Supplementary Information for the paper:

A Machine Learning architecture to forecast Irregular Border Crossings and Asylum requests for policy support in Europe: a case study

# Materials and Methods

## 

## Dataset List

The list of datasets used in our work is the following:

1. Detections of irregular border-crossings statistics (updated monthly), downloaded from: <https://www.frontex.europa.eu/what-we-do/monitoring-and-risk-analysis/migratory-map/>
2. Asylum applicants by type of applicant, citizenship, age and sex - monthly data downloaded from: <https://ec.europa.eu/eurostat/en/web/products-datasets/-/MIGR_ASYAPPCTZM>
3. Headline consumer price index monthly (hcpi\_m), is the raw measure of all aspects within economy that experience migration including highly volatile ones (food, energy) and those changing seasonally. It is partly related to the overall economic condition but often related to shifting costs of living. We used the data provided for all the countries by World Bank, downloaded from: <https://www.worldbank.org/en/research/brief/inflation-database>
4. Food price index, monthly (fcpi\_m), is the measure of food inflation rates over time compared to a reference period of an average basket of commonly consumed foods. We used the database provided by World Bank, downloaded from: [https://www.worldbank.org/ en/research/brief/inflation-database](https://www.worldbank.org/%20en/research/brief/inflation-database)
5. Foreign Direct Investment (FDI), net inflows, are the value of a purchase of interest or a project by a company or other entity from a different country. It reflects the interest that foreign business has in the country and is encouraged as a way to improve infrastructure and create new jobs. As such, it also establishes supranational ties between the countries, and these can translate into international migration. We used the World Bank dataset, downloaded from <https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD>
6. Consumer prices indices, monthly, general one, refers to the inflation in relation to consumer prices. We used database provided by FAOSTAT, downloaded from <https://fenix.fao.org/faostat/internal/en/#data/CP>
7. Monthly food price inflation estimates for fragile countries, is a special dataset computed by the World Bank for the countries for which the traditional data may not be available. In the list of the so-called 25 “fragile countries” there are: Haiti, Chad, Nigeria, Central African Republic, Republic of the Congo, Cameroon, Niger, Gambia, The, Sudan, Mali, Somalia, South Sudan, Mozambique, Guinea-Bissau, Liberia, Burkina Faso, Burundi, Myanmar, Lao People's Democratic Republic, Afghanistan, Syrian Arab Republic, Iraq, Lebanon. We downloaded the data from <https://datacatalog.worldbank.org/search/dataset/0060165/Monthly-food-price-inflation-estimates-by-country>. Fragile countries are identified based on the World Bank Country Policy and Institutional Assessment (CPIA). More on the typology: <https://thedocs.worldbank.org/en/doc/9b8fbdb62f7183cef819729cc9073671-0090082022/original/FCSList-FY06toFY22.pdf>

## Distance Correlation

To avoid the issue of multicollinearity and redundant features, we build a matrix of distance correlation coefficients among each pair of features (see Figure S.1) and remove those with high coefficients.

