Analysis of Electricity Price Results and Cross-border Areas Electricity Flow Balances in the Czech Republic

Seminar PaPER
MENE – Energetika, data a IT prostředí přenosových soustav, Winter 2023

Soňa Obůrková
sona.oburkova@email.cz

31th December 2023

INTRODUCTION

The aim of this seminar work is to apply statistical methods studied in the MENE practical seminar on the Czech electricity market data. This work follows given instructions and requirements and has six sections. After the Introduction follows presentation of used data and the data file characteristics. The third section focuses on visualization of data and hypotheses testing. Fourth section discusses results of regression analysis and the fifth one is the time series analysis and prediction. The last section summarizes obtained results. The dataset was downloaded from OTE and contains electricity market prices results and cross-border electricity flow balances. Data are available on a daily basis by hours for every day of January 2023. The seminar work was carried out in R language.

2. DATA AND DATA FILE CHARACTERISTICS

Dataset was dowloaded from OTE, a.s. that is a joint-stock company that serves as an energy market operator for the Czech Republic. On the daily electricity market, OTE is designated by the Nominated Market Organizer (NEMO), which ensures uniform interconnection of daily or intraday markets in accordance with Commission Regulation (EU) 2015/1222. (https://www.ote-cr.cz/cs[1])

We used OTE statistical data on electricity from „Rocni_zprava_o_trhu_2023_V0.xls“ (Source file) that provides large amount of data. We focused specifically on following data:

- Daily electricity market price results per MWh (EUR)

On the daily market, it is possible to anonymously offer or ask for electricity for any hours of the delivery day. The result is closed deals for a set amount of electricity and a uniform deal price for individual hours of the delivery day.(https://www.ote-cr.cz/cs[2])

Data is available by days and hours. We decided for all month of January 2023. Source file list: „DT ČR“, column „Price (EUR)“
→ variable in the seminar work: **‘Price_EUR‘** („Price per MWh (EUR)“ in graphs, pictures etc.)

- Cross-border electricity flow balances

Cross-border electricity flow balances were computed (see details below) using export and import data published in the Source file. Export signifies the electricity volume generated in the Czech Republic and sent to respective neighboring countries. Import, on the other hand, represents the electricity volume generated outside of the Czech Republic and brought into the country.

The data spans January 2023 and is available in hourly increments (measured in MWh). The variables were derived as the balance between the export and import columns for the following neighboring country pairs:

- Czechia and Austria (CZAT)
- Czechia and Germany (CZDE)
- Czechia and Poland (CZPL)
- Czechia and Slovakia (CZSK)

The Source file list "DT ČR Import-Export" encompasses columns "C-J," where the calculated variables 'CZAT', 'CZDE', 'CZPL', 'CZSK' are found.

1 https://www.ote-cr.cz/cs/kratkodobe-trhy/elektrina/files-informace-vdt-vt/trh_s_elektrinou.pdf

2 https://www.ote-cr.cz/cs/kratkodobe-trhy/elektrina/files-informace-vdt-vt/trh_s_elektrinou.pdf

The data are time series and were stored in a new Excel file named "**Seminar_paper_GitHub_20240121.xls**", which contains two lists. The "All_data" list includes data from the Source file along with the calculated new **variables** '**CZAT', 'CZDE', 'CZPL', 'CZSK**'. The list, "**Data_for_analysis,**" holds data specifically used for this seminar paper.

3. VISUALIZATION OF DATA AND HYPOTHESIS TESTING

First, we loaded the 'Data_CZ_price_export_import.xls' file into RStudio and examined all variables to gain a better understanding of their structure and characteristics.

Visualisation of data and hypothesis testing is divided into two sections. The first section is dedicated to descriptive analysis, focusing on detailing our dataset. The second section, exploratory analysis, deals with data exploration, involving the identification of outliers, testing hypotheses related to normal distribution, and investigating relationships between variables.

3.1. Descriptive Analysis

The dataset comprises a total of 744 observations. We'll be analyzing 7 variables, with the first one being in a time format and the remaining variables in numeric format (Table 1). There are no missing data points.

