Reproducibility Award Approach Description

Required files for an entry to be eligible for the Reproducibility Award:

- Zip file containing:
    1. the completed Reproducibility Award approach description
    2. complete documented code files

Approach *(https://inseefrlab.github.io/ESA-Nowcasting-2024/methodology.html)*

Please provide a detailed description of the approach used to calculate the point estimates for the selected countries. The description should contain:

(1) the data processing steps,

(2) the methods and models used,

(3) references to the scientific papers/sources that present the methods and models used

(4) the time it took to calculate the point estimates.

Bear in mind that the approach will also be evaluated by its originality, interpretability, simplicity and quality of assumptions.

Our team employed a comprehensive methodology utilizing both standard econometric models and machine learning type models to accurately nowcast the target variables. Specifically, we utilized a combination of three econometric models, including RegARIMA, Exponential Smoothing (ETS), and Dynamic Factor Models (DFM), as well as two machine learning models, XGBoost and Long Short-Term Memory (LSTM) models, to ensure robust and accurate forecasting. While we applied the same five model classes across all three challenges, including PPI, PVI, and Tourism, we tailored the specific datasets and parameters to each challenge to optimize model performance. This approach allowed us to leverage the strengths of both traditional econometric models and cutting-edge ML techniques to achieve the best possible forecasting results.

RegARIMA

Introduction

ARIMA modelling is a common type of time models used to capture internal information of a series, whether it is stationary or non stationary. ARIMA models offer some advantages in this exercise: they are relatively simple, easy to interpret, and therefore provide a useful benchmark for more complex models. As standard ARIMA models do not include external information, we use the extended version RegARIMA with external regressors (regression model with ARIMA errors):

$$y_t = c + \sum_{i=1}^{p} \alpha_i x_t^i + u_t$$

$$\phi(L)(u_t) = \theta(L)(\varepsilon_t)$$

Additional regressors used in this project are of two types:

- Economic variables presented in the data section. REGARIMA models must keep parsimonious (contrary to other methods implemented in this project), so relevant variables are selected through standard selection procedures or a priori.
- Outlier variables such as level-shift or additive outilers to control for atypical observations.

Models are applied to the first differentiation of the variable of interest.

Automatic identification and nowcasting

We use the RJdemetra package (the TRAMO part) that provide an easy way to identify and estimate RegARIMA models with high flexibility on parameters.

- we allow for automatic detection of outliers on the estimation period in order to avoid large bias on coefficients. Without adding outliers, “covid” points, for instance, would totally distort coefficients. Outliers identified through this procedure could in addition be used in other methods
- computation time is fast (a few seconds for the whole set of countries)
- external regressors are selected independently, a priori or through standard variable selection procedure.

The final nowcast value is provided by the projection of the model at horizon 1, 2 or 3, depending on the challenge and the position in the month.

Seasonal adjustment for electricity

For electricity, the seasonal component is very strong and may not be the same as the explanatory variables. An X13 pre-treatment is applied to seasonnally adjust and correct for trading days the target variable and the potential explanatory variables. This treatment also provides a prediction for the seasonal coefficient of the nowcasted month.

Next, seasonally adjusted variables are put in the REGARIMA model. The final step involves nowcasting the “raw value” by dividing the SA forecast by the projected seasonal coefficient.

Dynamic Factor Models (DFM)

Introduction

Dynamic Factor Models are a powerful and versatile approach for nowcasting, which involves extracting latent factors from a large set of observed economic or financial indicators. These latent factors capture the underlying dynamics of the data and are used to generate forecasts or predictions in real-time.

DFM are based on the idea that a small number of unobservable factors drive the behavior of a large number of observed variables. These factors represent the common underlying movements in the data and can be interpreted as representing the state of the economy or the financial system at a given point in time. By estimating these latent factors from historical data, DFM allows us to capture the relevant information embedded in the observed indicators and use it to generate accurate nowcasts.

A standard dynamic factor model involves 2 main equations.

- **The factor equation (Equation 1):** This equation represents the dynamics of the latent factors, which are unobservable variables that capture the common underlying movements in the observed data. The factor equation is usually specified as a dynamic system allowing the unobserved factors F_t to evolve according to a VAR(p) process, and can be written as:

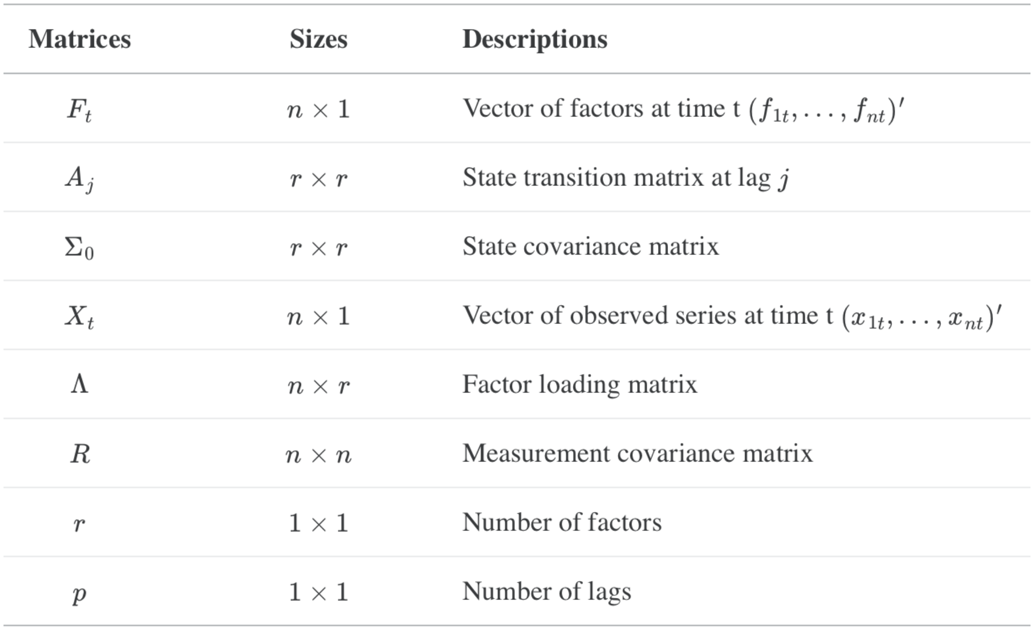
  $$F_t = \sum_{j=1}^{p} A_j F_{t-j} + \eta_t, \qquad \eta_t \sim N(0, \Sigma_0) \tag{1}$$

  where F_t represents the vector of latent factors at time t, A_j is the (state-) transition matrix capturing the dynamics of the factors at lag j, F_{t-1} is the vector of factors at time t-1, and Σ_0 is the (state-) covariance matrix.

