TALLINN UNIVERSITY OF TECHNOLOGY

Department of Economics and Finance

[redacted]

ESTONIAN ENERGY IMBALANCE MARKET RESEARCH AND FORECAST

Multivariate statistics (group) project

Juhendaja: Heili Hein

Tallinn 2022

RESEARCH PROBLEM

Chosen problem is a real-life business case from one of the biggest energy companies in Estonia – Eesti Energia. The ongoing energy crisis has dramatically changed energy business: electricity SPOT prices started rising in 2021 and have peaked at 4000 EUR per MWh on 17th August 2022 which affected millions of people. (Republic of Estonia Competition Authority, 2022). Dramatic increase of electricity prices is the cause of COVID19 pandemic and a sharp rise of a natural gas market price. Overall, more expensive coal consumption has raised demand for greenhouse gas emission permits, driving up prices (Eesti Energia, 2022, page 9).

The other "hidden" side of energy business is energy imbalance market, which directly affects energy traders (like Eesti Energia) as the costs for the imbalance volumes are high and indirectly affects customers with higher margins in final price.

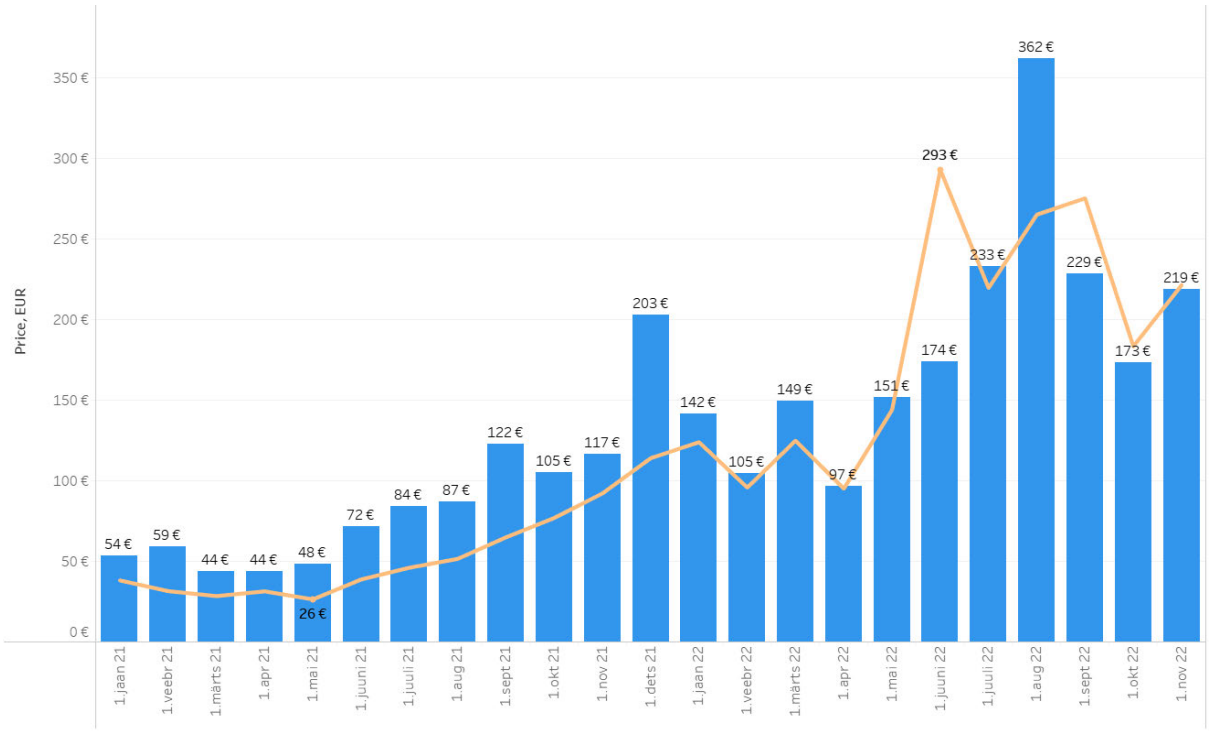
Energy market explanation in simple words.

