Documentation of the methodology used for the system imbalance forecast publications

October 2022

Content

1. Introduction

In a 2021 study[1], Elia developed and tested different machine-learning algorithms to forecast the average quarter-hourly system imbalance in the ongoing quarter hour and the next quarter hours. Forecasting the system imbalance is first of all relevant because the future system imbalance is key factor – along with others – for determining the volume of balancing reserves to activate. In addition, the study also analyzed the relevance of publishing the system imbalance forecasts and concluded that the publication of the system imbalance forecasts could bring a number of advantages, such as enabling a higher/more targeted implicit reaction from market parties and increasing transparency on the drivers of the system imbalance. Following the conclusions of the 2021 study, Elia proposed to target the publication of the system imbalance forecasts. To this end, Elia has first taken steps to further improve and develop the machine-learning algorithms to forecast the average quarter-hourly system imbalance, and has subsequently performed the necessary implementations. As of October 2022, Elia started the publication of its system imbalance forecasts on the EliaOpenData platform.

Specifically, the published forecasts consist of:

- a forecast of the average quarter-hourly system imbalance (in MW) (point forecast); and
- An estimation of the probability that the average quarter-hourly system imbalance will lie in certain pre-defined intervals.[2]

Every minute, these forecasts are published for both the ongoing quarter hour and the next quarter hour.[3]

For making these forecasts. Elia relies on different machine learning algorithm: one algorithm is used for the point forecasts (expected average quarter-hourly system imbalance expressed in MW), and a different algorithm is used for estimating the probability that the average quarter-hourly system imbalance will lie in each of the considered intervals (referred to as “estimation of the probability distribution” in the remainder of this document). These machine-learning models are first trained on historical data. The trained model is subsequently deployed to forecast the system imbalance near real-time.

This document describes the algorithms and data used by Elia for making the published system imbalance forecasts with the aim of increasing transparency of the published system imbalance forecasts. It must be noted that the methodology used by Elia for forecasting the system imbalance can change over time. When such changes occur, Elia aims

[1] The 2021 study can be found on the **Elia website**.
[2] The considered intervals are:]-∞ ; 400],]-400 ; -200],]-200 ; 0],]0 ; 200],]200 ; 400],]400 ; ∞[
[3] For instance, at 13h06, forecasts of the average quarter-hourly system imbalance in the quarter hour 13h00-13h15 and the quarter hour 13h15-13h30 are published. At 13h07, an update of the forecasts will be published for the same quarter hours.

to update the documentation in this document. However, under no circumstances the publication of this documentation (incl. possible missing elements or errors in the documentation) implies a shift in responsibility or liability towards Elia.

The remainder of this document is structured as follows:

- Section 2 describes the machine-learning algorithms used;
- Section 3 elaborates on the training of the algorithms on historical data;
- Section 4 describes the data that is used as input for the machine-learning algorithms.

2. Machine-learning algorithms

As mentioned in the introduction, separate algorithms are used for making on the one hand the point forecasts (expected system imbalance in MW), and on the other hand, the estimation of the probability distribution.

2.1 Point forecasts

For the point forecasts, multivariate linear regression models are used. In such type of models, the forecast of the average quarter-hourly system imbalance in a given quarter is hour is determined as a linear combination of different input variables (also referred to as predictors). For the system imbalance forecast, we distinguish three type of input variables[4]:

- variables v_i that represent quarter-hourly values of past quarter hours (e.g., the average quarter-hourly system imbalance or net regulation volume in past quarter hours)
- variables v_j that represent quarter-hourly values for the ongoing quarter hour and/or next quarter hours (e.g., the scheduled quarter-hourly net exports from the Belgian zone)
- variables v_k that represent minute-resolution values of past minutes (e.g., the instantaneous system imbalance in the past minutes)

An overview of the input variables actually used is provided in Section 4. However, the resulting model can be generically represented via the equation below:

$$SI^{Forecast}_{Qh+m,min} = a + \sum_{i} \sum_{n=1}^{lookback\ horizon} w_{i,Qh-n} \times v_{i,Qh-n} + \sum_{j} \sum_{n=0}^{lookforward\ horizon} w_{j,Qh+n} \times v_{j,Qh+n} + \sum_{k} \sum_{n=1}^{lookback\ horizon\ minutes} w_{k,min-n} \times v_{k,min-n}$$

, with:

- $SI^{Forecast}_{Qh+m,min}$: the forecast, made in minute “min”, of the average quarter-hourly system imbalance in quarter hour Qh+m (and Qh represents the ongoing quarter hour);
- a : the intercept;
- $v_{i,Qh-n}$: the value of the quarter-hourly input variable i in quarter hour Qh-n;
- $v_{j,Qh+n}$: the value of the quarter-hourly input variable j in quarter hour Qh+n;
- $v_{k,min-n}$: the value of the minute-resolution input variable k in minute min-n;

[4] Note that certain type of input data can be available for both the past quarter hours and the ongoing/next quarter hours.