- **The measurement equation (Equation 2):** This equation links the latent factors to the observed variables. It specifies how the observed variables are generated from the latent factors and can be written as:

  $$X_t = \Lambda F_t + \xi_t, \qquad \xi_t \sim N(0, R) \tag{2}$$

  where X_t represents the vector of observed variables at time t, Λ is the factor loading matrix, linking the factors to the observed variables, ξ_t is the vector of measurement errors at time t and R is the (measurement-) covariance matrix.



Matrices	Sizes	Descriptions
F_t	$n \times 1$	Vector of factors at time t $(f_{1t}, \ldots, f_{nt})'$
A_j	$r \times r$	State transition matrix at lag j
Σ_0	$r \times r$	State covariance matrix
X_t	$n \times 1$	Vector of observed series at time t $(x_{1t}, \ldots, x_{nt})'$
Λ	$n \times r$	Factor loading matrix
R	$n \times n$	Measurement covariance matrix
r	1×1	Number of factors
p	1×1	Number of lags

Data used and estimation

For the estimation of our Dynamics Factor Models (DFM), we utilized three main sources of data to capture different aspects of the economic and financial activity. These data sources include Eurostat data for economic activity, financial data obtained using the Yahoo API for financial activity, and Google Trends data for capturing more recent evolutions. By incorporating these three main sources of data, we aim to capture different aspects of the economic and financial activity, ranging from long-term trends to short-term fluctuations and recent evolutions. This multi-source data approach allows us to build a more comprehensive and robust DFM model, which can provide more accurate and timelier nowcasting predictions for the variables of interest.

For the estimation of our various Dynamics Factor Models (DFM), we relied on the “dfms” package in R, a powerful and user-friendly tool for estimating DFMs from time series data. The “dfms” package provides convenient and efficient methods for DFM estimation, making it an invaluable resource for our nowcasting project. Its user-friendly interface and optimized algorithms make it easy to implement and customize DFM models, while its efficient computational capabilities enable us to handle large datasets with ease. We really want to thank Sebastian Krantz, the creator of this package.

To ensure the robustness and accuracy of our DFM estimation, we took several steps beforehand the estimation pipeline of the “dfms” package. Since there are no trend or intercept terms in Equation 1 and Equation 2, we made X_t stationary by taking a first difference. Note that X_t is also standardized (scaled and centered) automatically by the “dfms” package.

We also pay attention to the availability of the data to ensure that each observed series has sufficient data for estimation. If any series have inadequate data, it is removed from the estimation process to prevent biased results. We also account for potential collinearities that could occur among several economic variables. Highly correlated series ($\rho > 0.9999$) are removed to mitigate multicollinearity issues and improve estimation accuracy.

To determine the optimal values for the number of lags and factors in our DFM models, we use the Bai and Ng (2002) criteria, which provides statistical guidelines for selecting these parameters. We set a maximum limit of 4 lags and 2 factors for computational efficiency reasons.

We initiate the estimation process from February 2005 for all challenges.

Nowcasting

Once the DFM is estimated (A_j, Σ_0, Λ, R), it can be used for nowcasting by forecasting the latent factors using the factor equation, and then generating nowcasts for the observed variables using the measurement equation. The predictions can be updated in real-time as new data becomes available, allowing for timely and accurate predictions of the variables of interest.

ETS

Exponential smoothing models are a class of models where forecasts are linear combinations of past values, with the weights decaying exponentially as the observations get older. Therefore, the more recent the observation is, the higher the associated weight is. Moreover, exponential smoothing models do not require any external data.