How are SPOT prices generated? Basically, energy companies order every day electricity from the market Nord Pool, and this is how the demand is formulated. Then energy suppliers start to offer electricity in the market for reasonable price, and this is how the supply is formulated. When demand meets supply the market prices are locked with marginal electricity supplier offered price. Prices can be even negative if there is a large supplier who desperately needs to sell a lot of energy. This can be a case for nuclear power plants which cannot be stopped for a few hours if there are no demand. Unfortunately, when there's a deficit electricity prices start rising out of control (Nord Pool, 2022). Currently there is a price ceiling 5000 EUR per MWh, which was 4000 EUR per MWh before 17th August (Republic of Estonia Competition Authority, 2022).

How are imbalance prices generated? When energy company forecasts electricity consumption for the next day which later is ordered from the market, they make mistakes. Forecasts cannot be perfect and that is where the penalty for the mistake comes from. If market imbalance is in either huge deficit or surplus, then usually imbalance prices are high. There can be a situation

when company ordered less than they consumed, and they need to purchase additional electricity for imbalance price. If the imbalance price is higher than market SPOT price, then it is a cost for the company. The opposite situation can occur if the imbalance price is lower than market SPOT price, then it is a gain for the company. But overall, the costs from imbalance market are high and are counted in millions of euros.

As it is shown on the graph 1 it is clearly seen that average SPOT prices are rising with average SPOT and imbalance price difference (which is the cost). In June 2022 for the first time SPOT and imbalance spread exceeded SPOT price and is remaining high.



Graph 1. Mean SPOT electricity price (blue bars) and mean SPOT-imbalance price spread (yellow line)
Source: Created by the authors in Tableau.

In this research we want to focus on Estonian imbalance volumes. Eesti Energia share of the Estonian market is rather large (around 50%), thus around half of the imbalance is generated with Eesti Energia forecast error. If we get a better picture of imbalance market, we can optimize day-ahead forecast by lowering it if the SPOT prices are higher than imbalance prices and increasing forecast it if the SPOT prices are lower than imbalance prices. The aim is to minimize costs, not to manipulate the market which is forbidden by law.

Main research questions are:

- What market factors or external factors are influencing Estonian imbalance surplus and deficit? (Inferential problem)
- Can Estonian imbalance surplus and deficit be predicted (assuming that it is not completely random)? (Forecasting problem)

To answer given questions statistical methods like logistic regression and random forest models were applied. More precisely methods are described in further sections.

DATA

Data for this project was taken from different variety of open datasets such as Elering LIVE, Baltic Transparency Dashboard and OpenWeather database. Authors have used 16 801 observations and 18 variables in period of 01.01.2021 3AM till 31.11.2022 11PM.

As energy prices are changing every hour so there is a data about every single hour of chosen period available. Furthermore, there is concrete definition about what day of the week it is and if it's a holiday or not.

Every time label has an information about:

- SPOT price of 1 MWh of electricity in Estonia (EUR per MWh), which Eesti Energia buys from Nord Pool daily;
- air temperature (°C). There is a strong connection between air temperature and energy consumption. So, in wintertime energy consumption is bigger because of the heating season; in summertime there can be a bigger consumption of energy due to extremely high temperatures, so people use air conditioners more than usual;
- wind speed (m/s). This variable has been chosen because of wind farms which are producing green energy and owned by Eesti Energia. Due to its lower cost price, electricity produced from wind is more competitive than electricity produced from fossil fuels. In 2021 Eesti Energia produced 5 217 GWh of electricity and 983 GWh from this amount was generated by wind farms;
- level of cloudiness (percentage rate from 0% to 100%). Every year the number of investments in solar parks are growing, Eesti Energia owns them and uses it to produce green energy, which lowers CO2 emission consumption and decrease the amount of energy, which Eesti Energia must buy from Nord Pool;
- Estonian forecasted electricity consumption and production in MWh;
- imbalance price of 1 MWh of electricity in Estonia (EUR per MWh);
- Estonian imbalance volumes in MWh.

Weather data is given for five different cities in Estonia – Jõhvi, Pärnu, Tallinn, Tartu, Võru, which are located in different regions.

Table 1 represents descriptive statistics. All values except for wind speed (in different cities) have quite an extensive range of standard deviation, minimum and maximum. Considering Estonian climate (we have all 4 seasons) it is expected that data is so spread out.