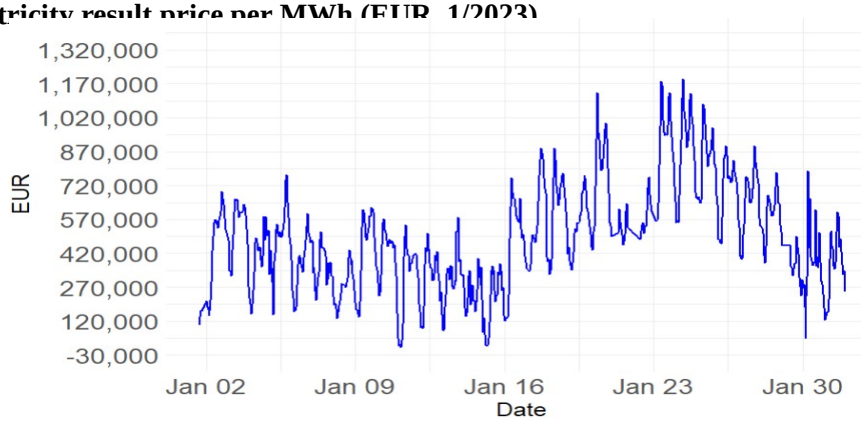
Table 1: Data frame details

```
> str(data_pcbf)
'data.frame':   744 obs. of  7 variables:
 $ Day      : POSIXct, format: "2023-01-01" "2023-01-01" "2023-01-01" ...
 $ Hour     : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Price_EUR: num  16916 -1218 -3418 -6887 -11153 ...
 $ CZAT     : num  2142 2252 2216 2163 2255 ...
 $ CZDE     : num  -62 -709 -751 -508 -790 ...
 $ CZPL     : num  -42 48.5 70.5 -43 45.2 ...
 $ CZSK     : num  871 972 1083 949 1122 ...
```

Source: Own calculation, Rstudio

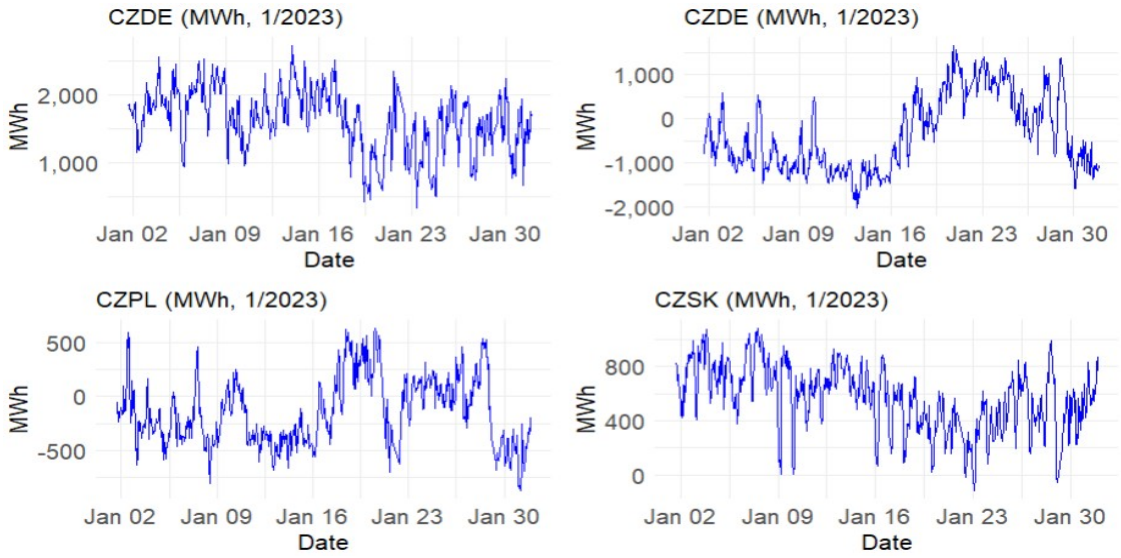
Graph 1 illustrates that the variable 'Price_EUR' exhibits a notably wide range, spanning from -12,709 EUR to 1,267,924 EUR. Regarding the cross-border electricity flow balance variables (depicted in Picture 1), 'CZDE' showcases the widest range of data, varying between -2,023 and 2,052 MWh, while 'CZSK' appears to have the narrowest range, fluctuating from -250 to 1,150 MWh. With the exception of 'CZAT,' all variables encompass both negative and positive values.

