World agricultural production analysis and country-level crop yield prediction

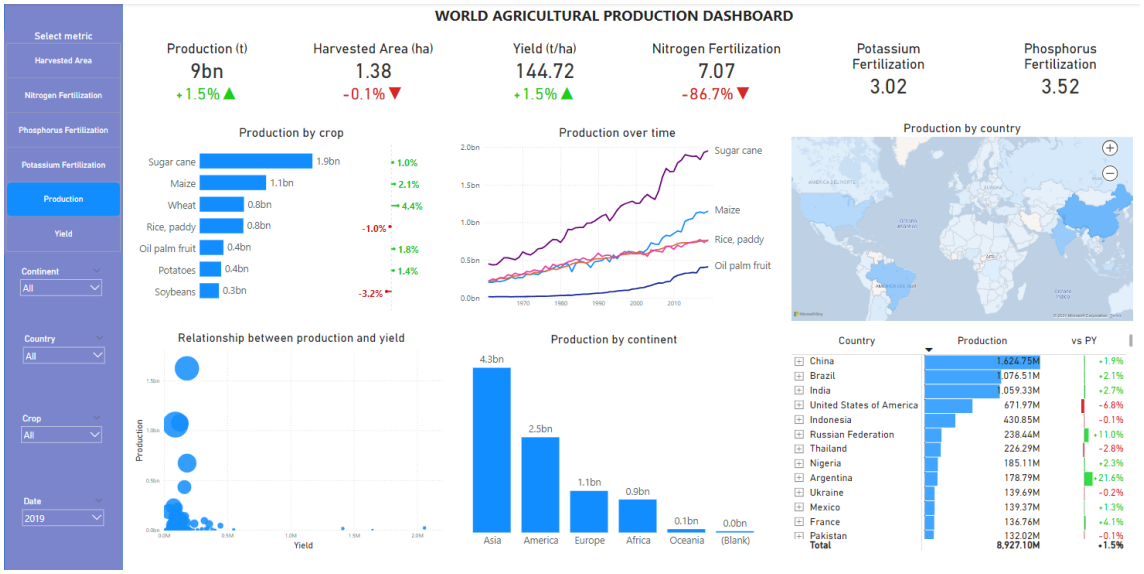
Author: Federico Garland Schiefer

OBJECTIVE

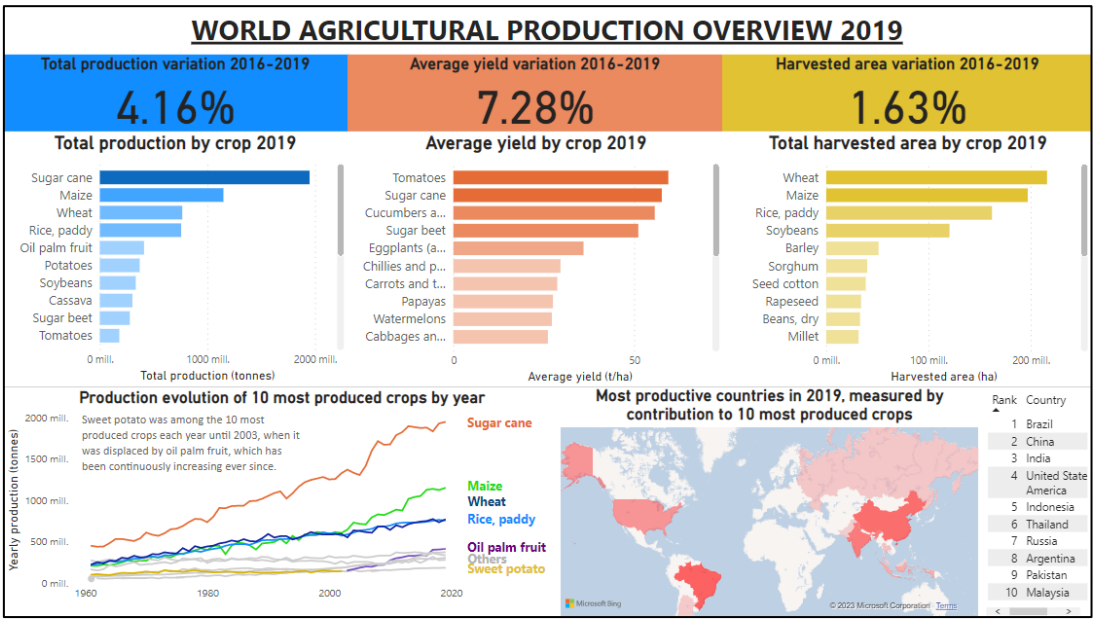
The objective of this project was to analyze world agricultural production to obtain useful insights for the decision-making process of agricultural businesses, farmers, investors, policymakers, researchers, and interested entities worldwide. The aim was to obtain relevant information such as the growth rate of worldwide agricultural production and it's two direct driving factors (yield and harvested area), identify the most important crops, time evolution of production by crop in the last 40 years, most influential countries, as well as generate crop-yield predictive models that allow for yields to be anticipated and for the most influential yield driving factors to be identified for proper management.