Table1. Descriptive statistics of chosen variables

	Mean	Median	Min	Max	Std.dv
price_EE	134.12	100.01	-1.41	4000.00	115.27
air_temperature_EE-Johvi	6.663	6.740	-29.220	32.860	9.6411
air_temperature_EE-Parnu	7.73	7.47	-19.33	32.84	8.9413
air_temperature_EE-Tallinn	7.589	7.280	-20.710	32.190	8.9757
air_temperature_EE-Tartu	7.082	6.840	-23.380	32.960	9.6600
air_temperature_EE-Voru	6.887	6.590	-25.430	32.940	9.6878
wind_speed_EE-Johvi	3.900	3.750	0.00	13.00	1.7989
wind_speed_EE-Parnu	4.462	4.210	0.00	13.930	2.2649
wind_speed_EE-Tallinn	4.231	4.120	0.00	16.460	2.1456
wind_speed_EE-Tartu	3.272	3.090	0.00	12.860	1.8945
wind_speed_EE-Voru	3.508	3.210	0.00	10.710	1.6971
cloudiness_EE-Johvi	66.9	84.00	0.00	100.0	36.328
cloudiness_EE-Parnu	66.87	80.00	0.00	100.0	35.740
cloudiness_EE-Tallinn	44.24	40.00	0.00	100.0	39.791
cloudiness_EE-Tartu	58.96	75.00	0.00	100.0	37.778
cloudiness_EE-Voru	68.54	90.00	0.00	100.0	36.470
Forecasted_consumption	996.1	975.8	361.5	1730.2	199.89
Forecasted_production	782.1	751.5	125.5	1666.3	254.23
Imbalance volumes EE	6.22	7.96	-315.76	201.42	44.87
Imbalance prices EE	153.59	80.40	-498.75	1785.32	208.28
Imbalance volumes Baltics	11.79	10.67	-415.66	350.70	75.69

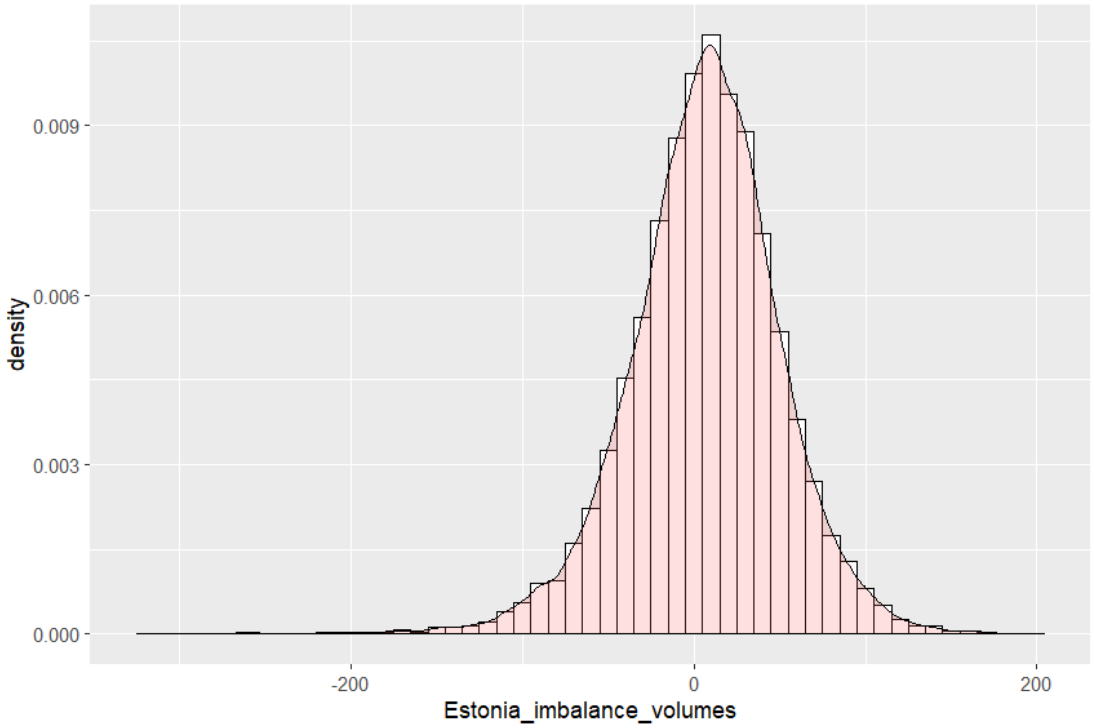
Source: author's calculations

Main variable, which authors try to explain and predict, is Estonian imbalance volumes. It was divided into two clusters:

0 – deficit where Estonian imbalance volume is less than 0;

1 – surplus where Estonian imbalance volume is more than 0.

