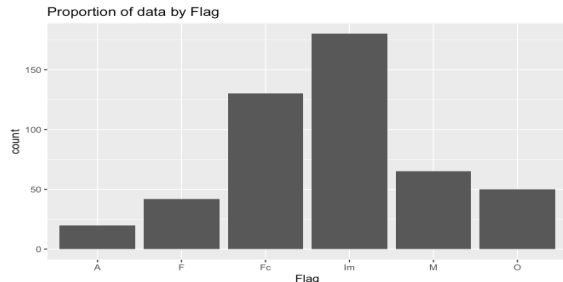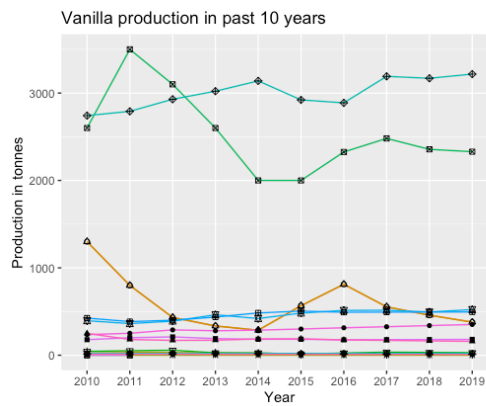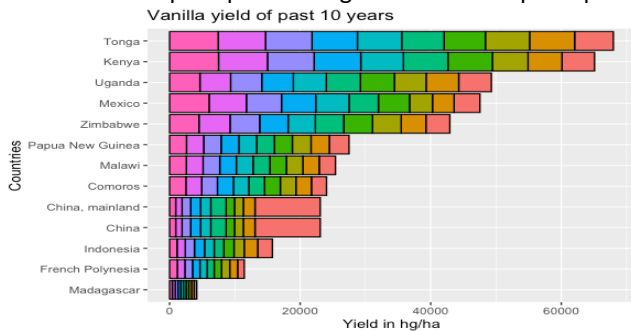# I – Presentation Methodology

Proportion of data by Flag

- We will seek to understand the nature of the data collected from the FAO website from a Data Mining perspective.
- Data: 'Area harvested', 'Production' and 'Yield' by 'Year' and 'Country' from 10 years of global Vanilla crop data.
- Official data 'O' is a very small fraction. The rest are filled in using 'Im' imputation models, 'Fc' yield calculations, aggregated 'A', FAO estimates 'F', and some contain missing 'M' flag.
- Yield stays consistent for the different countries. However, China sees a sharp drop indicating a shift in focus perhaps.

Vanilla yield of past 10 years

II – Statistics

Vanilla production in past 10 years

Production
- For most countries the vanilla production has remained fairly constant.
- However there is some fluctutation in Madagascar, Indonesia and mainland China.
- Madagascar has maximum average prod/year at about 50% of global production.
- Indonesia is 2nd at 14% and Tonga,Kenya,China, Mexico at 4% each.
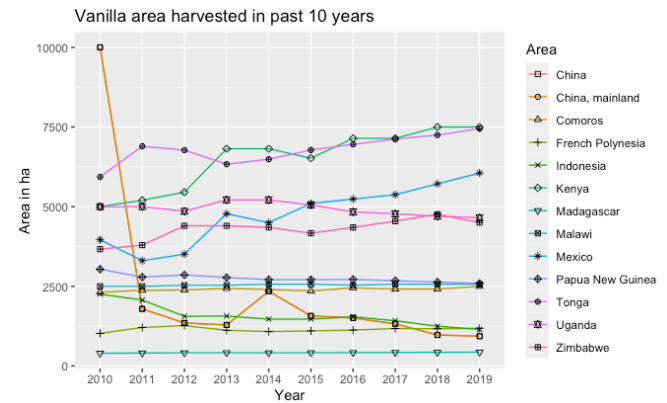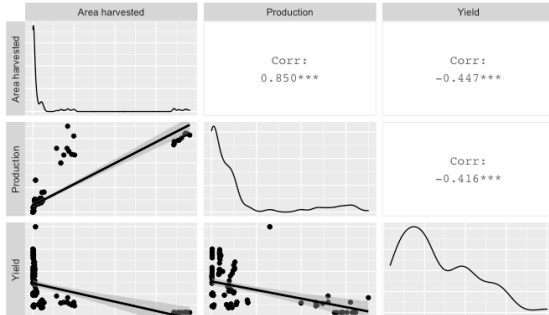
Area harvested
- We see that mainland China has recorded the sharpest drop in area harvested. This corresponds with the market drop in 2010 due to over supply.
- Most other countries have not seen much change over the last 10 years. Farmers become wary about volatile returns on their investment.
- However, Mexico and Kenya has seen a significant increase in this parameter.