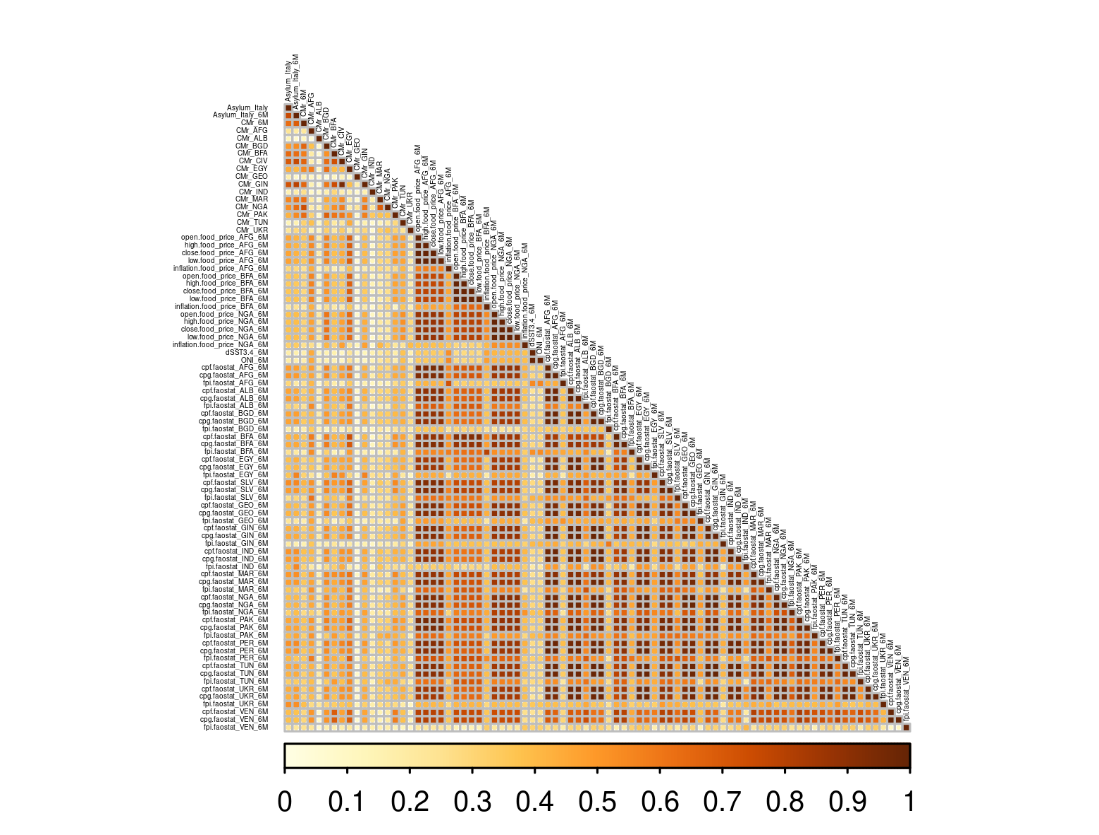


Figure S.1 – Distance correlation matrix (lower triangular).

## Variance Inflation Factor

The Variance Inflation Factor (VIF) measures the amount of multicollinearity between independent variables in a multivariate regression model. VIF can be defined as

eq.(1)

where is isthe R-squared value for the regression of the *i*-th regressor on the other regressors.

The following figures show the VIF values obtained for the selected variables for the cases of IBCs and asylum applications. Note that in all cases the values are lower than 3, meaning that there is little to no multicollinearity in the chosen predictors.