Estonian imbalance volumes vary from -316 to 201 MWh, averaging 6.22 MWh. There are 7011 cases of deficit imbalance volumes and 9790 cases of surplus imbalance volumes. Estonian imbalance volume distribution is shown below (see Graph 2):



Graph 2. Estonian imbalance volumes density
Source: Created by the authors in R.

METHODOLOGY

Research question (1) is analyzed by using logistic regression. It is a predictive model, but it can also be used to understand how different predictor variables influence the outcome variable. Logistic regression is supervised learning classification algorithm. We chose logistic regression, because our dataset has labeled information and it is divided into two clusters (0 – deficit, 1 - surplus). We fitted logistic regression model to understand if there are statistically significant variables in our data and what is the influence on Estonian imbalance. Analysis is conducted in R using base package function "glm", family "binomial" (Appendix 1). Logistic regression gives us coefficients in log odds, so to get the odds in percentages all the exponential of coefficients were calculated. For example hour[20] had coefficient of -6.449e-01, to get the odds we used exp(-6.449e-01) function in R, then as the coefficient is negative we subtract the outcome from 1 to get the answer 48%.

Research question (2) is analyzed with logistic regression and random forest models. Authors compare both models results to find a better solution for the given problem. Random Forest is a model that consists of several independent decision trees working together as an ensemble. Every tree in the random forest spits out a class forecast, and the classification that receives the majority of votes becomes the prediction made by our model. Forecasting is conducted in R using "nnet" package function "multinom" for logistic regression and "randomForest" package eponymous function for random forest model. Function "multinom" gives same results as other function "glm", family "binomial" with one significant difference. It allows type "class" as response in prediction which is very convenient. For random forest model arguments "mtry" was chosen to be 12 which means that model is run with 12 variables and "ntree" was chosen to be 50 which means that number of merged trees is 50.

Forecasting models are validated using common method of backtesting. "Backtesting is a term used in modeling to refer to testing a predictive model on historical data. Backtesting involves moving backward in time, step-by-step, in as many stages as is necessary. [In our research, backtesting with refit and fixed training size (rolling origin) is used for forecasting model validation.] It is a technique, where the forecast origin rolls forward and the size of training remains constant (Appendix 3). This is also known as time series cross-validation or walk-forward validation. Training set period length is chosen to be 365 days. Forecasting models are tested 365 times back in time, which is equal to one year" (Skforecast, 2022).

RESULTS AND CONCLUSIONS

To answer research question (1) logistic regression analysis was conducted. For logistic regression we are only focusing on the variables that showed statistical significance, we chose confidence level as 95%. R squared was 0,082 which means that our model explains 8,2% of variance (Appendix 4).

Logistic regression results displayed that air temperature is statistically significant only when it is measured in Jõhvi, it has positive effect on imbalance surplus. +1 °C raises the odds of having an imbalance surplus by 3%. Next is windspeed, measurements taken in Tartu and Võru have very low p-value, but they have opposite effect on imbalance. When windspeed rises on average in Võru by 1 m/s it is 8% more likely to cause imbalance deficit and vice versa for Tartu by 6%. Cloudiness as a predictor is not due to chance, except when it is measured in Võru, then the p value is not small enough to make it statistically significant. In our model cloudiness has a negative impact on imbalance, but for all the cases it is very small, less than 1%.

Next, we added days of the week into our model. Compared to Monday, which is our baseline for a day of the week, there is a 45% increase in the odds of having Estonian imbalance surplus on Saturday and 48% on Sunday.