This project was first made in January 2023, and then a new improvement was made in June 2024 creating a fully automated ETL from the SQL database to Power BI, and creating an interactive dashboard using DAX. The 2023 dashboard was made for storytelling while the 2024 dashboard was made for user interactive use, in both cases in order to communicate key insights and drive impactful business action.

1. Interactive Dashboard 2024



2. Storytelling Dashboard 2023



METHODOLOGY

Tools used

- Excel
- SQL
- Power BI
- R

Process

- Dataset selection and download
- Database creation and data upload (SQL)
- Data cleaning (SQL, Excel and R)
- Data analysis (SQL)
- Dashboard design (Power BI)
- Creation of yield predictive models (R)

Detailed methodology description

1. **Data selection and download**

Seven different datasets were downloaded for the analysis. These datasets contained worldwide information at the country and yearly level on agricultural production (crops table), average temperature (temperature table), minimum and maximum temperature (minmaxtemperature table), rainfall (rainfall table), irrigated area (irrigation table), fertilizer use (fertilizer table), pesticide use (pesticides table) and tractor density (tractor

table). Crops table contained data up to year 2019, whereas the other tables varied in this regard. Crops and fertilizer tables also contained crop-level information. The sources are FAO, World Bank, International Fertilizer Association, among others. These datasets were found in Kaggle, World Bank website, scientific papers and others.

2. Data cleaning

All datasets were uploaded to a SQL database (using pgadmin, PostgreSQL) and were subjected to data cleaning to obtain tables with useful and interpretable data, and consistent keys that allow for them to be joined. Irrigation and tractor tables were first structured in R because the raw files had data displayed in rows instead of columns. Also, minmaxtemperature table required prior cleaning in Excel because temperatures where recorded as text in the format "- 13.2 °C", which had to be transformed to numeric in the form of "-13.2", and countries where recorded as " Algeria *" which had to be converted to "Algeria" using text functions such as LEFT, EXTRACT, FIND and LENGTH. After upload to SQL, the first data cleaning was conducted on the agricultural production dataset (named crops table on this project). For this, all rows containing groups of items instead of individual items where removed, since they would duplicate the data already contained in the individual item rows and distort general results. Therefore, items such as "Cereals, Total" (that contain the accumulated values of crops such as maize, wheat, etc. that are already individually present in the dataset), "Fibre Crops Primary", "Vegetables" and others where removed. Also, the item "Rice, paddy (rice milled equivalent)" was removed and "Rice, paddy" was utilized, given that the data of the latter is proportional to the values of the former with a relation of 1.49, with one being the net weight and the other the gross weight. Gross weight was chosen because all other crops are also measured in those terms. Furthermore, all areas comprised of groups of countries such as Africa, Europe, etc. were removed because they contain duplicate information already present in the individual country data that would distort the results.

For the tables temperature, minmaxtemperature, rainfall, pesticides, irrigation and tractor, countries that were written differently from the crops table but referred to the same nation were identified and re-written in order to be able to join the tables. As for the fertilizer table, both country and crop items having the aforementioned issue were re-written to match the crops table.

3. Data analysis

After data cleaning, several SQL queries were executed to obtain relevant information from the crops table. The results of the queries were summarized in an Excel book and a Power BI dashboard was designed to communicate the most important insights from world agricultural production up to year 2019. DAX functions were used to obtain the variation rates for global production, average yield and harvested area in the period 2016-2019.