Vanilla area harvested in past 10 years

III. CORRELATION MATRIX | MCA

Distibution of the variables and pair correlations

MCA factor map

Correlation and MCA
- Area harvested and Production have strong positive correlation. There is an exponential year on year increase in vanilla flowering till it reaches a stable level.
- Area harvested has a significant negative correlation with Yield.
- In the MCA we observe that dimension 1 and 2 account for 40% of the variance.
- Official data is most associated with Mexico, Indonesia, mainland China and French Polynesia
- Aggregated data is most associated with China.
- Calculated data 'Fc' is mainly for 'Yield' which is to be calculated for all countries.
- For countries which have some imputation model 'im' values, they are also likely to have FAO estimated values 'F'.
- 'M' missing values are most associated with Belgium, Seychelle, Fiji, and Portugal.
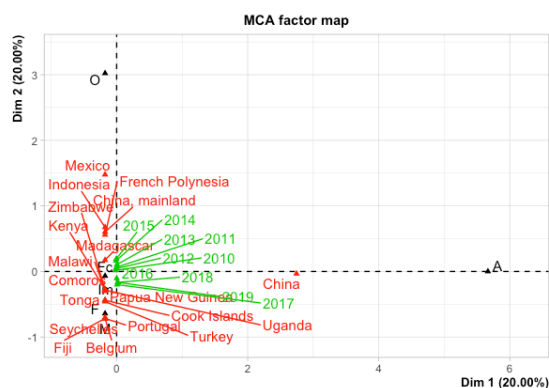
IV. HIERARCHICAL CLUSTERING

Factor map
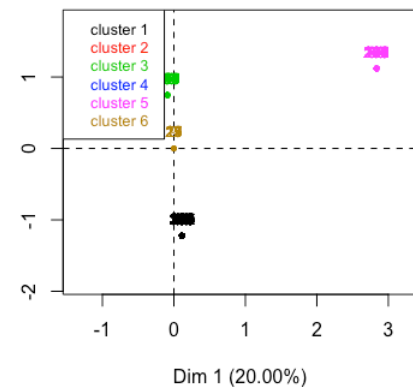
Hierarchical clustering
- We can consider 6 clusters and 5 axes to define the associations within the data and the flag types.
- The MCA interpretations observed in the previous section are properly classified in this model and we get some additional information as well.
- For example, flag 'F' (FAO estimates) seems most associated with the years 2010, 2012, and 2015.

V- Results

Mean of production, area harvested, and yield ,per year per country in the dataset are: 552 tonnes, 7081 ha, and 3287 hg/ha. There are 19 countries in the dataset.
Mean of the global sum of production, area harvested, and yield, per year in the dataset are: 8395 tonnes, 99130 ha, and 42732 hg/ha.
As of 2019, Madagascar is the top producer of Vanilla, followed by Indonesia and then Mexico. Countries with the best yield are Tonga, Kenya and Mexico. These are also the same countries with the most harvested area. Yield data may be unreliable because of scarcity and improper reporting from some countries.
Maximum amount of Official data is from Mexico, Indonesia, mainland China and French Polynesia; whereas missing data is from Belgium, Seychelle, Fiji, and Portugal.

Vi – Conclusion

This exercise gives us an insight into the nature of the data currently available on the FAO website. The different data flag types allow us to interpret the reliability and current data practices of different countries. We can observe that official data is still a very small part of the vanilla production and there is scope for implementing improved data collection methods. With the advent of big data and advanced image recognition algorithms, we can soon expect a drastic improvement in the type of data available and a more controlled (less volatile) production. This will open the doors for more jobs and additional income for farmers.

References

https://globalrisksinsights.com/2013/08/the-world-vanilla-market-is-anything-but-vanilla/
https://github.com/Appymp/Vanilla-Analysis
http://www.fao.org/faostat/en/#data/QC