In addition to days of the week we also added hours to the day to model. Baseline for hours is the hour[0], so 00:00 – 01:00. Out of the 23 hours 15 were had a statistically significant effect on Estonian imbalance. One out of 15, hour[2] increased odds for imbalance surplus by 29%. Rest of the hours all increased the odds for imbalance deficit: hour[7] by 27%, hour[8] by 39%, hour[9] by 36%, hour[12] by 33%, hour[13] by 34%, hour[14] by 31%, hour[15] by 32%, hour[16] by 37%, hour[17] by 50%, hour[18] by 54%, hour[19] by 58%, hour[20] by 48%, hour[21] by 51%, hour[22] by 32%.

Forecasted consumption, forecasted production, imbalance price lag and Estonia imbalance volume lag all had positive effect on Estonia imbalance surplus. Out of the 4 variables the first 3 had a less than 1% increase per unit. The 4th, Estonia imbalance volume lag, increased the odds for imbalance surplus by 1% per 1 MWh. Estonia's SPOT price 24h lag, Baltic imbalance volume lag all had a negative effect on the outcome variable, but also it was recorded as less than 1%.

To forecast Estonian imbalance surplus or deficit (research question 2) two models were used: logistic regression and random forest models. Later both models were compared in terms of

accuracy on average, on holidays and regular days, on all days of the week and all hours of the day to determine created models' weaknesses and strengths. As a result, we managed to achieve 64.12% accuracy with logistic regression forecasting model and 64.40% random forest forecasting model.

Both logistic regression and random forest models showed significant improvement during holidays equaling 70.5% and 68.1% accuracy respectively compared to 63.9% and 64.3% during regular days (Table 2). Thus, holiday variable is important and essential in energy imbalance market analysis.

Table 2. Forecasting models accuracy on holiday and regular day

Holiday	Logistic regression	Random forest
No	63,9%	64,3%
Yes	70,5%	68,1%

Source: author's calculations

Comparing forecasting models' accuracy in days of the week shows that both models are better at predicting Saturday and Sunday. The difference in accuracy is much smoother for random forest model, whereas logistic regression shows noticeable improvement during the weekend compared to working days. (Table 3)

Table 3. Forecasting models accuracy on different days of the week

Day of week	Logistic regression	Random forest
Monday	64,5%	63,9%
Tuesday	63,3%	64,2%
Wednesday	63,0%	63,1%
Thursday	62,7%	63,6%
Friday	62,3%	64,3%
Saturday	66,1%	65,5%
Sunday	67,0%	66,2%

Source: author's calculations

Forecasting models were tested on different day hours (Table 4). Both models performed better at night and early morning hours (23:00 – 6:00) and the worst accuracy occurred during second half of day (15:00-22:00) and in the morning hours (8:00-9:00). Authors find it logical, because night and early mornings are usually less exposed to external factors, not influenced by solar panels production (there's no sun) and electricity consumption is the smallest at those hours.

On the other hand, morning 8 to 9 are usually when humans start their day, wake up and go to work and evening 15 to 22 are the time when people go home and do activities which is hard to predict, thus electricity consumption fluctuates a lot during those hours.

Table 4. Forecasting models accuracy on different hours of the day

Hour	Logistic regression	Random forest
0	64,8%	66,8%
1	67,3%	67,0%
2	75,3%	73,9%
3	67,7%	64,0%
4	70,1%	73,4%
5	72,1%	72,6%
6	69,8%	72,5%
7	61,9%	64,4%
8	59,2%	58,6%
9	59,7%	60,8%
10	67,7%	64,7%
11	67,7%	64,9%
12	68,5%	68,2%
13	66,8%	66,3%
14	65,8%	63,3%
15	63,6%	60,8%
16	61,1%	61,1%
17	57,5%	57,5%
18	57,5%	57,8%
19	57,3%	58,1%
20	55,9%	59,2%
21	55,3%	59,7%
22	59,5%	61,6%
23	66,8%	68,2%

Source: author's calculations

To conclude, both research questions were answered. Factors that influence Estonian imbalance deficit and surplus are air temperature in Jõhvi, windspeed in Tartu and Võru, cloudiness in Võru, days of the week – Saturday and Sunday, hours – 7, 8, 9, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22, forecasted consumption, forecasted production, imbalance price lag and Estonia imbalance volume lag. Estonian imbalance deficit and surplus is possible to predict with 64.12% accuracy with logistic regression forecasting model and 64.40% random forest forecasting model which is rather promising.