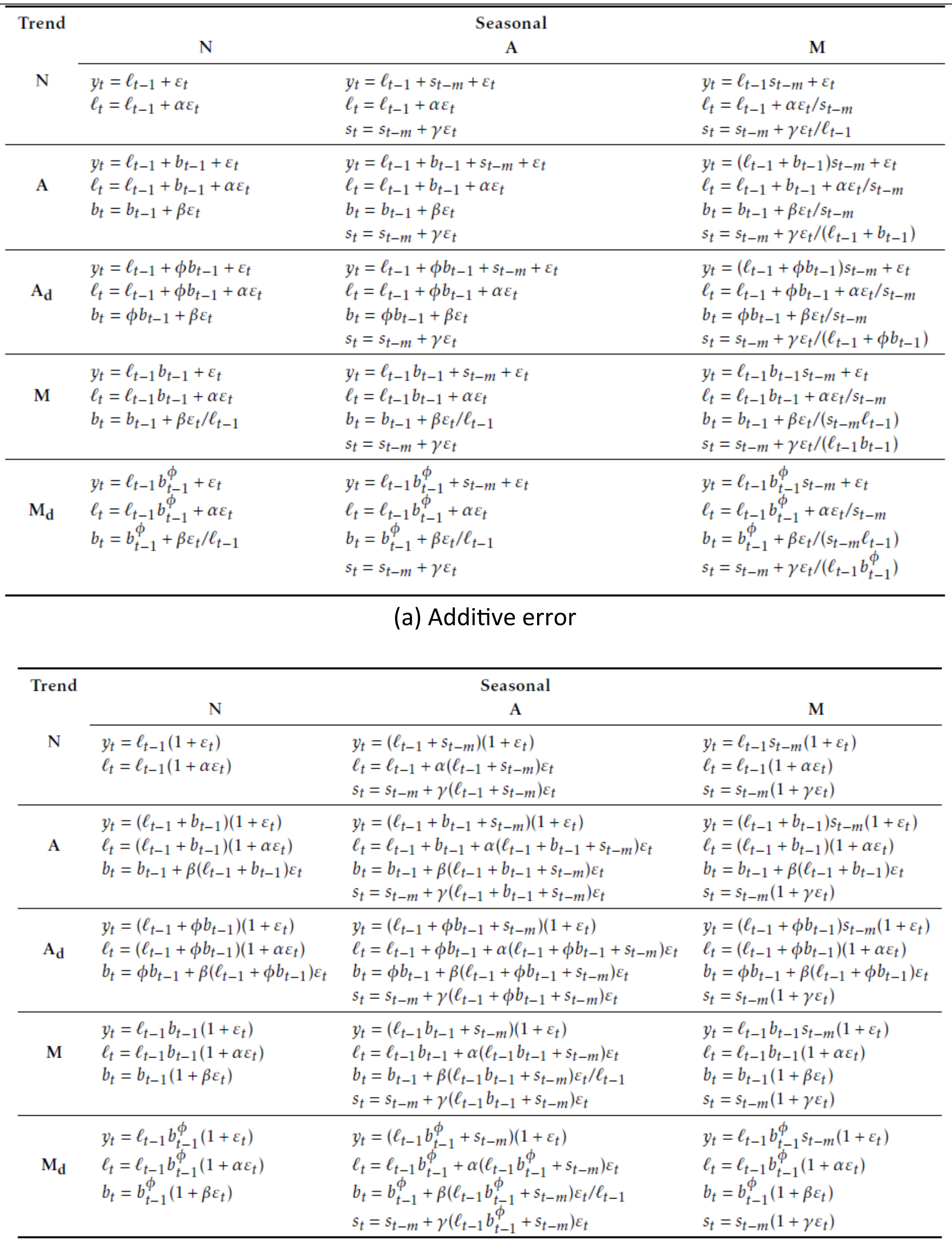
The exponential smoothing models used are a combination of three components:

- An error: additive (A) or multiplicative (M).

- A trend: additive (A), multiplicative (M), damped (A_d or M_d) or absent (N).

- A seasonality: additive (A), multiplicative (M) or absent (D).

See Figure 1 for the description of the model.



Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathrm{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$y_t = \ell_{t-1}b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
$\mathrm{M_d}$	$y_t = \ell_{t-1}b_{t-1}^{\phi} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$	$y_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1}^{\phi})$

(a) Additive error

Trend	Seasonal		
	N	A	M
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathrm{A_d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
M	$y_t = \ell_{t-1}b_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathrm{M_d}$	$y_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1 + \beta\varepsilon_t)$	$y_t = (\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1 + \beta\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

(b) Multiplicative error

Source: Hyndman, R.J., & Athanasopoulos, G. (2018).

Figure 1: Exponential smoothing models.

For each series, the model is selected minimizing the Akaike's Information Criterion (AIC) and parameters are estimated maximizing the likelihood.

XGBoost

Introduction

XGBoost is a powerful algorithm that has gained popularity in the field of machine learning due to its ability to handle complex interactions between variables and its flexibility in handling various types of data. In the context of Eurostat's nowcasting competition, we utilized the XGBoost algorithm to predict the values of the Producer Price Index, the Producer Volume Index and the Number of nights spent at tourist accommodation establishments for most European countries. We will delve here into the technicalities of the XGBoost approach, and how we tailored it to our specific nowcasting problem.

XGBoost Algorithm

XGBoost is a gradient boosting algorithm that is particularly well suited for regression and classification problems. It works by building a sequence of decision trees, each tree trying to correct the errors of the previous tree. During the training process, the algorithm iteratively adds decision trees to the model, where each new tree is fit on the residuals (i.e., the errors) of the previous trees. The final prediction is made by adding the output of all the trees.

To control for overfitting, XGBoost uses a combination of L1 and L2 regularization, also known as "lasso" and "ridge" regularization, respectively. These regularization methods add a penalty term to the loss function, which forces the algorithm to find simpler models. L1 regularization shrinks the less important features' coefficients to zero, while L2 regularization encourages the coefficients to be small, but does not set them to zero. By using both methods, XGBoost is able to produce models that are both accurate and interpretable.

Another key feature of XGBoost is its ability to handle missing values. Rather than imputing missing values with a fixed value or mean, XGBoost assigns them a default direction in the split, allowing the algorithm to learn how to handle missing values during the training process.

Overall, the XGBoost algorithm has proven to be a powerful tool in the field of machine learning, and its ability to handle large datasets and complex interactions between variables make it well-suited for nowcasting problems like the Eurostat competition.

Transforming Time Series

To apply the XGBoost algorithm to our nowcasting problem, we first transformed the time series data into a larger dataset tailored for the algorithm. We gathered several sources of data, including financial series, macroeconomic series, and surveys, and created a dataset where each row corresponds to a value per country and per date with many explanatory variables. We added lagged versions of the target variable and some of the explanatory variables as additional features. By doing so, we captured the time series properties of the data and made it suitable for the XGBoost algorithm.

Grid Search of Hyperparameters

To obtain optimal results from the XGBoost algorithm, we used a grid search technique to find the best combination of hyperparameters for each model. We experimented with various values of hyperparameters, including learning rate, maximum depth, and subsample ratio, to determine which combination of parameters resulted in the best performance. The grid search enabled us to identify the best hyperparameters for the model, allowing us to obtain the most accurate predictions. We did not differentiate the hyperparameters for each country as it would have likely caused even more overfitting.

Training XGBoost for Nowcasting

To predict our 3 indicators for each country, we trained an XGBoost model for each country independently. We randomly split the data into training and testing sets and trained the model on the training set using the optimal hyperparameters obtained from the grid search. We evaluated the model’s performance on the testing set using various metrics such as mean squared error and mean absolute error.