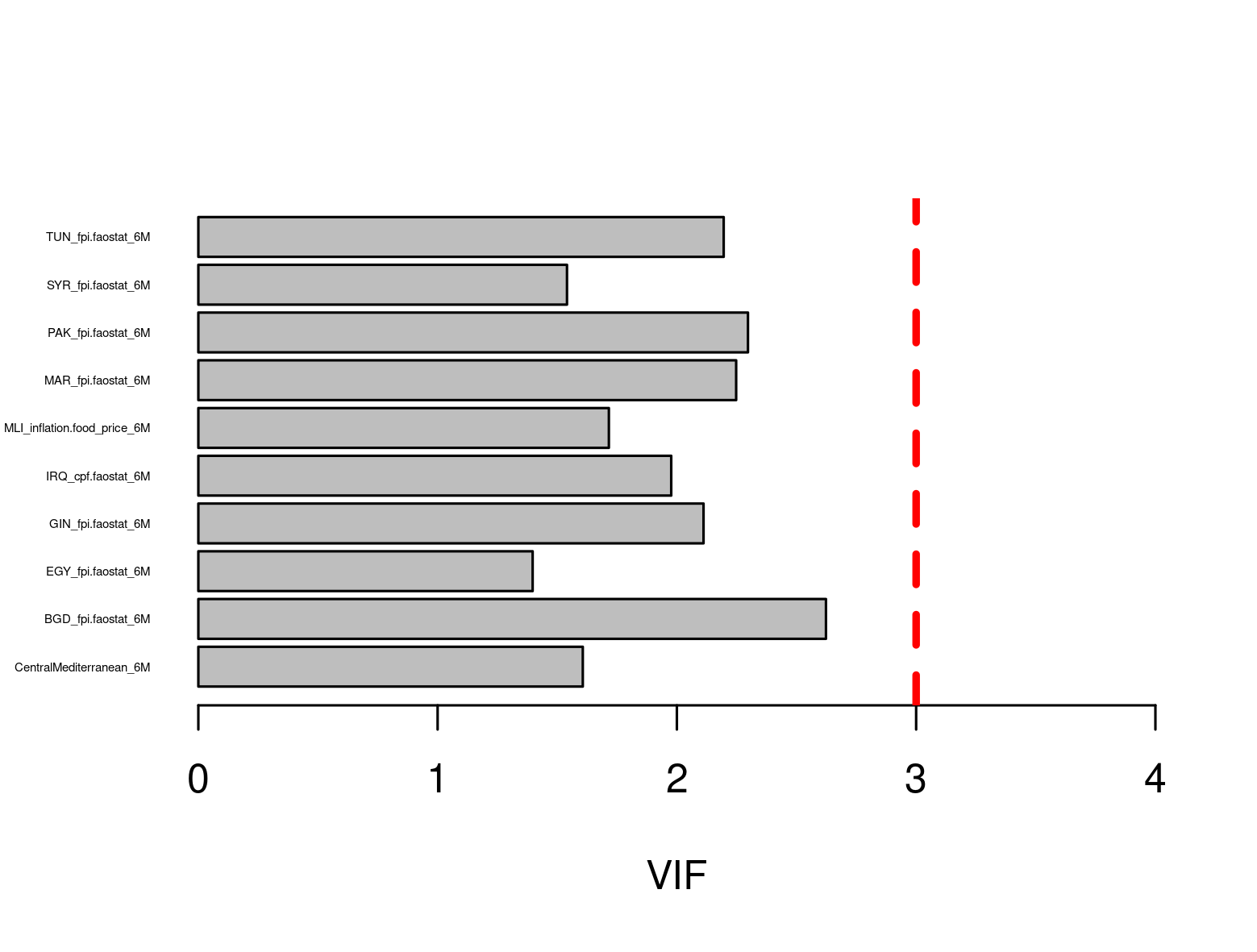


Figure S.2 – Variance Inflation Factor plot for the selected predictors of Irregular Border Crossings. All values are below 3 (red dashed vertical line).