USED SOURCES

Republic of Estonia Compeetition Authority. (25.08.2022). Baltic NRAs call for a review of the automatic rise in the price cap for the wholesale electricity market. Used 27. December 2022. https://www.konkurentsiamet.ee/en/news/baltic-nras-call-review-automatic-rise-price-cap-wholesale-electricity-market

Elering LIVE (2022). Used 02. December 2022 https://dashboard.elering.ee/et/system/with-plan/production-consumption?interval=minute&period=years&start=2020-12-31T22:00:00.000Z&end=2021-12-31T21:59:59.999Z

Elering LIVE (2022). Used 02. December 2022 https://dashboard.elering.ee/et/nps/price?interval=minute&period=years&start=2021-12-31T22:00:00.000Z&end=2022-12-31T21:59:59.999Z

Nord Pool (2022). Price calculation https://www.nordpoolgroup.com/en/trading/Day-ahead-trading/Price-calculation/

OpenWeather (2022). Used 02. December 2022 https://openweathermap.org

Eesti Energia annual report 2021.

Eesti Energia III Q 2022 annual report

Baltic Transparency Dashboard (2022). Used 02.December https://baltic.transparency-dashboard.eu/node/44

Baltic Transparency Dashboard (2022). Used 02.December https://baltic.transparency-dashboard.eu/node/42

Skforecast. (2022). Backtesting. Used 28. December 2022. https://joaquinamatrodrigo.github.io/skforecast/0.4.3/notebooks/backtesting.html

APPENDIX

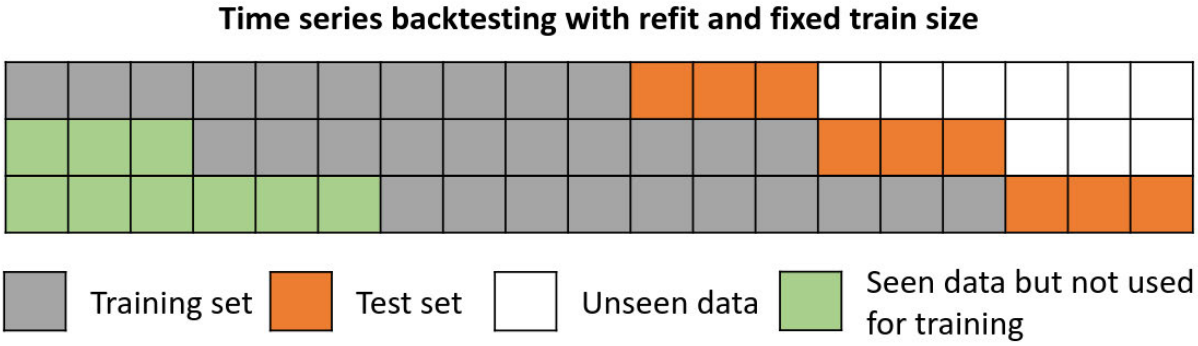
Appendix 1. Logistic regression – code

```
glm.fits <- glm(
  clusters ~ . -time_EET ,
  data = data_for_forecast, family = binomial
)
```

Appendix 2. Forecasting back testing – code

```
library(randomForest)
library(tidyverse)
library(nnet)
#last day in data
max_date <- max(data$time_EET %>% as.Date(), na.rm = TRUE)

#Model test
results_df <- data.frame()
for (day in c(1:365)) {

  #setting dates
  date_of_forecast          <- max_date +1 - day
  end_date_training_set  <- date_of_forecast - 1
  start_date_training_set <- end_date_training_set – 365

  #filter forecasting and training set
  forecast_set <- data_for_forecast %>%
                    filter(as.Date(time_EET) == date_of_forecast)
  training_set <- data_for_forecast %>%
                   filter(as.Date(time_EET) <= end_date_training_set &
                   as.Date(time_EET) >= start_date_training_set)
  #run model
  Logistic_regression_model <- multinom(as.factor(clusters) ~ . -time_EET, training_set)
  Random_forest_model <- randomForest(as.factor(clusters) ~ .-time_EET,
                                  data = training_set, mtry = 12, ntree = 50)

  #predict based on model
  forecast_set$fcst1 <- predict(Logistic_regression_model, forecast_set,  type = "class")
  forecast_set$fcst2 <- predict( Random_forest_model, forecast_set,  type = "class")

  #add result to results_df
  results_df <- rbind(results_df, forecast_set)
}
```

Appendix 3. Time series backtesting with refit and flexed train size



Source: Skforecast (2022).