4. Yield prediction models

Furthermore, the crops table data was joined with the data from all the other tables to obtain 4 different datasets displaying the all-crop, maize, wheat, and potato average yield for each country, as well as variables such as minimum, maximum and average

temperature, average yearly rainfall, average pesticide rate per hectare, total pesticide use by country, average nitrogen, phosphorus and potassium fertilizer rate per hectare (net dose and gross dose), irrigated area, tractor density, among others. The difference between fertilizer net dose and gross dose is that net dose is the total fertilizer use by country divided between the number of hectares that actually received fertilizers, while gross dose is the proportion between total fertilizer use by country and the total arable hectares, fertilized or not. These datasets were used to generate 4 multiple linear regression models for yield prediction. Stepwise regression was used to obtain the models, and residual analysis was conducted to modify and adjust the models to fulfill all regression assumptions (linearity, independence of residuals, normality of residuals, homoscedasticity and lack of multicollinearity among predictors). Assumptions were verified both graphically (using residual plots) and statistically with different hypothesis tests at the 0.05 level of significance. Independence of residuals was confirmed through Durbin-Watson test, normality of residuals through Shapiro-Wilk, homoscedasticity through Breusch-Pagan and linearity through the Residuals vs Fitted graph. Multicollinearity was evaluated with the Variance Inflation Factor (VIF). When one or more assumptions were not fulfilled, certain predictors were removed, added or transformed. Response variable (yield) was also transformed when necessary. Transformations used were log (natural logarithm), square root, potentiation and Box-Cox. Models were adjusted until all assumptions were satisfied and a significant model with $p < 0.05$, high R^2, low residual standard error and all significant predictors was obtained. Standardized β coefficients were computed for each model to rank the predictors by their degree of influence on yield. Finally, goodness of fit graphs (Measured vs Predicted) were produced to assess and communicate the precision of the models.

RESULTS

1. World agricultural production

Here, a few relevant insights from the analysis will be mentioned:

- In the period 2016-2019, total agricultural production grew 4.16%. It's two direct drivers, average yield and total harvested area, grew 7.28% and 1.63% respectively. This indicates that increasing yields are the main factor driving the growth of agricultural production.

- The 5 most produced crops in period 2019 were sugar cane, maize, wheat, rice and oil palm fruit.

- Sugar cane production has been steadily increasing for the last 40 years. Maize, wheat and rice increased at a similar rate until around year 2000, when maize started to increase its production at a higher rate.

- Sweet potato was among the 10 most produced crops each year between 1961 and 2002, until oil palm fruit displaced it in year 2003. Further on, oil palm fruit

continued to increase in production and rose to be the fifth most produced crop in year 2019.

- Among the 5 most produced crops of 2019. oil palm fruit has experienced the highest harvested area growth, with a 21% increase in hectares harvested between 2016 and 2019. On the other hand, it's yield only increased by 0.6% in the same period, which indicates that the increase of oil palm production is driven by more area being cultivated with the crop.

- On average, sugar cane was the highest yielding crop for 57 years (1961 to 2017), until tomato displaced it, becoming the highest yielding crop in years 2018 and 2019.

- On the other hand, sugar cane was only the 15th ranking crop sorted by harvested area. This along with the previous insight, reveal that the high worldwide sugar cane production is driven by high yields, and not by cultivated area.

- The 5 countries that contributed the most to the production of the 10 most produced crops worldwide in 2019 were Brazil, China, India, United States of America and Indonesia, in descending order. Among these, USA has the highest yields whereas China has the highest harvested area.

- Sugar cane is the most produced crop in Brazil and India, whereas maize is the most produced crop in China and the USA.

- Among these countries, USA has the highest maize yield (10.5 t/ha) and India has the highest sugar cane yield (80.1 t/ha).

- Globally, the 10 crops with the highest harvested area growth rates between 2016 and 2019 are mushrooms and truffles, hemp, anise, pyrethrum, karite nuts, seed cotton, okra, areca nuts, cashew and currants.

2. Yield prediction models

The next table summarizes the results from the 4 multiple linear regression models used to predict all-crop, maize, wheat and potato yields.