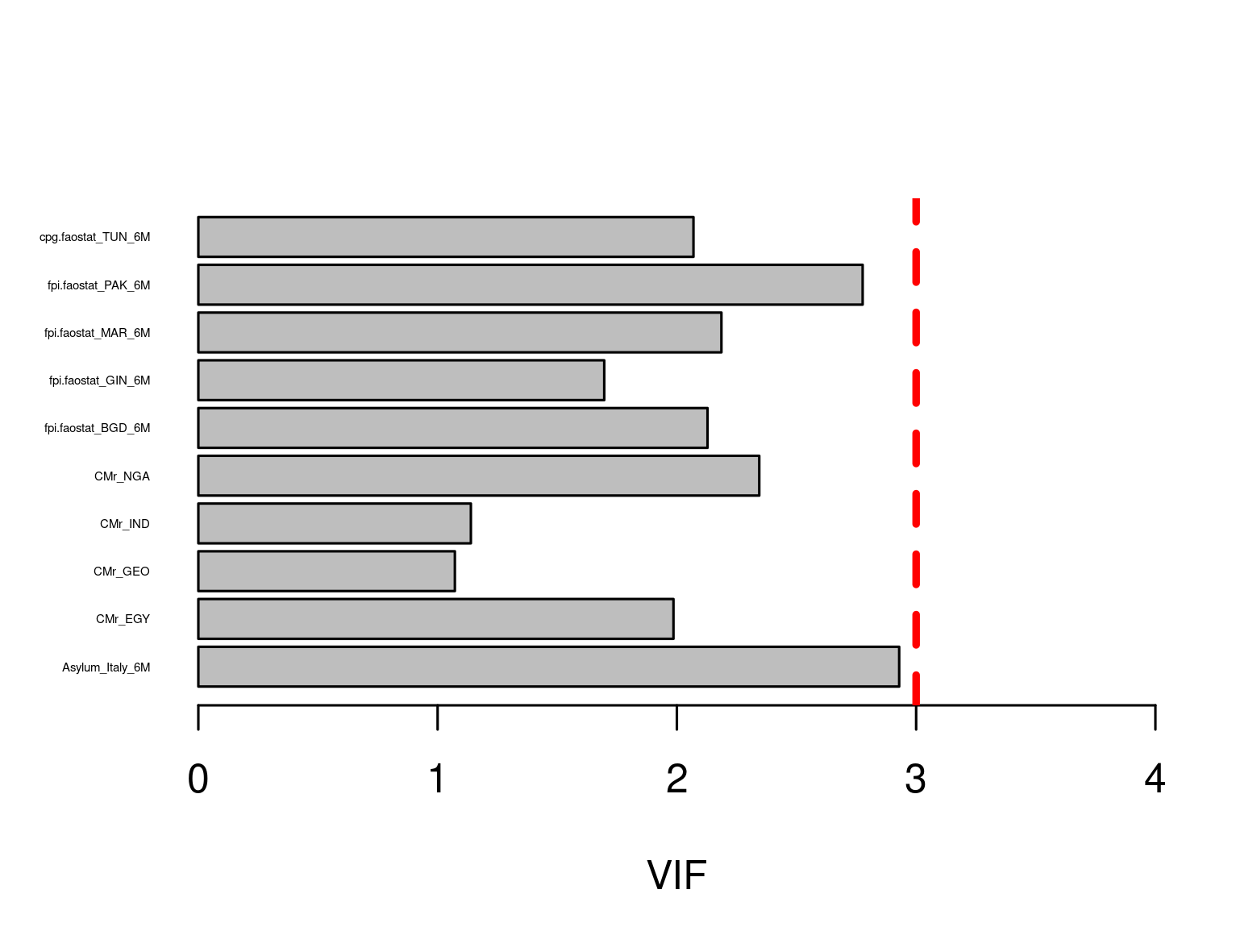


Figure S.3 – Variance Inflation Factor plot for the selected predictors of asylum applications. All values are below 3 (red dashed vertical line).

## Data Partitioning

To include in our model the great variability of the whole time series due to the variability characterizing migration and the various shocks to the migration system since 2009, we split the data so that one random observation by quarter is used for validation (around 25% of the data), and the rest is used as training set (around 75%, see Figures S.4-5). We have used the same split for all the models we used (trivial, RF, XGBoost, and ANN).

