Crop Constellations : A comprehensive analysis of yield instability and pattern recognition among Top Producing Items in Global Agricultural Systems

Xavier MARTIN and Fabiola MARTINEZ

I – Presentation Methodology

This study aims to characterize the agricultural systems of countries by examining the yield instability of the top 10 most produced items from 1996 to 2018. One of the study's objectives, and an assumed bias, is to use a specific production item—namely, the umpteenth agricultural production of each country—as a descriptive variable for their agricultural policies.

This data was extracted from FAOSTAT records for the top 11 countries in the world. The 10 most produced items from these countries were determined by aggr egating production volumes in tons from 1996 to 2018.

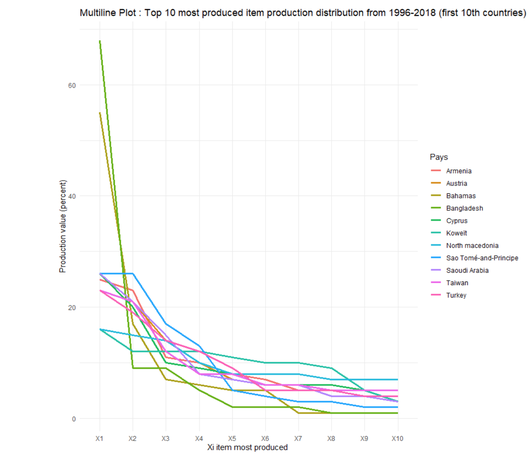


Figure 2 : Production distribution between the ten first items produced in a country. Bahamas and Bangladesh seems to have a strong unequal repartition production between crops while Kuwait and North Macedonia seems to have roughly equal repartition of their productions within the 10 most produced crops in the country.

Figure 1 : Representing 11 countries very diverse ; from practically 2 tones produced each year per inhabitants for Austria, to less than 200 kilos produced for Kuwait. In general terms, size, density, economy and human culture of this 11 countries there represent, more or less, world diversity

To ensure unique variables and make country data available for descriptive statistical analysis, two new variables were introduced after the 4th most produced item in each country. For each of these variables, the 5th, 6th and 7th umpteenth yields were calculated based on the centralized mean for each year. The same procedure was applied to the second set of variables, corresponding to the 8th, 9th, and 10th most produced items in the analyzed countries.

II – Description/Statistics

The eleven countries were chosen based of the presence of data of yield for their 4 most produced items. For the two new variables X_5_6_7 and X_8_9_10, countries were selected because they have at least one dataset of yield in the 5, 6, 7th most produced item of the countries. For example, Austria have in the variable X5_6_7, only its 7th items yield that is used because there is no data for skimmed cow's milk (its 5 one) and malted barley beer of yield (its 6 one) but there is yield data for barley, so that will be the only data used to make variable X_5_6_7 for Austria.

This mean yield of this diverse crops that we can show in fig3 is hiding a temporal series. In fact, each value that make each color of the fig3 is taken from the mean of a specific item's yield from 1980 to 2020. This specific evolution of the first column of the fig3 is shown on the fig4. The legend show for each curve country with its first production. The most produced item within the country in first national rank its cowmilk. It's regarding 4 countries that have been colored in thicken, continue and colored lines.

The value have been before centered to minimize crop-specific impact on data and get a result that reflecting how a first national produced crop is managed and the resulting yield of this management throughout the period.

A good fingerprint of this management can be the variation throughout the period. The variance from 1980 to 2020 of the centered yield have been taken for each of the item's shown in fig 3 to make a PCA that can assess the countries agricultural management.

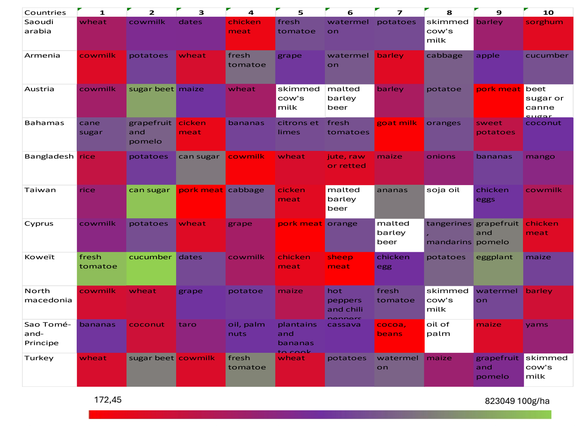


Figure 3 : Same values represented in a table with crop names. We can see that countries have very diverse crop cultivated but there is general trend to have cereals and milk in the first places for most of the countries.