- $w_{i,Qh-n}$: the weight/coefficient for the input variable i in quarter hour Qh-n;
- $w_{j,Qh+n}$: the weight/coefficient for the input variable j in quarter hour Qh+n;
- $w_{k,min-n}$: the weight/coefficient for the input variable k in minute min-n.

Note that a separate linear regression model is developed and trained for every minute of the quarter hour in which the forecast is made.[5] This is done for two reasons. First, the input data that is available and can be used as a predictor can be different depending on when the forecast is made (e.g., during the first minutes of the current quarter-hour, certain data related to the past quarter hour might not yet be available). Therefore, the amount of input variables/predictors might differ for the models used for different minutes. Second, as additional close-to-real-time data becomes available as time progresses (e.g., the observed system imbalance in the minutes before the forecast is made), the weights/coefficients that would lead to the most accurate forecasts could change depending on when the forecast is made. For instance, if the forecast is made later during the quarter hour, it is possible that the impact of certain input variables on the forecasted system imbalance is reduced as this impact would possibly already be reflected in the system imbalance measured close to real time.

2.2 Estimation of the probability distribution

For estimating the probability distribution of the average quarter-hourly system imbalance in a given quarter hour, Elia considers the following intervals: (-∞, -400 MW], (-400 MW,-200 MW], (-200 MW,0 MW], (0 MW,200 MW], (200 MW,400 MW], (400 MW, ∞).

For each of these intervals, a binomial logistic regression model is developed that estimates the probability that the average quarter-hourly system imbalance lies within that interval. Specifically, the expected probability that the system imbalance in a given quarter hour “Qh” lies in a certain interval “int”, is calculated by the binomial logistic regression model as follows:

$$P\left(SI_{Qh}^{actual} \in int\right) = \frac{1}{1+e^{ß^T x_{Qh}}},$$

where x_{Qh} represents the vector of predictors used when making a forecast for the quarter hour $Qh,$ and ß represents the vector of weights assigned to these predictors. The weights are determined based on the model training (cfr. Section 3).

[5] In addition, a different linear model is used for making the forecast for the ongoing quarter hour (Qh0) and for making the forecast for the upcoming quarter hour (Qh+1). Therefore, 30 different linear regression models are trained and deployed.

The estimated probabilities of these different binomial logistic regression models are then scaled such that the sum of the probabilities assigned to the different intervals equals 1.

3. Model training

This section describes how the machine-learning algorithms are trained on historical data.

3.1 Target / Objective

For the point forecasts, the coefficients of the linear regression model (i.e., the variables a, $w_{i,Qh-n}$, $w_{j,Qh+n}$ and $w_{k,min-n}$ in the equation in Section 2.1 are determined in such a way that the forecast errors are minimized. This objective function corresponds to minimizing the mean square error of the system imbalance forecast errors:

$$Mean\ square\ error\left(a, w_{i,Qh-n}, W_{j,Qh+n}\right) = \frac{\sum_{Qh} e_{Qh}^{2}}{\sum_{Qh} 1}$$

, where the forecast error for a given quarter hour Qh (e_{Qh}) is calculated as follows:

$$e_{Qh} = SI_{Qh}^{Forecast} - SI_{Qh}^{actual}$$

, with SI_{Qh}^{actual} the measured cumulative SI in the last minute of the quarter hour.[6]

For the model used for estimating the probability distribution of the average quarter-hourly system imbalance in a given quarter hour, the different binomial logistic regression models are trained by selecting the weights ß in order to minimize the following cost function:

$$J(\text{ß}) = \sum_{Qh}\left[-y_{Qh}\log\left(\frac{1}{1+e^{\text{ß}^{T}x_{Qh}}}\right) - \left(1-y_{Qh}\right)\log(1-\frac{1}{1+e^{\text{ß}^{T}x_{Qh}}})\right],$$

$$\text{with } y_{Qh} = \begin{cases} 1\ if\ SI_{Qh}^{actual} \in int \\ 0\ otherwise \end{cases}$$

3.2 Training period and frequency

The models deployed are re-trained on a monthly basis. The model used to forecast the system imbalance in a given month M is trained on a historical data set comprising data from months M-2, M-3, M-4, M-5, M-6, M-7 and M-12[7]. This process is schematically presented in Figure 1 below.

[6] As published on the EliaOpenData platform: Activated balancing energy volumes per minute (Near real-time) — Elia Open Data Portal

[7] For example, the model used for making forecasts in November 2022 will be trained on data from [November 2021; April 2022-September 2022].