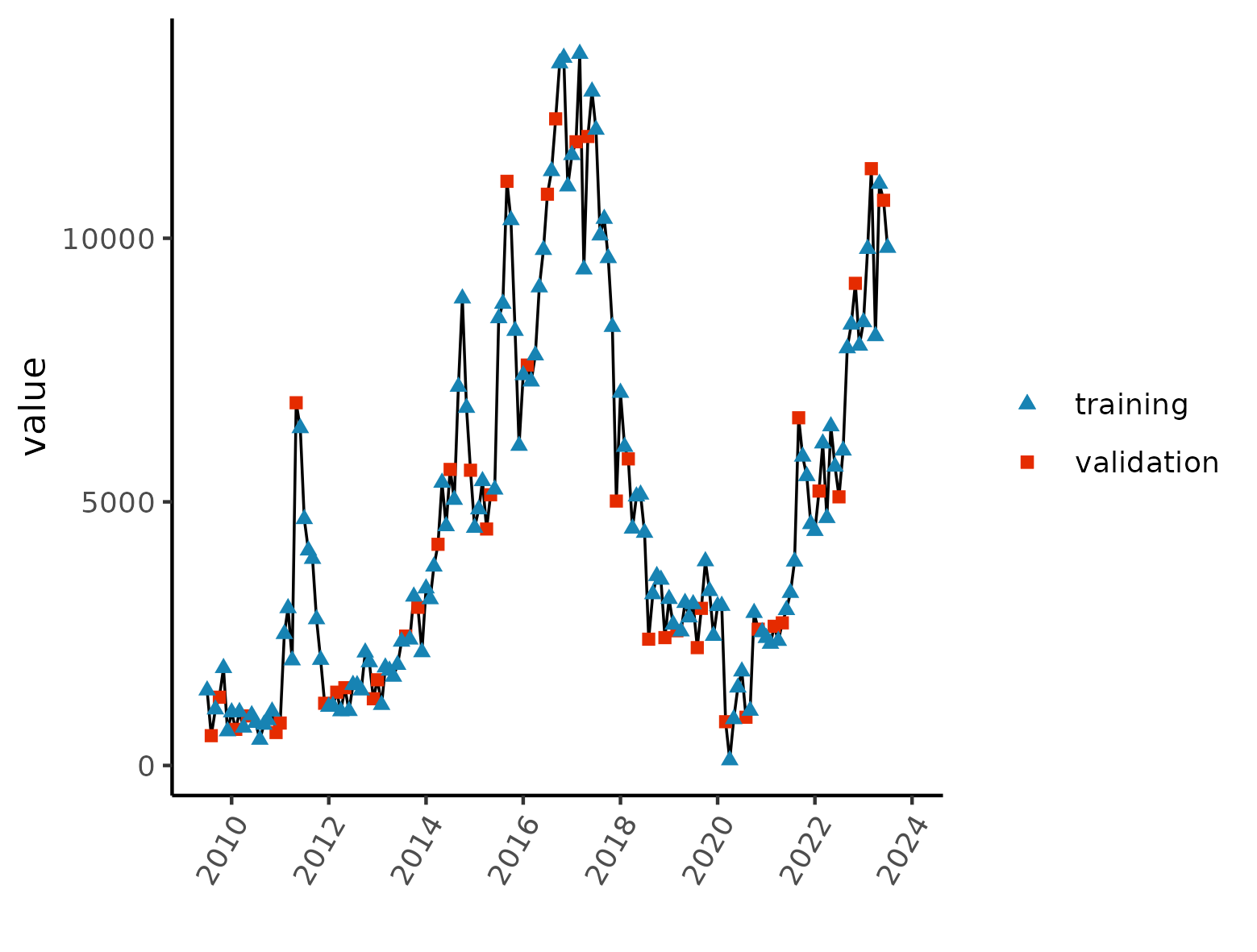


Figure S.4 – Data partitioning into training and validation set for asylum data.