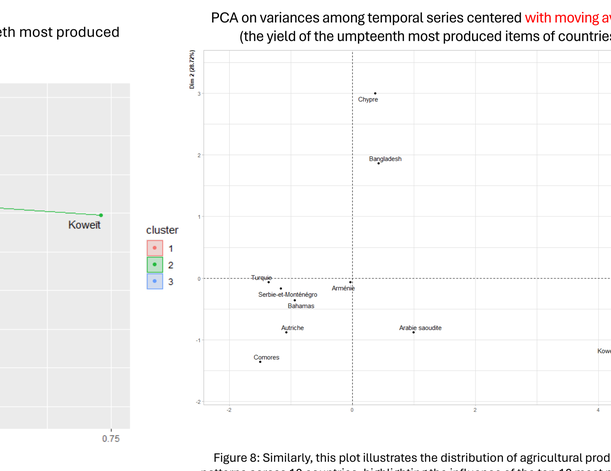


Figure 4: Variables time line of yield centered from 1980 to 2020. We can see that the productivity is increasing for all item umpteenth number

III. CORRELATION MATRIX AND PCA

Positive correlations are observed between X1 and X2 (0.35), X1 and X3 (0.31), and X2 and X3 (0.59). Negative correlations are present between X1 and X4 (-0.14), X1 and X5_6_7 (-0.05), and X1 and X8_9_10

Strong Correlations: the positive correlation between X2 and X3 (0.59) indicates a relatively strong relationship between these variables. The negative correlation between X1 and X8_9_10 (-0.17) suggests an inverse relationship.

Weak Correlations: X3 and X4, as well as X4 and X5_6_7, show relatively weak correlations (-0.20 and 0.01, respectively).

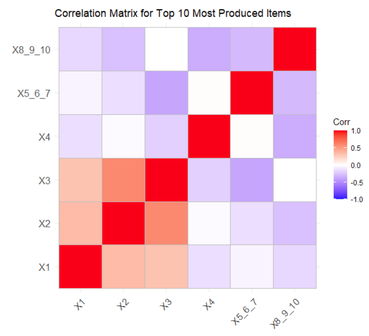


Figure 5 : Representation of the correlation matrix for the variance of the top 10 most produced items from 1996 to 2018

Figure 6: Variables contribution for each dimension

After conducting the PCA analysis, a comprehensive understanding emerges as follows:

Individuals: Representing the 11 countries, each delineated by their top 10 most produced items.

Dimensions: Dim.1 and Dim.2 play significant roles in the observed variance, accounting for 36.13% and 29.88%, respectively.

Variables Contribution:

- Countries making notable contributions to these dimensions include "Kuweit," "Bangladesh," and "China, Taiwan Province."
- "Kuweit" particularly stands out with a substantial contribution to Dim.1, indicating a distinctive pattern in agricultural production.
- "Bangladesh" makes a significant contribution to Dim.2, suggesting a pattern distinct from other countries.
- Variables like X3 exhibit high contributions, signifying their importance in shaping the observed patterns.

IV – DISCRIMINANT ANALYSIS AND CLUSTERING

Cluster Sizes:
Cluster 1: 2 countries
Cluster 2: 2 countries
Cluster 3: 7 countries

Percentage of Variance Explained:
The between-cluster sum of squares (between_SS) compared to the total sum of squares (total_SS) is 73.9%, suggesting a substantial amount of variance between the cluster.

Cluster 1 and Cluster 2 have fewer countries but display distinct patterns in their principal components (PC1 and PC2).

Cluster 3 is the largest, containing seven countries with a different pattern in the principal components.

The within-cluster sum of squares indicates how closely the countries within each cluster align with the cluster mean, with Cluster 1 having the lowest and Cluster 3 the highest.