LSTM

Introduction

Long Short-Term Memory (LTSM) networks are a particularly interesting kind of recurrent Neural Network (NN) when it comes to time series forecasting. It allows for learning long-term dependencies in the data without losing performances in grasping short term relations. They overcome the main flaw addressed to recurrent NN models which is the unboundedness of the lookback time window that implies limitations in long-term dependencies. LSTM enable to cope with this problem thanks to the incorporation of a cell space that stores long term information that is updated at each step. This update implies incorporating but more importantly getting rid of some information which regulates for the long-term dependence. This is done using a particular structure using repeated modules, each of which is composed of four layers that convey information in a particular.

LSTM model

During the training process, the LSTM model is fed a sequence of inputs, with each input representing a timestep in the time series. The model then generates a corresponding output for each timestep, which is compared to the actual value to compute a loss function. The weights of the model are updated through backpropagation, where the gradient of the loss function is propagated backwards through the network.

LSTMs models are designed to remember long-term dependencies by maintaining a cell state and using gates to control the flow of information. These gates include:

- Input gate: Decides which values from the input to update the memory state.
- Forget gate: Determines what details are discarded from the cell state.
- Output gate: Decides what the next hidden state should be.

So, one of the challenges of using LSTMs for time series forecasting is selecting an appropriate window size, or the number of previous timesteps that the model should consider when making each prediction. A larger window size can capture longer-term trends in the data but may also introduce more noise and complicate the training process. A smaller window size, on the other hand, may be more sensitive to short-term fluctuations in the data, but may not capture longer-term trends as effectively.

In the context of the Eurostat competition, the time series data for each country was transformed into a format that was suitable for LSTM training (in particular, missing values in main and exogenous variable must be processed, as LSTM cannot work with). During training, the LSTM network learns by adjusting the weights of the connections based on the error of the predictions. This is done through a process called backpropagation through time.

Data and estimation

We gathered indicators of the macroeconomic environment from different sources. These data include hard: macro variables, financial indicators, economic surveys, prices, and soft indicators. The data is transformed into a large dataset, we included macroeconomic series and their lags. The series are all scaled.

The dataset is split into 2 parts: the first contains the data that will be used to train the model, while the second, known as the ‘future’ part, contains the months to be predicted and the values of all the available exogenous variables.

The LSTM model is trained independently for each country, with a grid search used to find the optimal hyperparameters for each country.

Once trained, the LSTM can make predictions about the current state of the time series. It uses the most recent data points to forecast the present or near future values.

Overall, the LSTM approach proved to be a powerful tool for time series forecasting, and its ability to capture long-term dependencies in the data made it particularly well-suited for nowcasting problems.

LSTM can deal with various frequencies of data inputs and is capable of managing large numbers of input features. But, designing and implementing an LSTM model can be complex and requires careful tuning of parameters. LSTMs like other neural networks, are often considered “black boxes” because it’s difficult to interpret how they make predictions.

Some references that were used

- Hopp, Daniel, Economic Nowcasting with Long Short-term Memory Artificial Neural Networks (LSTM) (March 23, 2021). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3855402
- Woloszko, N. (2020), « Tracking activity in real time with Google Trends », Documents de travail du Département des Affaires économiques de l'OCDE, n° 1634, Éditions OCDE, Paris, https://doi.org/10.1787/6b9c7518-en
- Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3
- Gómez, Víctor and Maravall, Agustin, (1998), Automatic Modeling Methods for Univariate Series, Working Papers, Banco de España, https://EconPapers.repec.org/RePEc:bde:wpaper:9808

- Chang, Ih & Tiao, George & Chen, Chung. (1988). Estimation of Time Series Parameters in the Presence of Outliers. Technometrics. 30. 193-204. 10.1080/00401706.1988.10488367

Similarities/differences to State-of-the-Art techniques (optional)

Please provide a list of similarities and differences between the approach and the state-of-the-art techniques. See https://inseefrlab.github.io/ESA-Nowcasting-2023/methodology.html.

A few characteristics and specificities of our work:

- A significant portion of our database comprises classic macroeconomic indicators, including prices, surveys, and Brent. All of them are open source.
- For low-dimensional methods, we ensure minimal control over the consistency of selected variables or signs of estimated coefficients.
- We have compared different methods and included traditional methods in our analysis, but our research also incorporates methods that rely on recent developments.
- We exclusively use open data sources, avoiding non-free data aggregators, which ensures high reproducibility. However, data retrieval can be more expensive, and some non-free data cannot be utilized.
- We made a conscious effort to identify new data sources or indicators. We also incorporate soft data such as Google Trends in our analysis.
- The methods we use combine data with diverse frequencies, up to weekly indicators (created from daily data), allowing us to improve the precision of our predictions up until the last day of each month.

Moreover, we at Insee Innovation Team place a paramount value on scientific reproducibility, particularly in the realm of data science, and as such, we have made **all our codes easily accessible to the public through Github** (https://github.com/InseeFrLab/ESA-Nowcasting-2024). **The comprehensive collection of codes has been available to the public since the commencement of the competition**, and we take great pride in our unwavering commitment to transparency and openness. A website to present our results in real time was also developed and is available here: https://inseefrlab.github.io/ESA-Nowcasting-2024/.

Lessons Learned (optional)

Please state any lessons learned during the competition.

See https://inseefrlab.github.io/ESA-Nowcasting-2024/lessons-learned.html.