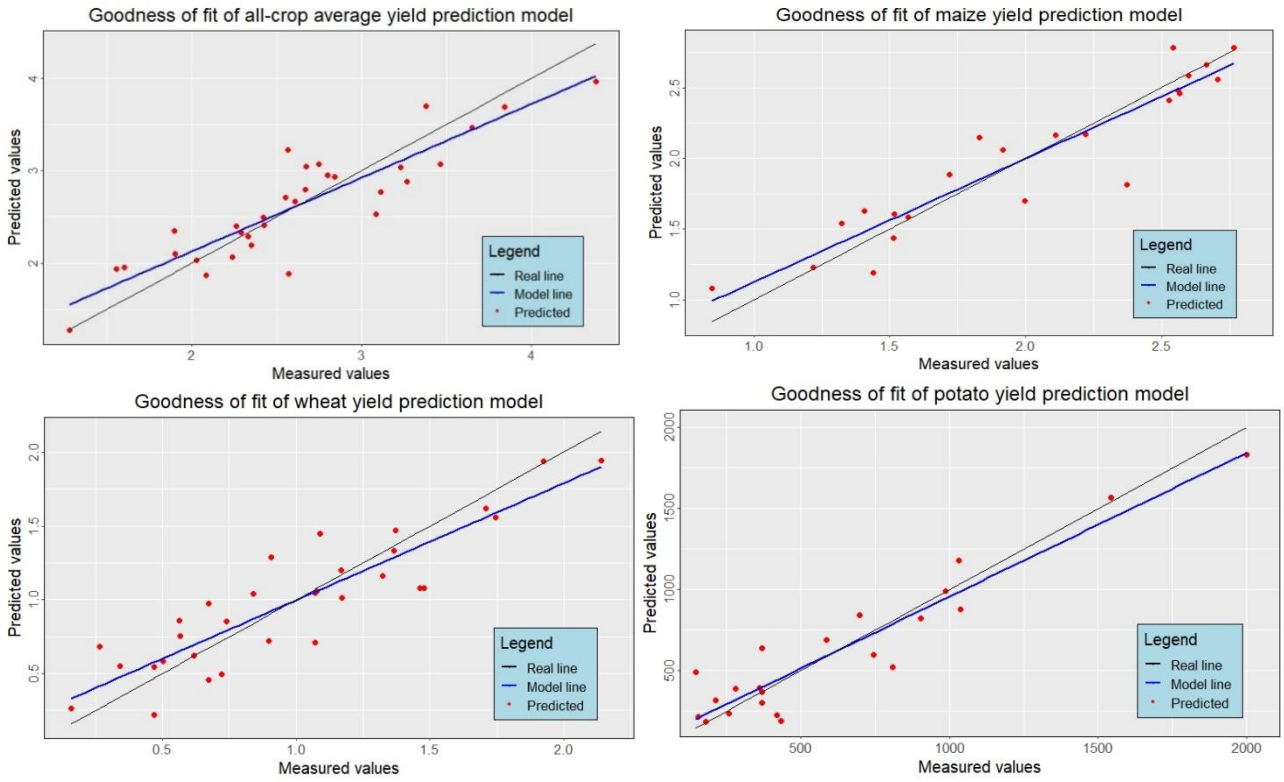
Model	R^2	p-value	Response (transformation)	Predictors (transformation) *
All-crop average yield	0.76	<0.0001	- Yield (log)	- Nitrogen fertilizer net dose (log) - Phosphorus fertilizer gross dose - Potassium fertilized area (squared) - Pesticide rate per hectare (square root) - Maximum temperature (log)
Maize yield	0.85	<0.0001	- Yield (square root)	- Nitrogen fertilizer net dose (log) - Average temperature - Tractor density - Total country-level pesticide use
Wheat yield	0.76	<0.0001	- Yield (log)	- Nitrogen fertilizer gross dose - Phosphorus fertilizer gross dose - Potassium fertilizer gross dose (squared)

				- Nitrogen fertilizer net dose
Potato yield	0.84	<0.0001	- Yield (squared)	- Pesticide rate per hectare - Irrigated area (squared) - Potassium fertilizer net dose (square root) - Average temperature - Nitrogen fertilizer gross dose (elevated to -1) - Total phosphorus country-level consumption (log)

*All predictors are statistically significant at the 5% level. Log transformation refers to natural logarithm (ln). Predictors for each model are sorted by decreasing absolute value of standardized β coefficients, which measures their degree of influence on yield. Predictors having a positive relationship with yield are written in black, while those having a negative relationship are marked red.

Here, goodness of fit plots (Measured values vs Predicted values) are shown for every yield prediction model.

Figure 2. *Goodness of fit plots for the yield prediction models*



The graphs confirm that the precision of all the yield prediction models is acceptable.