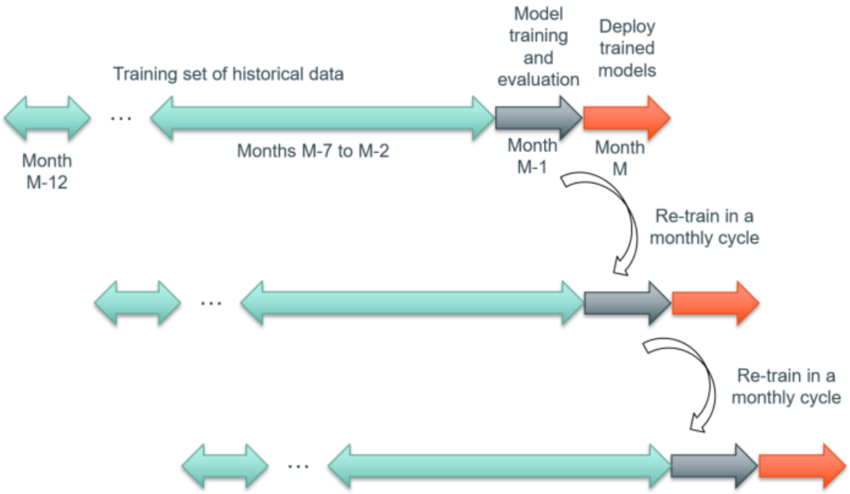
Model training and evaluation
Deploy trained models
Training set of historical data
...
Months M-7 to M-2
Month M-12
Month M-1
Month M
Re-train in a monthly cycle
...
Re-train in a monthly cycle
...

Figure 1: Schematic overview of the model training

3.3 Handling of specific cases

During the model training, certain periods are excluded from the training set. Specifically, the following periods are excluded from the training set:

- Moments with a forced outage of a unit > 100 MW in the Belgian zone: the quarter hour during which the forced outage started as well as the subsequent quarter hour;[8]
- Periods for which one of the predictors is missing.

[8] An overview of the forced outages is published on the **Elia website** and the ENTSO-E Transparency Platform.

4. Input data (predictors)

4.1 Overview of input data

An overview of the input variables used in the machine-learning models is provided in Table 1 below. Identical input data is used for the model making the point forecasts and the model estimating the system imbalance probability distribution.

Table 1: Overview of input variables used

	Input variable	Lookback horizon	Look-forward horizon	Lookback horizon minutes	Source to where the data can be found[9]
SI	Cumulative SI in the last minute of the quarter hour	4 quarter hours	/	/	Activated balancing energy volumes per minute (Near real-time) — Elia Open Data Portal
	Cumulative SI	/	/	last available minutes	Activated balancing energy volumes per minute (Near real-time) — Elia Open Data Portal and Imbalance prices per minute (Near real-time) — Elia Open Data Portal
NRV	Average quarter-hourly net regulation volume	4 quarter hours	/	/	Activated balancing energy volumes per minute (Near real-time) — Elia Open Data Portal and Imbalance prices per minute (Near real-time) — Elia Open Data Portal

[9] Note that the machine-learning models used by Elia to forecast the average quarter-hourly system imbalance do not retrieve the input data from the EliaOpenData portal, but directly from the source applications. As such, it cannot be fully excluded that the data used by Elia would differ from the data published on the EliaOpenData portal. In addition, for certain data sources, there might be a slight delay before the data is available on the EliaOpenData portal.

	Cumulative net regulation volume	/	/	4 minutes	Imbalance prices per minute (Near real-time) — Elia Open Data Portal
Imbalance tariff	Quarter-hourly imbalance tariff	4 quarter hours	/	/	Imbalance prices per quarter-hour (Near real-time) — Elia Open Data Portal
XB Nominations (External Commercial Trade Schedules)	Sum of long-term and day-ahead scheduled quarter-hourly import/export into the Belgian zone	4 quarter hours	4 quarter hours[10]	/	Day-ahead commercial schedule - by border — Elia Open Data Portal
	Intraday scheduled quarter-hourly import/export into the Belgian zone	4 quarter hours	4 quarter hours[10]	/	Intraday implicit net position (Belgium's balance) — Elia Open Data Portal
	Sum of long-term and day-ahead scheduled quarter-hourly import/export between BE and GB	4 quarter hours	4 quarter hours[10]	/	Day-ahead commercial schedule - by border — Elia Open Data Portal
	Intraday scheduled quarter-hourly import/export between BE-GB in the intraday timeframe	4 quarter hours	4 quarter hours[10]	/	Intraday implicit net position (Belgium's balance) — Elia Open Data Portal
Load	Last available intraday total load forecast	4 quarter hours	4 quarter hours[10]	/	Load and load forecasts (elia.be)
	Aggregated day-ahead Offtake Nominations from all BRPs	4 quarter hours	4 quarter hours[10]	/	/[11]

4.2 Data treatment and handling of missing data

The machine-learning algorithms used rely on significant amounts of input data. In that regard, it is possible that not all input data is available in a given moment in time. This could be the case for the model training or when deploying the trained model close to real-time. When training the models, periods with missing data are currently excluded from

[10] Including the ongoing quarter hour: Qh0, Qh1, Qh2 and Qh+3

[11] This data is currently not publically available

the training set (as discussed in Section 3.3). When deploying the trained model near real-time, gaps in the data are filled based on a linear interpolation.

No feature reduction techniques are applied to reduce the number of input data/predictors.[12]

[12] Principal component analysis (PCA) and recursive feature extraction (RFE) techniques have been tested but did lead to a (limited) degradation of the performance and hence have not been applied.