- Retrieving data is a costly task, as we exclusively use open data sources and avoid non-free data aggregators. In addition to classic macroeconomic indicators that are common to most European countries, identifying interesting indicators specific to certain countries can be expensive. Unfortunately, the short duration of the competition limited our ability to acquire new data sources, such as payment card data, which could have been useful for the challenge. Moreover, for a goal of reproducibility, we decided to exclude non open source data from our scope.
- Post-mortem analysis on errors is crucial. However, in the real-time context of nowcasting challenges, having a track record of past residuals before the start of the challenge is not always straightforward. Economic variables availability can move throughout the month, making it difficult to establish a true track record. For more information and visualizations about our post-mortem analysis, you can go to the following website: https://inseefrlab.github.io/ESA-Nowcasting-2024/post-mortem-gas.html.
- Depending on the model, taking into account the impact of COVID-19 on estimation is relevant. Otherwise, coefficients could be strongly biased, with the variance of COVID-19 points dominating the total series variance.
- Our approach is mainly neutral regarding the choice of variables, with an automatic selection procedure and a focus on treating all countries. This mainly neutral approach is partially due to a lack of time, but fine-tuning country by country can also be a useful approach.
- "Soft" data, such as Google Trends, appears to provide some information for the tourism challenge, but less so for production prices and production, at least during a "stationary" period.
- Using nowcasting techniques on disaggregated variables is an interesting option, particularly for prices that have exhibited distinct dynamics across different products in recent times. However, implementing this approach can be expensive as it necessitates the use of different models for each disaggregated level and appropriate re-aggregation for obtaining the final nowcast value. Given our constraints with respect to time, we were unable to explore this approach thoroughly.
- For most of our models, the last available value of the indicator often has a very big influence, more than we would have thought. Because of this, even in our most recent results we may observe a lag between the true value of the indicators and our predictions based on past data. This shows that we were not able to identify all the external factors influencing the indicators. With more resources and a larger time window, we would still be able to identify some more explicative variables to improve the predictions.

List of Data Sources with Descriptions

For each country, list the data sources (and their description) that were used to calculate the point estimates for the selected country. Please use the template below to provide the information for each source. **If multiple data sources were used, please copy paste the template below and fill it in.**

Bear in mind that the data sources will also be evaluated based on its openness, availability, coverage and consistency.

Introduction

We take great pride in our dedication to using exclusively open data sources in our modeling efforts, which was a fundamental aspect of our approach during the challenge. While some proprietary data sources may have had greater predictive power, we firmly believed that utilizing open data sources was crucial to promoting the principles of transparency and reproducibility in our modeling efforts. By leveraging publicly available data, we were able to derive nowcasts of key economic indicators while ensuring that our work can be easily replicated and validated by others in the official statistics community. This approach not only provided us with a robust foundation for our models but also served to promote the values of open science, data transparency, and reproducibility.

During the challenge, we utilized three primary sources of data to inform our modeling efforts. The first source was **economic data** from the Eurostat database, which provided us with a comprehensive overview of the economic situation in the European Union. The second source of data was **financial data**, which provided us with valuable insights into the financial context surrounding each of the challenges. This data included stock prices, exchange rates, and other financial indicators that were useful in predicting economic trends and identifying potential risks. Finally, we also used **Google Trends** data to capture the most recent trends and shifts in consumer behavior. This data enabled us to monitor changes in search volume for specific keywords, which served as an early warning system for sudden changes in consumer sentiment and preferences. Overall, our use of these three distinct sources of data allowed us to develop a comprehensive understanding of the economic landscape and to generate nowcasts of the 3 target variables.

All the information about the data we used is available here: https://inseefrlab.github.io/ESA-Nowcasting-2024/data.html. Unless specified otherwise, all the data has been retrieved for all the countries in the scope of the challenge (or is not country-dependant).

Eurostat data

Supply, transformation and consumption of gas	The NRG_CB_GASM belongs to the monthly European statistics that cover the most important energy commodities. For inland gas consumption, the data are collected by the reporting countries via separate dedicated questionnaires and subsequently aggregated and transferred to Eurostat.
Supply, transformation and consumption of oil and petroleum	NRG_CB_OILM belongs to the monthly European statistics that cover the most important energy commodities. For oil and petroleum product deliveries, the data are collected by the reporting countries via separate dedicated questionnaires and subsequently aggregated and transferred to Eurostat.
Supply, transformation and consumption of electricity	NRG-CB_EM belongs to the monthly European statistics that cover the most important energy commodities. For electricity availability, the data are collected by the reporting countries via separate dedicated questionnaires and subsequently aggregated and transferred to Eurostat.
Producer prices in industry	Producer Price in Industry (PPI) refers to the average price that domestic producers receive for the goods and services they produce. This indicator measures changes in the price of goods and services at the producer level, and it is considered an important leading indicator of inflation.
Industrial import price index	Import prices in industry, also known as industrial import price index (IPI), refer to the cost of goods and services imported into a country for use in production. This indicator reflects changes in the prices of imported raw materials, intermediate goods, and capital equipment, and it is influenced by factors such as exchange rates, global commodity prices, and trade policies.
Production index in industry	The Production Index in Industry (or Production Volume in Industry - PVI) is a measure of the physical output of the industrial sector of an economy. It tracks changes in the volume of goods produced over time, and it is considered an important indicator of the health and performance of the manufacturing sector. The production index can be used to assess trends in productivity, capacity utilization, and competitiveness.
Harmonised Index of Consumer Prices on a few products	The Harmonised Index of Consumer Prices (HICP) is a measure of inflation that is used to compare price changes across the European Union. It tracks the average change over time in the prices of goods and services that households consume, including food, housing, transportation, and healthcare. The HICP is calculated using a harmonised methodology that ensures comparability across EU member states, and it is published on a monthly basis by Eurostat. It is a key indicator of price stability.
Business survey in industry	The Business Survey in Industry is a survey conducted by Eurostat to gather information on the business conditions and expectations of companies in the manufacturing sector. The survey covers a range of topics, including production, new orders, inventories, employment, prices, and investment, and it is conducted on a monthly basis across the European Union. The data collected from the survey can be used to assess the current and future state of the manufacturing sector, to identify sector-specific challenges and opportunities, and to inform policymaking and business decision-making.
Consumer survey in industry	The Consumer Survey in Industry is a survey conducted by Eurostat to gather information on the consumer sentiment and behavior in the European Union. The survey covers a range of topics, including household income, savings, spending intentions, and major purchases, and it is conducted on a monthly basis. The data collected from the survey can be used to assess consumer confidence, to identify trends in consumer spending and saving patterns, and to inform policymaking and business decision-making. The Consumer Survey in Industry is an important indicator of the overall health of the economy, as consumer spending is a major driver of economic activity.

