**Introduction**

The first ESCB/SSM hackathon on climate change took place from 14th to 16th November 2023 at the ECB in Frankfurt, Germany. The event was organized by ECB, Banca d’Italia, Banque de France, and the European Commission (Copernicus Programme).

The research topic “Beat the heat with cool insights!” focused on assessing the impact of extreme weathers (heat waves and droughts) on agricultural and industrial production. Teams had a 24-hours programming time to solve the challenge. The following day was dedicated to the presentation of the team results, and a price ceremony in the presence of President Christine Lagarde.

The Climate Change Challenge provided a good opportunity to expand technical knowledge and skills in manipulating climate data, extracting relevant insights, programming, and using advanced analytics within the ECB public cloud. As for all the previous hackathons, this was hands-on learning with enriching teamwork and networking sessions.

Around 60 participants from almost all the EU central banks with different experience, knowledge, or background in data science, statistics, computer science, climate and environmental sciences, economics, and/or macro- and micro-prudential risks gathered in Frankfurt and worked in eight mixed teams for three days. The project had to be delivered in 24 hours, so people worked day and night, some without any breaks. The atmosphere was exciting and at the same time stressful due to the limited time available and the complexity of the project required.

President Lagarde was present at the awards ceremony demonstrating a deep and profound interest in climate change.

# Deep dive in the work of the winning team

# *Data and variables*

To tackle this challenge, the team used data on economic activity and climate variables.

On the one hand, the economic activity were retrieved from National Accounts in the “General and regional statistics” of the Eurostat Database. National Accounts are well-established statistics and are harmonized up to the NUTS-3 level. NUTS-3 region are relatively small as depicted by the small regions on Figure 1. For most regions, yearly data are available since 2003. The main variables of interest are the gross value added in the agriculture and in the manufacturing sectors[[1]](#footnote-1). Gross value added is defined as output value at basic prices less intermediate consumption valued at purchasers’ prices.

[map of max temp in 2001 and 2018 for comparison illustration ? ]

On the other hand, drought and heat stress monthly data were retrieved from the European Drought Portal and Copernicus Climate Change Service. Variables included relevant drought and heat stress indicators, such as Standardized Precipitation Evapotranspiration Index (SPI), Soil moisture anomaly (SMA), low river flow index (LFI), a heatwave index (HWI), the Fraction of Absorbed Photosynthetically Active Radiation anomaly (FAPAR) and the Average maximum temperature (AMT). Most of them (i.e. SPI, LFI, FAPAR, HWI) were based on deviations from long-term averages. This particularity is essential to detect climate anomalies and quantity their severity since, as an example, while a summer mean air temperature average of 18°C in Luxembourg[[2]](#footnote-2) would be considered as normal while 18°C in winter would obviously considered as an anomaly.

|  |  |
| --- | --- |
| Figure 1. NUTS-3 regions for Belgium | |
|  |  |

# *Exploratory analysis*

From the exploratory work using descriptive statistics, we found that sectorial economic variables alone were good predictors of the sectorial gross value added. As a result, since the importance of the challenge is to understand the mechanisms under which climate data affects production, it was decided to include these economic variables as controls to isolate the effect of climate variables on production.

Then, short-term and long-term drought were characterized using respectively the precipitation anomalies based on a comparison of observed total precipitation amounts for an accumulation of respectively 1 month and 12 months. Similarly, the heat wave index)[[3]](#footnote-3), indicating the severity of a heatwave, and the number of heat waves in a specific year were used as a detector of a heat wave. Lastly, to capture the compounding effect of a drought and a heatwave, we created variables determining short and long term, mild, moderate and severe droughts associated with a heatwave.

Because regional characteristics play a crucial role, the model includes spatial controls to assess whether a region is distinguished by mountains, a coastal boundary, or elevated levels of urbanization. Then, since many climate events are shared within a neighboring region, for each month of each region, we created new “neighbor” variables (NN) by taking the average of a variable (typically the Standardized Precipitation Index, Low-Flow Index and Heat Wave index) of a neighboring region as illustrated by Figure 2.

|  |  |
| --- | --- |
| Figure 2. Example of Nearest Neighbors aggregation of the SPI at time *t* | Figure 3. Correlation Matrix of the median SPI01, SPI03, SPI06, SPI12 |
|  |  |

It is important to note that a groundwork on feature selection was required to end-up with the above list of relevant variables. As an example, the SPI was available for precipitations accumulated for 1, 3, 6 and 12 months. On Figure 3, the correlation matrix emphasize that there are high correlations between periods. As a result, to ensure to capture different effects using the SPI variables, only SPI01 and SPI12 were included in the model. In addition, SPI01 gives more information about soil moisture while SPI12 is more an indicator of ground water and reservoir storage. Among other techniques, OLS was ran to detect the most important features.

## *Model*

Finally, the Histogram-based Gradient Boosting Regression Tree algorithm was chosen to train data. The primary advantage of this technique is that the estimator has native support for missing values. When algorithms cannot handle missing values, even if only one value out of ten is missing in a row, no prediction can be estimated. Therefore, with an algorithm handling missing values, we were able to include the FAPAR variable that is available only from 2012, without discarding, nor imputing the 2001-2011 data. Other advantages concern binning, boosting and regularization[[4]](#footnote-4). In addition, to gain time in the cross-validation, we used the Halving Grid Search[[5]](#footnote-5) for the tuning of hyperparameters in the model.

To assess the model, testing data were provided for the years 2019 and 2021. The model (trained on 2001-2018 data) provided good predictions of the gross value added in agriculture and manufacturing on testing data with and out of sample R-Squared of 85% and 83%.