Graph 1: Electricity result price per MWh (EUR, 1/2023)



Source: Own calculation, RStudio

Picture 1: Cross-border electricity flows balances (in MWh, 1/2023)



Source: Own calculation, RStudio

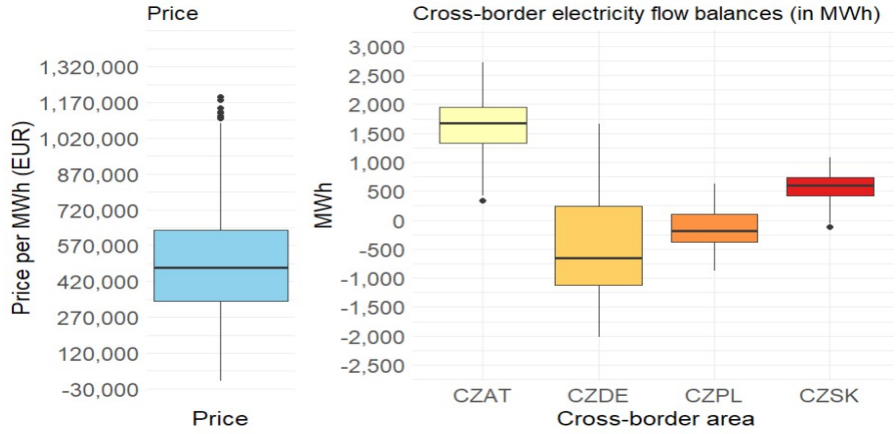
3.1. Exploratory Analysis

After gaining a comprehensive understanding of the dataset through descriptive analysis, we can advance to the exploratory phase. Here, we'll undertake the following steps systematically: identifying and eliminating outlier values (including comparing the datasets before and after the removal of outliers), testing hypotheses on the normal distribution of variables, and analyzing relationships between these variables.

- Identification and Removal of Outlying Values

Let's initiate the identification of outlier values. Picture 2 exhibits boxplot graphs for all five variables. The left side displays the boxplot for the 'Price_EUR' variable (Boxplot “Price”), while the right side illustrates four variables related to cross-border electricity flow balance. These graphs are separated due to variations in the y-axis scales, ensuring clear and readable boxplots. From these representations, it appears that there are outliers in the 'Price' boxplot, as well as in the 'CZAT' and 'CZSK' variables.

Picture 2: Boxplots by variables



Source: Own calculation, RStudio

There are various methods to detect outliers, and in our analysis, we applied Mahalanobis distance. After detecting outliers, observations with a Mahalanobis distance greater than 13 were removed. The RStudio output in Table 2 shows that through 10 iterations, 59 observations were identified and subsequently removed. As a result, the new total number of observations in the dataset is now 685

Table 2: Results of Mahalanobis distance code

```
[1] "Iteration: 1 ; Outlying records removed: 21 ; Cummulative sum of removed records: 21"
[1] "Iteration: 2 ; Outlying records removed: 14 ; Cummulative sum of removed records: 35"
[1] "Iteration: 3 ; Outlying records removed: 7 ; Cummulative sum of removed records: 42"
[1] "Iteration: 4 ; Outlying records removed: 3 ; Cummulative sum of removed records: 45"
[1] "Iteration: 5 ; Outlying records removed: 2 ; Cummulative sum of removed records: 47"
[1] "Iteration: 6 ; Outlying records removed: 2 ; Cummulative sum of removed records: 49"
[1] "Iteration: 7 ; Outlying records removed: 3 ; Cummulative sum of removed records: 52"
[1] "Iteration: 8 ; Outlying records removed: 5 ; Cummulative sum of removed records: 57"
[1] "Iteration: 9 ; Outlying records removed: 1 ; Cummulative sum of removed records: 58"
[1] "Iteration: 10 ; Outlying records removed: 1 ; Cummulative sum of removed records: 59"
```

Source: Own calculation, RStudio

In the next step, we will compare the datasets before and after the removal of outliers.