Yahoo Finance data

Yahoo Finance is a popular online platform for financial information and investment tools. It provides a wide range of financial data, including real-time stock prices, historical price charts, news articles, analyst ratings, and financial statements for publicly traded companies. We used its API to get the latest financial data to improve our short-term predictions.

Euro/Dollar exchange rate	The Euro/Dollar exchange rate represents the value of one euro in terms of US dollars. It is a widely followed currency pair in the foreign exchange market, as it reflects the relative strength of two of the world's largest economies. Movements in the exchange rate can be influenced by a range of factors, such as interest rate differentials, inflation expectations, political developments, and global economic trends. The exchange rate can impact international trade, investment flows, and the competitiveness of exports and imports, making it a key indicator for businesses, investors, and policymakers alike.
Brent Crude Oil Stock Price	The Brent Crude Oil Last Day Financial Futures Stock Price is a benchmark for the price of crude oil from the North Sea, which is used as a pricing reference for approximately two-thirds of the world's traded crude oil. As a financial futures contract, it allows investors to trade the price of oil without actually buying or selling the physical commodity. The stock price reflects the market's perception of supply and demand dynamics, geopolitical risks, and other macroeconomic factors that impact the oil market.
Crude Oil Futures Price	Crude oil futures price refers to the agreed-upon price for the delivery of a specified quantity of crude oil at a future date. It's determined through trading on futures exchanges, where buyers and sellers speculate on future oil prices based on various factors such as supply and demand dynamics, geopolitical events, and market sentiment.
Natural Gas Futures Price	Natural gas futures price is the agreed-upon price for the delivery of a specified quantity of natural gas at a future date. Like crude oil futures, it's determined through trading on futures exchanges, where participants speculate on future natural gas prices based on factors such as supply and demand dynamics, weather patterns, storage levels, and geopolitical developments.
S&P 500 Index Stock Price	The S&P 500 Index stock is a market capitalization-weighted index of 500 leading publicly traded companies in the United States. It is widely considered to be a barometer of the US stock market's performance, providing investors with a broad-based measure of the economy's health and direction. The S&P 500 index includes companies from a range of sectors, such as technology, healthcare, finance, and energy, making it a diversified indicator of the US equity market.
Euro stoxx 50 Index Stock Price	The Euro Stoxx 50 Index stock is a market capitalization-weighted index of 50 leading blue-chip companies from 12 Eurozone countries. It is designed to reflect the performance of the Eurozone's most liquid and largest companies across a range of industries, including banking, energy, consumer goods, and healthcare. As a widely recognized benchmark of the Eurozone equity market, the Euro Stoxx 50 Index stock is used by investors and analysts to track market trends, benchmark portfolio performance, and identify investment opportunities. Movements in the index are influenced by a range of factors, such as economic growth prospects, monetary policy decisions, geopolitical risks, and corporate earnings announcements.
CAC40 Index Stock Price	The CAC 40 Index Stock is a benchmark index of the top 40 companies listed on the Euronext Paris Stock Exchange, representing a broad range of industries such as energy, finance, healthcare, and technology. It is the most widely used indicator of the French equity market's performance and is considered one of the leading indices in Europe. The CAC 40 Index Stock is weighted by market capitalization and is closely monitored by investors and analysts as an indicator of economic health and growth prospects in France. Movements in the index can be influenced by a variety of factors, such as geopolitical risks, macroeconomic indicators, and company-specific news.

Google Trends data (https://trends.google.fr/)

Google Trends is a free online tool provided by Google that allows users to explore the popularity of search queries over time and across different regions and languages. It provides valuable insights into the behavior of internet users, the topics they are interested in, and the evolution of search trends over time.

Nevertheless, the use of Google Trends data as a tool for economic analysis needs to be done carefully. Google Trends provides Search Volume Indices (SVI) based on search ratios, with the initial search volume for a category or topic at a given time divided by the total number of searches at that date. However, changes in the denominator (total searches) can induce biases as internet use has evolved since 2004.

We implemented an approach to address this downward bias by extracting a common component from concurrent time series using Principal Component Analysis (PCA) on the log-SVI series long-term trends filtered out using an HP filter. The rescaled first component obtained from the long-term log-SVIs is assumed to capture the common long-term trend, and it is subtracted from the log-SVIs. This approach can help to remove the downward bias common to all Google Trends variables and improve their economic predictive power. More information in this paper.

The categories we used for this challenge are:

Fuel Economy & Gas Prices
Electricity
Oil & gas

Other data

Electricity prices	Ember.org provides European wholesale electricity price data that can be used to analyze electricity market trends, monitor price volatility, and inform investment decisions. The data is sourced from various market operators and exchanges across Europe and covers a wide range of countries and regions.	All countries

Hardware Specifications

Please describe the hardware specifications of the machines that were used to calculate the point estimates.

See https://inseefrlab.github.io/ESA-Nowcasting-2023/reproducibility.html for how to run the code.

Machine 1: Github Actions

CPUs	2 – Core CPU (X86_64 – 7Go RAM / 14Go Disk)
GPUs	0
TPUs	0
Disk space	500Mo

Short description of the team and all team members – area of expertise (optional)

Please provide a description of the team, all team members, their area of expertise and contact information.

