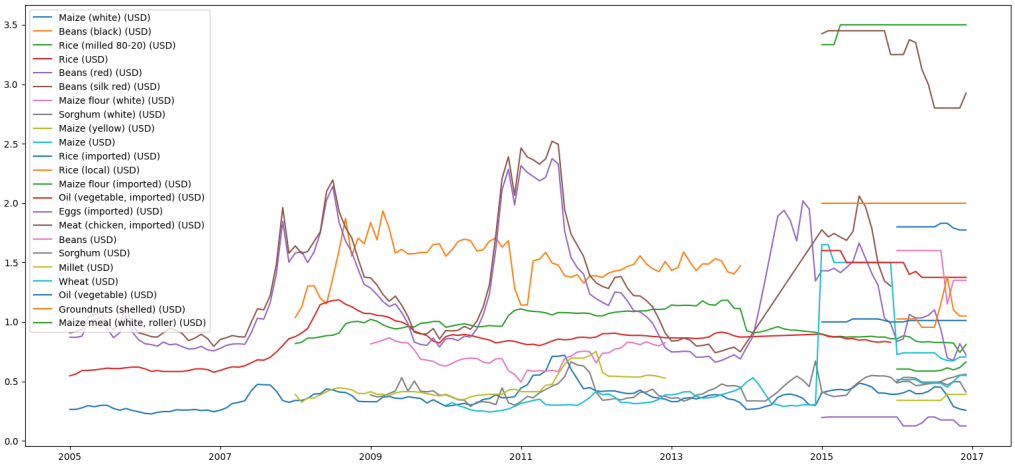


Figure 17: Example plot (query: currency=="USD")

Appendix E Spearman correlations tables

E.1 Configuration 1: minYears=5, minCountries=1

Table 1: Configuration 1: Positive correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Beans (red)	Beans (silk red)	0.8283471145	El Salvador
2.	Maize	Cornstarch	0.7655666439	Cote d'Ivoire
3.	Wheat flour (first grade)	Wheat flour (high quality)	0.7439557634	Kyrgyzstan, Tajikistan
4.	Beans	Fish (smoked)	0.7415213492	Democratic Republic of the Congo
5.	Beans	Cassava (cossette)	0.739014411	Democratic Republic of the Congo
6.	Groundnuts (large, shelled)	Groundnuts (small, shelled)	0.7169884034	Mozambique
7.	Cassava (cossette)	Plantains	0.7069642328	Democratic Republic of the Congo
8.	Cassava flour	Salt	0.7051941328	Democratic Republic of the Congo
9.	Maize (white)	Maize (yellow)	0.703545205	Colombia, Guatemala, Nigeria
10.	Cassava (cossette)	Fish (smoked)	0.7019014784	Democratic Republic of the Congo

Table 2: Configuration 1: Negative correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Beans	Cassava (chikwangue)	-0.8547056048	Democratic Republic of the Congo
2.	Plantains	Cassava (chikwangue)	-0.7483397708	Democratic Republic of the Congo
3.	Wheat flour	Peas (green, dry)	-0.6245418649	Colombia
4.	Chickpeas (imported)	Meat (beef, minced)	-0.6169819631	Colombia
5.	Spinach	Papaya	-0.5818436686	Colombia, Rwanda
6.	Oil (palm)	Fish (dry)	-0.5742511421	Cote d'Ivoire
7.	Cassava (cossette)	Rice (local)	-0.5370189699	Cameroon, Democratic Republic of the Congo
8.	Cassava flour	Spinach	-0.5099637091	Rwanda
9.	Fish (appolo)	Peanut	-0.5048363095	Cote d'Ivoire
10.	Maize	Spinach	-0.5005140462	Rwanda