More interesting is the importance of the different features on the sectorial gross value added, in particular the climate data.

When focusing on agriculture, the shapely value estimated by the model are depicted on Figure 4. The economic variables largely contribute to the performance as shown in Figure 4. Looking at the top 10 largest absolute values of the average Shapely coefficients of climate variables displayed in Table 1, we observe that an increase in urbanization, the number of heat waves, an increase in the FAPAR dispersion, droughts (short-term and long-term, mild moderate and severe) associated with a heat wave and a low flow long-term average reduce agricultural production.

|  |  |
| --- | --- |
| Figure 4. Top 10 Shapely values for the Agricultural sector | Table 1. Top 10 Absolute Shapely values of climate variables for the Agricultural sector |
| Employment per capita | |  |  |  | | --- | --- | --- | | **Features** | **Shap Value** | **Abs Shap Value** | | Urbanization index | -5.11 | 5.11 | | Count of Heat Waves | -5.1 | 5.1 | | FAFAR standard deviation | -3.41 | 3.41 | | Heat Mild Drought L-T | -3.24 | 3.24 | | Heat Moderate Drought L-T | -3.1 | 3.1 | | Heat Moderate Drought S-T | -2.82 | 2.82 | | Heat Mild Drought S-T | -2.81 | 2.81 | | Low Flow L-T Average | -2.69 | 2.69 | | Heat Severe Drought S-T | -2.46 | 2.46 | | Heat Extreme Drought L-T | -2.34 | 2.34 | |

Similarly, when we concentrate on the Manufacturing sector, economic variables still have the largest effect on industrial production as shown by Figure 5.. Then, when aiming attention at climate variables (Table 2.), highly urbanized area, an increase in the median soil moisture anomalies, an increase in the long-term low flow severity, a heatwave associated Mild (Moderate) and long-term (short-term) droughts and a higher dispersion of the FAPAR tend to decrease manufacturing production. Surprisingly, the higher number of heatwaves, an increase in the SPI01 for a neighboring region, and short-term extremes droughts associated with a heatwave positively affect manufacturing production. As the relation between the manufacturing process is highly heterogeneous, and the extreme heat waves relatively rare, the latter effects should be carefully considered.

|  |  |
| --- | --- |
| Figure 5. Top 10 Shapely values for the Manufacturing sector | Table 2. Top 10 Absolute Shapely values of climate variables for the Manufacturing sector |
|  | |  |  |  | | --- | --- | --- | | **Features** | **Shap Value** | **Abs Shap Value** | | Count of Heat Waves | 144.06 | 144.06 | | Neighbours' Median SPI01 | 113.34 | 113.34 | | Urban Type | -50.28 | 50.28 | | Average Median SMA | -44.85 | 44.85 | | Low Flow L-T Average | -41.12 | 41.12 | | Heat Mild Drought L-T | -39.32 | 39.32 | | Heat Moderate Drought S-T | -38.02 | 38.02 | | Heat Extreme Drought S-T | 37.47 | 37.47 | | Low Flow L-T Average | -36.4 | 36.4 | | FAFAR standard deviation | -33.77 | 33.77 | |

# Conclusion

The team opted for a histogram gradient boosting model to assess the effect of heat waves and droughts on agricultural and manufacturing production. Their model provides good nowcasting prediction for agriculture and manufacturing, with an out-of-sample R-squared of respectively 85% and 83%. For both sectors, the results are primarily driven by the economic variables (GDP, Employment) used as macroeconomic controls and followed by the climate variables.

In agriculture, they found that an increase in urbanization, the number of heat waves, the FAPAR dispersion, droughts (short-term and long-term, mild moderate and severe) associated with a heat wave and a low flow long-term average reduce agricultural production.

For manufacturing, highly urbanized area, an increase in the median soil moisture anomalies, an increase in the long-term low flow severity, a heatwave associated Mild (Moderate) and long-term (short-term) droughts and a higher dispersion of the FAPAR tend to decrease manufacturing production. It should be kept in mind that external factors in manufacturing could not be controlled for, and that results for this particular sector should be cautiously considered as the effect of climate is more indirect than for agriculture.

1. NACE A and C according to the European statistical classification of economic activities. [↑](#footnote-ref-1)
2. Summer long-term average in Luxembourg is based on the period 1991-2020, and computed by MeteoLux. <https://www.meteolux.lu/fr/filedownload/565/2022_informations_sur_le_climat_au_luxembourg_en_2022_anglais.pdf/type/pdf#:~:text=In%202022%2C%2068%20summer%20days,days%20it%20is%207.4%20days.&text=The%20winter%202021%2F2022%20showed,K%20from%20the%20climate%20normal>. [↑](#footnote-ref-2)
3. The methodology is developed by Lavaysse et al. (2018) is based on the persistence for at least three consecutive days of events with both daily minimum and maximum temperatures (Tmin and Tmax) above the 90th percentile daily threshold. [↑](#footnote-ref-3)
4. The estimator bins input samples into integer-valued bins, reducing splitting points and utilizing integer-based data structures (**Binning**), builds an additive model in a forward stage-wise fashion, optimizing differentiable loss functions (**Boosting**), includes an L2 regularization parameter to prevent overfitting by adding a complexity penalty to the loss function (**Regularization**), supports early stopping to prevent overfitting by halting training if the model's performance on a validation set doesn't improve for a specified number of iterations (**Early stopping**). [↑](#footnote-ref-4)
5. The search strategy starts evaluating all the candidates with a small amount of resources and iteratively selects the best candidates, using more and more resources. [↑](#footnote-ref-5)