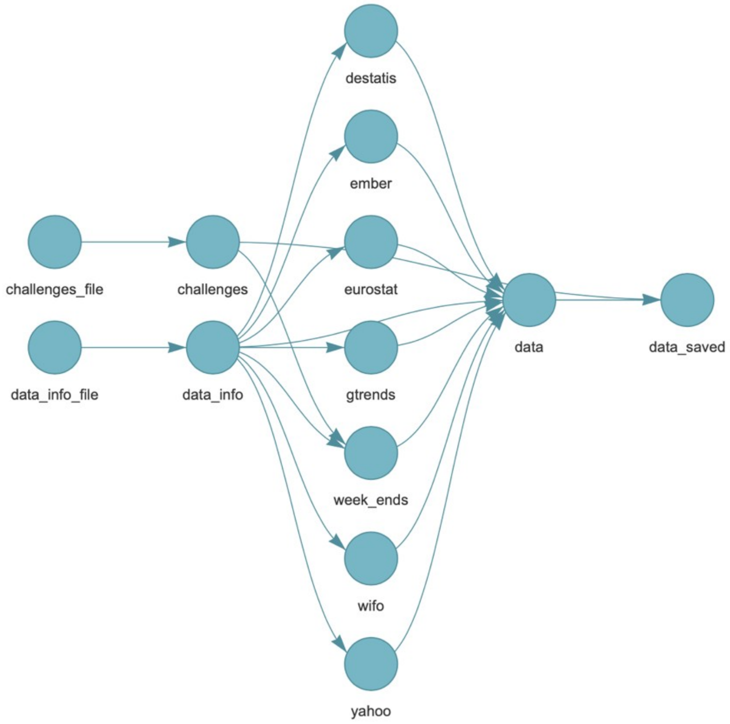
We are all members of **INSEE**, the French National Institute of Statistics and Economic Studies, for this challenge. Our team is composed of 3 members:

- Yves-Laurent Benichou, Senior Data Scientist in the Innovation Lab in Data Science
- **Thomas Faria**, Data Scientist in the Innovation Lab in Data Science
- **Antoine Palazzolo**, Data Scientist in the Innovation Lab in Data Science

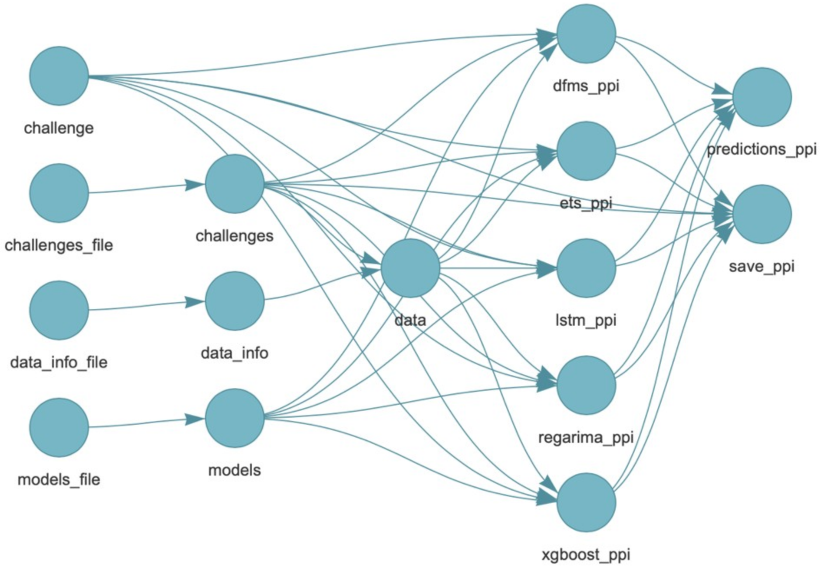
For more information about our team and our work, you can visit the following website: https://inseefrlab.github.io/ESA-Nowcasting-2024/.

Appendix

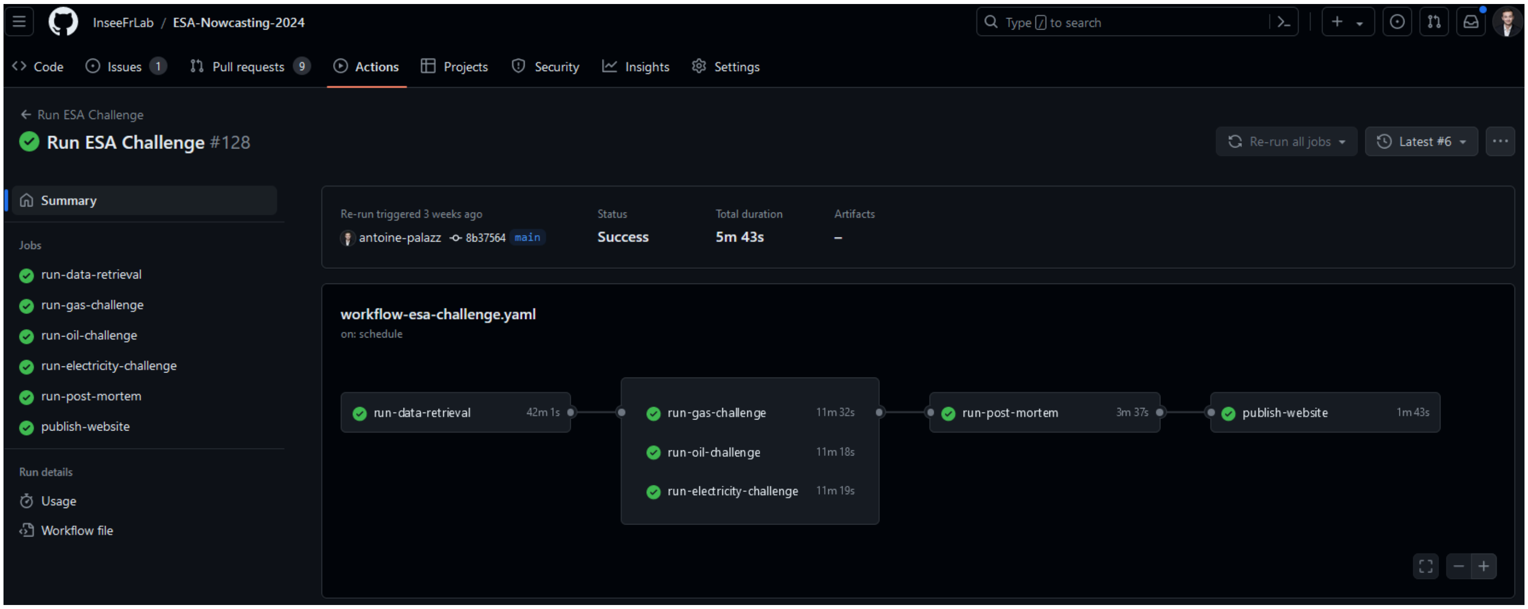
Example of a possible pipeline visualization with Targets: *the data pipeline*