Formulas for each yield prediction model

a. Average all-crop yield

$$\ln(Yield) = 3.94 + 0.8192 * \ln(Nitrogen.dose) - 0.0091 * \text{Phosphorus.gross.dose} + 0.0001009 * K.fertilized.area^2 + 0.2131 * \sqrt{Pesticide.rate} - 1.4813 * \ln(Max.temperature)$$

b. Maize yield

$$\sqrt{Yield} = -0.07554 + 0.5094083 * \ln(Nitrogen.dose) - 0.0300201 * Temperature + 0.0003079892 * Tractor.density + 0.000001355661 * Total.pesticides$$

c. Wheat yield

$$\ln(Yield) = 0.1548 + 0.006018 * Nitrogen.gross.dose - 0.01026 * Phosphorus.gross.dose + 0.0001388 * Potassium.gross.dose^2 + 0.004938 * Nitrogen.dose$$

d. Potato yield

$$Yield^2 = -94.43 + 153.75 * Pesticide.rate - 0.9396 * Irrigated.area^2 + 74.09 * \sqrt{Potassium.dose} - 20.91 * Temperature - \frac{18992.62}{Nitrogen.gross.dose} - 109.86 * \ln(Phosphorus.gross.dose)$$

Conclusions and recommendations

- Increasing yields seem to be the main factor driving agricultural production growth worldwide. This suggests that to keep expanding the global agricultural output to feed the growing population, research and product development efforts should be focused on increasing crop yields, instead of habilitating further arable lands.

- According to the average all-crop yield prediction model obtained in this project, research and product development focus should be specifically placed in variables that greatly influence yields, such as NPK fertilization and plant nutrition (with nitrogen fertilizer dose being the most important yield-driving factor worldwide), pest management and crop tolerance to high temperatures (specially considering climate change).

- The production and area cultivated with oil palm fruit is increasing quickly. This means that research regarding oil palm production chain variables such as seeds and varieties, fertilization, pest management, irrigation, post-harvest, logistics, etc. could be beneficial for farmers and traders of said crop. Furthermore, this opens an opportunity for businesses to develop products that optimize oil palm fruit production, given that the amount of area and farmers of this crop are increasing, thus creating more demand for products like improved seeds and propagules, biostimulants, pesticides and others that specifically enhance oil palm production and yields. This could also help meet global demand using the already existing oil palm crops, in turn reducing the further expansion of oil palm fields, which is causing environmental problems in some areas like Borneo and the Amazon Forest.

- Brazil, China, India, USA and Indonesia are the countries contributing the most to the production of the most important crops worldwide. Therefore, agricultural research should include conditions similar to those of these countries in order for the results to be usefully extrapolated. Economically, these countries represent an important opportunity for the market of products related to the production of the most important crops worldwide.

- Mushrooms, hemp and anise, among others, are the agricultural items with the highest harvested area growth in the period 2016-2019. Therefore, these could be interesting investment, research and product development opportunities, since the increase in harvested area suggests the presence of an expanding market for these products.

- According to the maize yield predictive model, nitrogen fertilization is the most important yield-driving factor, followed by average temperature, tractor density and pesticide use. Therefore, efforts should be allocated to optimize nitrogen nutrition, heat-tolerance, tecnification (tractor use) and pest management to increase maize yields. As for temperature, it presents an inverse relationship with yield, meaning that higher temperatures cause lower yields. This is concerning in the context of climate change and calls for research related to generating heat-tolerant maize varieties and products that reduce heat-induced stress, like biostimulants.

- As for the wheat yield predictive model, the most important predictor was also nitrogen fertilization, and all 4 predictors were related to NPK fertilization. This indicates that to increase wheat yields focus should be placed on optimizing plant nutrition and fertilization. The inverse relationship of yield with phosphorus gross dose (which was also found for all-crop yield model) is strange considering that phosphorus is an essential element for plant development and usually increases yields. However, this relationship could be due to phosphorus being overused causing toxicity and antagonism with micronutrients like zinc, or due to phosphorus gross dose being

correlated with another variable not considered in this project which is inversely related to wheat and all-crop yield. This relationship should be further investigated.

- Regarding the potato yield predictive model, the most important predictor was pesticide rate, followed by irrigated area, potassium dose, average temperature, nitrogen gross dose and phosphorus consumption by country. Of these, only pesticide rate and potassium dose have a positive relationship with yield. This suggests that efforts should be made to optimize pest management and potassium fertilization in order to increase potato yields worldwide. As for the negative relationship with temperature (which was also found for maize and all-crop models), this further reinforces the need to generate heat-tolerant varieties and products that help potato crops tolerate heat, especially in the context of climate change. As for the negative relationship with irrigated area, nitrogen gross dose and phosphorus consumption by country, more investigation is needed to determine the factores responsibe for this.

- Finally, more investigation is needed to better undestand the economic and social implications of world agricultural production. It would be useful to include data related to prices, economic indexes, demographics and other phenomenon that could improve the decision-making process and impact of agriculture worldwide.