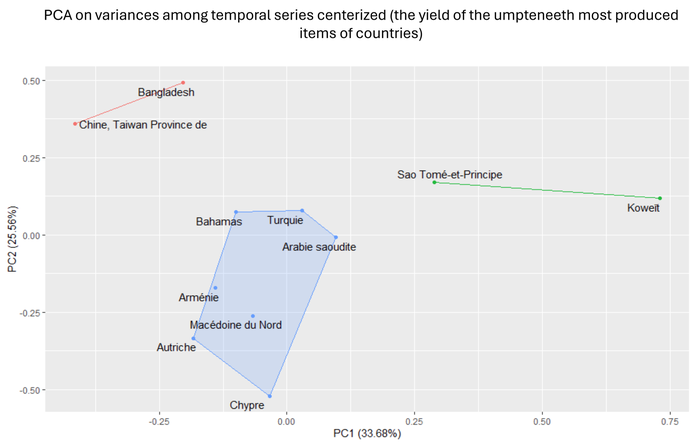


Figure 7: This plot illustrates the distribution of agricultural production patterns across 11 countries, highlighting the influence of the top 10 most produced items. Each point on the plot represents a country, with its position determined by its scores on Dim.1 and Dim.2, reflecting the variability captured by these principal components.
The clustering overlay adds further insight into the grouping of countries based on their agricultural production profiles. Three distinct clusters emerge, delineated by different colors or markers, indicating shared characteristics in production patterns within each cluster.

Figure 8: Similarly, this plot illustrates the distribution of agricultural production patterns across 10 countries, highlighting the influence of the top 10 most produced items. Each point on the plot represents a country, with its position determined by its scores on Dim.1 and Dim.2, reflecting the variability captured by these principal components.
The variances of yield was not directly catch among their centered yield for the period from 1980 to 2020, but the standardization by centralization of the yield of the temporal series was made on moving average of 8 years to hide global trend yield increase between 1980 to 2020 and get only the instability as fingerprinting of the country's agricultural system.

Figure 9: This figure represents the variance of the centered data from 1980 to 2020. This variance data have been calculated for the 4th crops mostly produced in each countries and then the variable of the data mean of the 3 next produced items of each countries have been used from 5 to 10th crops mostly produced

Figure 10: This figure is equivalent to the previous one. The difference is that the variance was made among the centered on the moving average of the temporal series of the yield.

Figure 11: Temporal series of the variable used to calculate the variance used in the PCA.

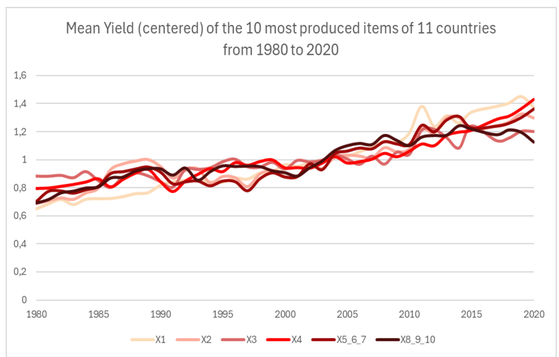
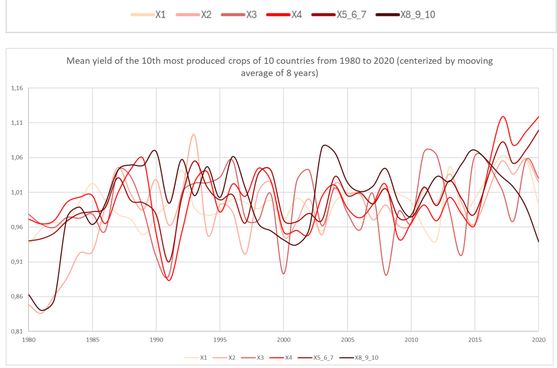


Figure 12: This figure is equivalent to the previous one. The difference is that the general trend was suppressed by the moving average instead of the general average.