Example of a possible pipeline visualization with Targets: *the GAS pipeline*



The whole pipeline visualized with Github Actions



← Run ESA Challenge

Run ESA Challenge #128

Summary

Jobs

- run-data-retrieval
- run-gas-challenge
- run-oil-challenge
- run-electricity-challenge
- run-post-mortem
- publish-website

Run details

- Usage
- Workflow file

Re-run triggered 3 weeks ago — antoine-palazz 8b37564 main

Status: Success

Total duration: 5m 43s

Artifacts: –

workflow-esa-challenge.yaml

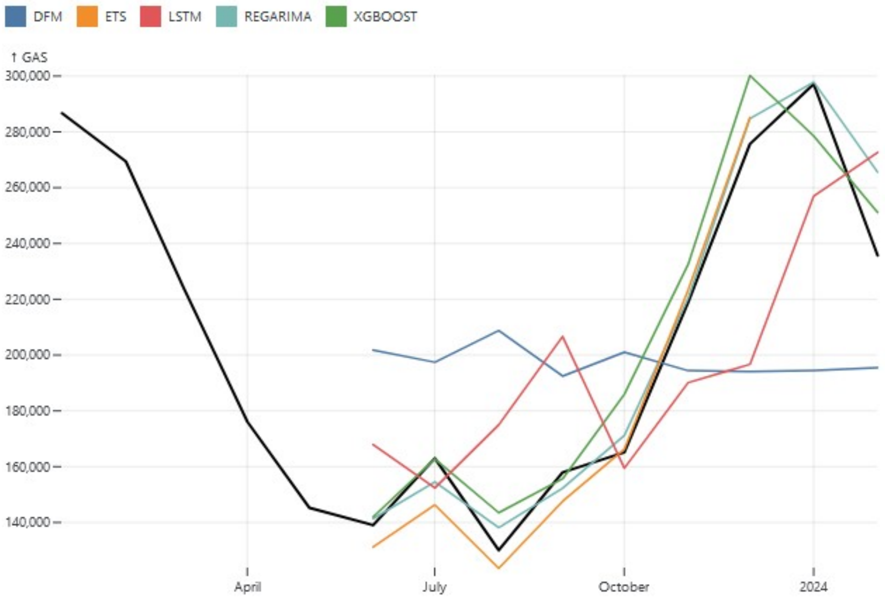
on: schedule

run-data-retrieval 42m 1s → run-gas-challenge 11m 32s, run-oil-challenge 11m 18s, run-electricity-challenge 11m 19s → run-post-mortem 3m 37s → publish-website 1m 43s

Examples of post-mortem visualizations (available here):

Past predictions

This interactive graph displays the historical forecasts generated by all of our models, as well as the actual observed value for the selected country.

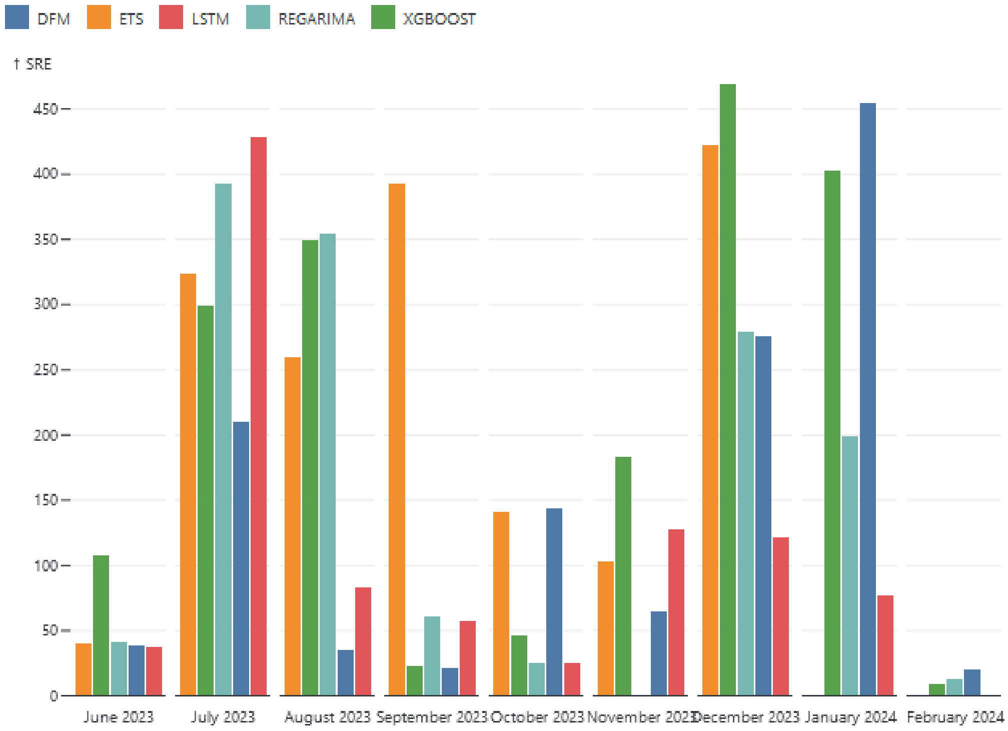


Square relative error per month

This graph illustrates the square relative error for each of the models used in the challenge. The square relative error is a measure of the accuracy of a forecast that takes into account the magnitude of the error, as well as the level of the first official release being predicted.

$$SRE = \left(\frac{Y - R}{R}\right)^2$$

where R is the first official release and Y the nowcasted value.


DFM
ETS
LSTM
REGARIMA
XGBOOST
↑ SRE
June 2023
July 2023
August 2023
September 2023
October 2023
November 2023
December 2023
January 2024
February 2024