E.2 Configuration 2: minYears=5, minCountries=3

Table 3: minYears=5, minCountires=3: Positive correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Maize (white)	Maize (yellow)	0.703545205	Colombia, Guatemala, Nigeria
2.	Onions (red)	Onions (white)	0.6863814159	Colombia, Egypt, Philippines, Rwanda
3.	Garlic	Onions (white)	0.6080469032	Colombia, Egypt, Philippines, Rwanda
4.	Oil (palm)	Plantains	0.59952273891	Congo, Cote d'Ivoire, Democratic Republic of the Congo
5.	Meat (beef)	Meat (mutton)	0.596852849	Gambia, Kyrgyzstan, Rwanda, Tajikistan
6.	Milk	Oil (vegetable)	0.5592121243	Armenia, Gambia, Guatemala, Tajikistan
7.	Wheat flour	Meat (chicken)	0.5566065047	Armenia, Colombia, Democratic Republic of the Congo, Turkey
8.	Meat (chicken)	Plantains	0.545303096	Colombia, Democratic Republic of the Congo, Guatemala
9.	Sorghum	Millet	0.5261361061	Benin, Burkina Faso, Gambia, Guinea-Bissau, Mali, Niger, Nigeria, Senegal, Sudan, Uganda, Zambia, Zimbabwe
10.	Sorghum	Maize	0.5126597114	Benin, Burkina Faso, Gambia, Guinea-Bissau, Mali, Niger, Nigeria, Rwanda, Zimbabwe

Table 4: minYears=5, minCountries=3: Negative correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Potatoes	Oil (sunflower)	-0.4543193482	Colombia, Egypt, India, Kyrgyzstan, Turkey, Ukraine
2.	Meat (beef)	Rice (imported)	-0.4470171008	Democratic Republic of the Congo, Guinea, Jordan
3.	Maize	Cassava (cossette)	-0.4465180844	Cameroon, Central African Republic, Democratic Republic of the Congo
4.	Wheat flour	Bananas	-0.415306023	Colombia, Rwanda, Turkey
5.	Oil (palm)	Rice (local)	-0.4006096958	Cote d'Ivoire, Democratic Republic of the Congo, Guinea, Guinea-Bissau, Nigeria
6.	Meat (chicken)	Maize	-0.375965546	Democratic Republic of the Congo, Gambia, Kyrgyzstan, Myanmar
7.	Tomatoes	Onions (white)	-0.3704187391	Colombia, Egypt, Philippines, Rwanda
8.	Cabbage	Oil (sunflower)	-0.3650731945	Colombia, Kyrgyzstan, Lebanon, Ukraine
9.	Wheat flour	Spinach	-0.3545249843	Colombia, Lebanon, Rwanda
10.	Maize	Sweet potatoes	-0.3122308646	Democratic Republic of the Congo, Gambia, Rwanda

E.3 Configuration 3: minYears=3, minCountries=1

Table 5: minYears=3, minCountires=1: Positive correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Meat (beef, first quality)	Meat (buffalo), first quality	0.9393959287	Laos
2.	Oil (palm)	Oil (soybean)	0.9055763063	Laos, India
3.	Meat (pork, first quality)	Meat (pork, second quality)	0.8882461714	Laos
4.	Meat (beef, first quality)	Meat (beef, second quality)	0.8737089164	Laos
5.	Beans (red)	Beans (silk red)	0.8283471145	El Salvador
6.	Sugar (brown)	Fish (snake head)	0.8248436851	Laos
7.	Rice (coarse)	Rice (basmati, broken)	0.8163238019	Pakistan
8.	Rice (ordinary, second quality)	Rice (ordinary, first quality)	0.7988958438	Guatemala, Laos
9.	Meat (beef, second quality)	Meat (buffalo, first quality)	0.7943483018	Laos
10.	Meat (beef, second quality)	Meat (buffalo, second quality)	0.7816749438	Laos

Table 6: minYears=3, minCountires=1: Negative correlations

	Product 1	**Product 2**	**Spearman's correlation**	**Countries**
1.	Fish (snake head)	Fish (catfish)	-0.9997705896	Laos
2.	Sugar (brown)	Fish (catfish)	-0.8266783749	Laos
3.	Rice (ordinary, second quality)	Meat (pork, first quality)	-0.8079430336	Laos
4.	Rice (ordinary, second quality)	Meat (pork, second quality)	-0.791571994	Laos
5.	Rice (coarse)	Ghee (artificial)	-0.7748339951	Pakistan
6.	Eggs	Sweet potatoes	-0.7274405741	Gambia, Philippines, Rwanda
7.	Eggs	Meat (buffalo, second quality)	-0.7065692187	Laos
8.	Meat (chicken)	Meat (buffalo, second quality)	-0.6996164505	Laos
9.	Meat (pork, second quality)	Rice (glutinous, second quality)	-0.6871430966	Laos
10.	Eggs	Meat (beef, second quality)	-0.6618496479	Laos