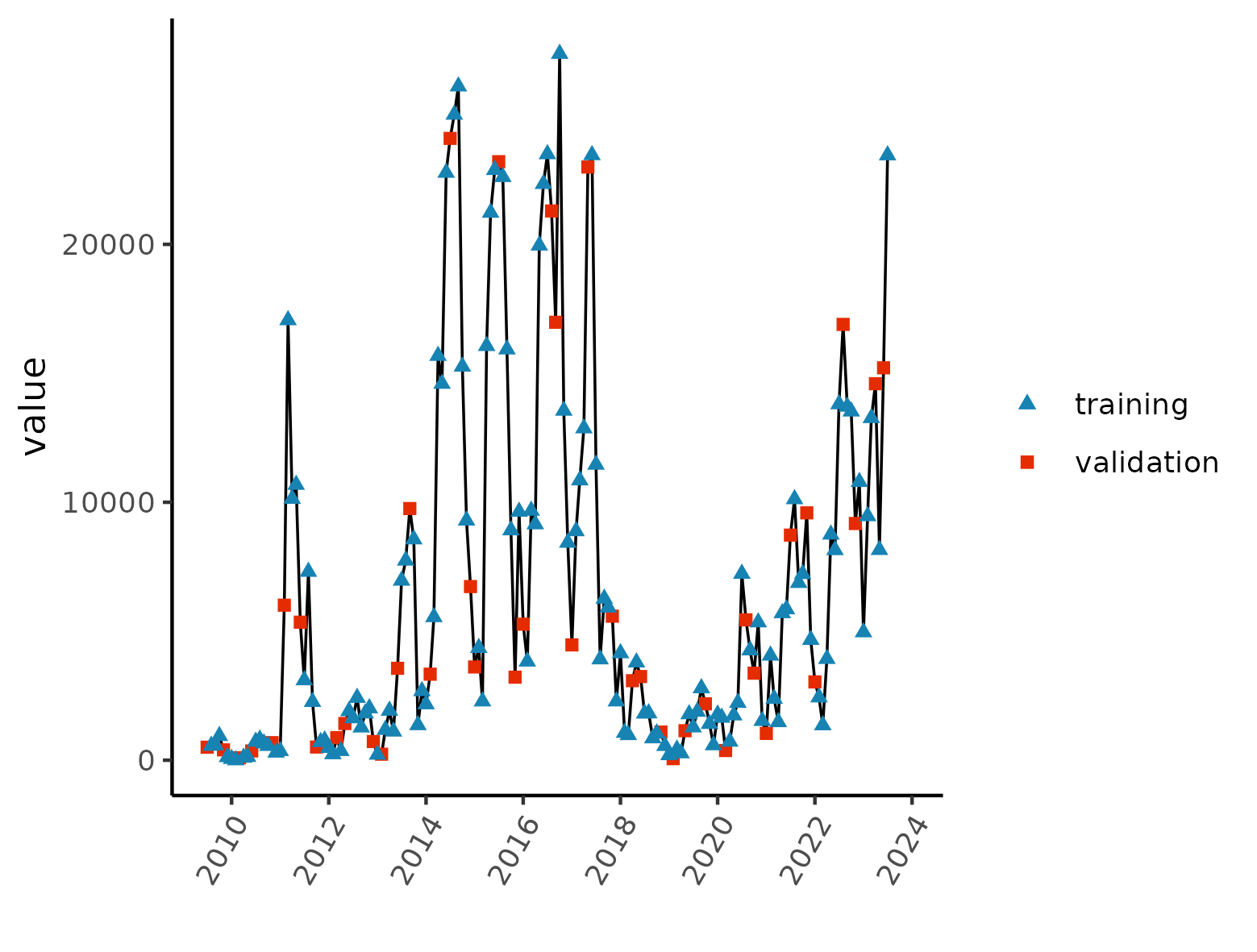
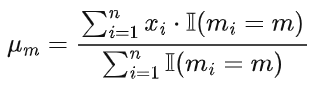


Figure S.5 - Data partitioning into training and validation set for IBCs data.

## Trivial Model

We used a trivial model as the baseline to benchmark the performance of our machine learning models and the final ensembling. The trivial model uses the same training and validation data of the other more complex models. It calculates the mean by month of the training set and assigns this mean to the future months. Equation 2 shows the trivial model specification.

 eq.(2)

where:

* is the mean of the time series *X* (training set) for month *m*
* is the value of the time series *X* (training set) at time point *i*
* *I(mi=m)* is the indicator function, which is 1 if *mi=m* and 0 otherwise
* *n* is the total number of data points in the time series *X* (training set)

# Results

This section includes the results in validation for all the models we applied both for predicting asylum applications and irregular border crossings.