V- RESULTS

Understanding the relationships between these dimensions and additional context about agricultural practices and policies can provide insights into the factors influencing yield instability. One of the way to understand the clustering is to see the climatic aspect of the represented countries. The cluster one with Bangladesh and Taiwan are respectively monsoon and chinese climate. These represent different climate but that have some similarities like the rain season. The cluster 2 is by far, the largely spaced of the three clusters. Saint Thomas and Prince and Kuweit have equatorial climate and desert climate respectively. Even if there are very different climate and they represent a cluster very spaced, they have both huge differences of yield variability among their culture from 1980 to 2020 maybe the extreme climatic weather create instability of yield among some crops that are very dependent of climatic conditions.

In the third cluster, there are no countries of these extreme climates (chinese, desert, equatorial and monsoon) except for Saudi Arabia. The presence of Saoudi Arabia in this cluster far away from Kuweit can be explained by a long history of agricultural policy among agriculture in Saudi Arabia since 1970. Indeed, the governement of the country wanted to reach food auto-sufficiency from this date, but that was officialy give-up in 2008 because of the lack of water. Nevertheless, Saoudia Arabia have invested far away before Koweït in huge technical machinery for irrigation and more widely in controlled agricultural systems. This can explain the huge separation between the two countries in PCA.

All of the countries in the cluster 3 seems to have also a long history of agriculture that can explains a more resilient and exhaustive approach of their agricultural policy among all crops cultivated in the country.

If we analyze the second PCA made on a most representative data of the yield instability, we see that there is a better separation of the data within the two first dimensions (73,41 % of the total variances).Additionally, the individuals seem to remain in similar positions. The individuals that were in the cluster 3 appears to be better separated from other individuals. Saudi Arabia seems to be nearer to Kuweit (as Saint Thomas and Prince do not remain in this dataset for comparaison) but there is still a huge gap between the two countries. As we can see in the data, Comoros have a very little general variances among their most produced items. This explains its presence within the country with strong agricultural stability as Austria, Bahamas, Serbia-and-Montenegro, Turkey and Armenia.

VI – CONCLUSION

Through an examination of yield instability in the top 10 most produced agricultural items from 1996 to 2018, the study effectively delineates the diverse agricultural systems prevalent in various countries. The data preparation demonstrates a robust approach, encompassing the aggregation of production volumes, the introduction of new variables, and meticulous selection criteria for both countries and variables. Principal Component Analysis (PCA) has been instrumental in uncovering significant dimensions of variance within the study, particularly highlighting Dim.1 and Dim.2 as pivotal in elucidating a substantial portion of the observed variability. This methodological approach has allowed for the identification of noteworthy contributions from specific countries and variables, thereby accentuating distinct patterns in agricultural production. Furthermore, the Cluster Analysis conducted in this study has yielded valuable discernment, revealing three distinct clusters of countries based on their agricultural production patterns. Through meticulous examination of cluster sizes, explained variance, and within-cluster sum of squares, the research has unveiled nuanced similarities and differences among these clusters.

In summary, this study provides a comprehensive understanding of agricultural systems across different countries, offering valuable implications for policy-making, resource allocation, and future research in the field of agriculture and food security. To go further in the analysis, it is possible to change the approach of a given umpteenth production of a country, but to work with group like we do for the two last variables. For example, we could make 6 groups of 4 items within the 24 most produced items of each country. It will enable us to include more countries by reporting each instance of instability when data for a particular item's yield is absent in the other items within the group. We can also add, to takes into account the different types of distributions between crops in the countries, the importance in % of tones (or % turnover) of the given item of the country to weight better instability values.

BIBLIOGRAPHY

FAOSTAT: Food and Agriculture Organization Statistical Database. https://www.fao.org/faostat/en/#data/QCL
Horikoshi, M. (ggfortify): Enhancing data visualization with ggplot2. https://cran.r-project.org/web/packages/ggfortify/
Husson, F., Josse, J., Le, S., & Mazet, J. (FactoMineR): Multivariate Exploratory Data Analysis and Data Mining with R. https://cran.r-project.org/web/packages/FactoMineR/
Kassambara, A. (factoextra): Extract and Visualize the Results of Multivariate Data Analyses. https://cran.r-project.org/web/packages/factoextra/