Appendix 4. Logistic regression model

	clusters		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.89	0.63 – 1.25	0.493
air temperature EE Johvi	1.03	1.01 – 1.06	**0.005**
air temperature EE Parnu	1.02	1.00 – 1.04	0.122
air temperature EE TALLINN	1.01	0.99 – 1.03	0.419
air temperature EE Tartu	0.97	0.93 – 1.01	0.094
air temperature EE Voru	0.99	0.96 – 1.02	0.532
wind speed EE Johvi	0.98	0.95 – 1.01	0.254
wind speed EE Parnu	1.00	0.98 – 1.03	0.787
wind speed EE TALLINN	1.02	0.99 – 1.04	0.167
wind speed EE Tartu	1.06	1.03 – 1.08	**<0.001**
wind speed EE Voru	0.92	0.89 – 0.95	**<0.001**
cloudiness EE Johvi	1.00	1.00 – 1.00	**0.023**
cloudiness EE Parnu	1.00	1.00 – 1.00	**<0.001**
cloudiness EE TALLINN	1.00	1.00 – 1.00	**0.016**
cloudiness EE Tartu	1.00	1.00 – 1.00	**0.001**
cloudiness EE Voru	1.00	1.00 – 1.00	0.539
days of week [2]	1.06	0.94 – 1.20	0.307
days of week [3]	1.04	0.92 – 1.18	0.502
days of week [4]	1.13	1.00 – 1.27	0.053
days of week [5]	1.07	0.95 – 1.21	0.253
days of week [6]	1.45	1.28 – 1.64	**<0.001**
days of week [7]	1.48	1.31 – 1.67	**<0.001**
hour [1]	1.06	0.84 – 1.33	0.610
hour [2]	1.29	1.03 – 1.63	**0.030**
hour [3]	0.97	0.78 – 1.22	0.820
hour [4]	1.04	0.83 – 1.31	0.725

hour [5]	1.17	0.93 – 1.47	0.193
hour [6]	1.04	0.83 – 1.31	0.722
hour [7]	0.73	0.58 – 0.91	**0.006**
hour [8]	0.61	0.49 – 0.76	**<0.001**
hour [9]	0.64	0.51 – 0.81	**<0.001**
hour [10]	0.83	0.66 – 1.04	0.111
hour [11]	0.83	0.66 – 1.04	0.104
hour [12]	0.67	0.53 – 0.84	**<0.001**
hour [13]	0.66	0.52 – 0.83	**<0.001**
hour [14]	0.69	0.55 – 0.87	**0.001**
hour [15]	0.68	0.54 – 0.86	**0.001**
hour [16]	0.63	0.50 – 0.79	**<0.001**
hour [17]	0.50	0.40 – 0.63	**<0.001**
hour [18]	0.46	0.37 – 0.58	**<0.001**
hour [19]	0.42	0.34 – 0.53	**<0.001**
hour [20]	0.52	0.42 – 0.66	**<0.001**
hour [21]	0.49	0.39 – 0.61	**<0.001**
hour [22]	0.68	0.54 – 0.85	**0.001**
hour [23]	0.96	0.77 – 1.21	0.727
holiday [1]	0.92	0.76 – 1.12	0.430
Forecasted consumption	1.00	1.00 – 1.00	**<0.001**
Forecasted production	1.00	1.00 – 1.00	**<0.001**
price EE lag24hours	1.00	1.00 – 1.00	**0.002**
Baltic imbalance volumes lag24hours	1.00	1.00 – 1.00	**0.001**
Estonia imbalance volumes lag24hours	1.01	1.01 – 1.01	**<0.001**
Imbalance price lag24hours	1.00	1.00 – 1.00	**<0.001**
Observations	16775		
R^2 Tjur	0.082		