## 

## Asylum applications

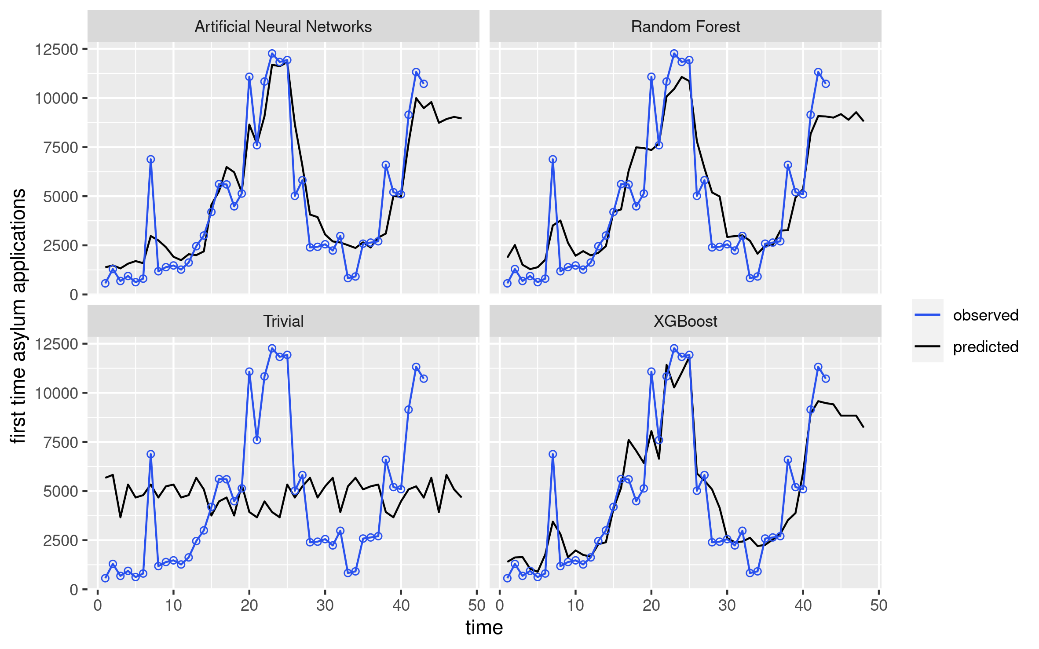


Figure S.6 - Results of the three ML models and the trivial model to predict asylum application in Italy (validation set of data). In blue the observed values and in black the modelling predictions.

## Irregular Border Crossings

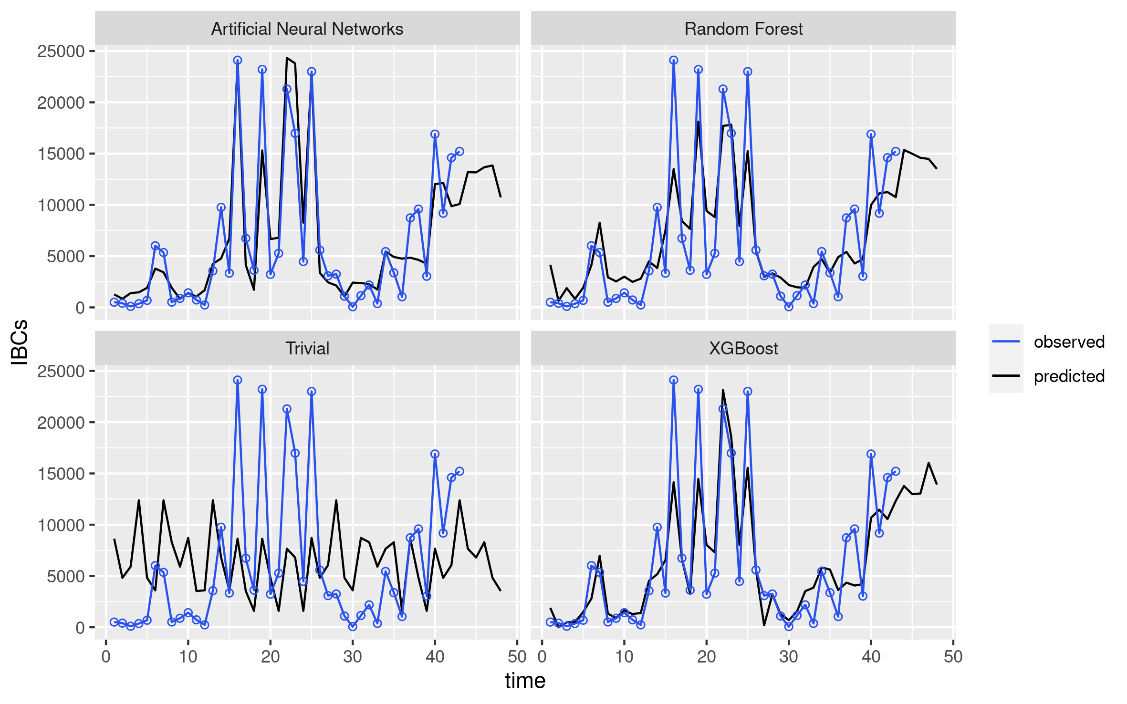


Figure S.7 - Results of the three ML models and the trivial model to predict IBCs over the Central Mediterranean Route (validation set of data). In blue the observed values and in black the modelling predictions.