Graph 2 and Picture 3 depict a comparison between the original and cleaned datasets using boxplots. These visual comparisons are complemented by Table 3 and Table 4, which provide descriptive statistics.

Table 3: Descriptive statistics of original and cleaned data

```
> summary(data_pcbf_original)
   Price_EUR            CZAT          CZDE             CZPL          CZSK
 Min.   : -12709   Min.   : 336   Min.   :-2023.0   Min.   :-872.8   Min.   :-250.2
 1st Qu.: 328371   1st Qu.:1348   1st Qu.:-1089.4   1st Qu.:-389.8   1st Qu.: 401.9
 Median : 462042   Median :1671   Median : -543.4   Median :-181.2   Median : 594.7
 Mean   : 479553   Mean   :1630   Mean   : -333.5   Mean   :-145.6   Mean   : 558.2
 3rd Qu.: 617930   3rd Qu.:1950   3rd Qu.:  353.0   3rd Qu.:  92.5   3rd Qu.: 737.7
 Max.   :1267924   Max.   :2718   Max.   : 2052.6   Max.   : 633.1   Max.   :1150.4
> summary(data_pcbf_clean[3:7])
   Price_EUR            CZAT          CZDE             CZPL          CZSK
 Min.   :   4547   Min.   : 336   Min.   :-2023.0   Min.   :-872.8   Min.   :-114.3
 1st Qu.: 337282   1st Qu.:1334   1st Qu.:-1117.7   1st Qu.:-383.8   1st Qu.: 420.9
 Median : 479090   Median :1672   Median : -653.9   Median :-184.9   Median : 598.4
 Mean   : 495144   Mean   :1623   Mean   : -412.6   Mean   :-137.6   Mean   : 580.7
 3rd Qu.: 636008   3rd Qu.:1949   3rd Qu.:  247.0   3rd Qu.: 102.1   3rd Qu.: 739.4
 Max.   :1192861   Max.   :2718   Max.   : 1654.6   Max.   : 633.1   Max.   :1083.5
```

Source: Own calculation, RStudio

Table 4: Standard deviation of variables before and after outliers removal

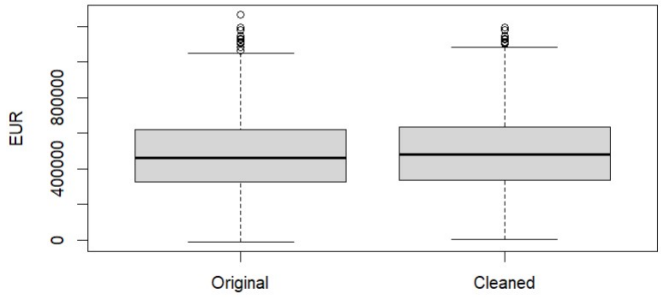
	Price_EUR	CZAT	CZDE	CZPL	CZSK
Original dataset	238,305	453	888	323	268
Cleaned dataset	231,579	459	832	318	233
Difference[3]	6,736	- 5	56	5	35

Source: Own calculations

Graph 2 illustrates that the number of outlier observations for the 'Price_EUR' variable decreased after the data cleaning process. The range between the minimum and maximum values also reduced. Tables 3 and 4 highlight that while the median/mean values slightly increased, the standard deviation only marginally grew.

3 Difference of standard deviations between original and cleaned datasets

Graph 2: Electricity result price per MWh: original vs. cleaned data (EUR, 1/2023)

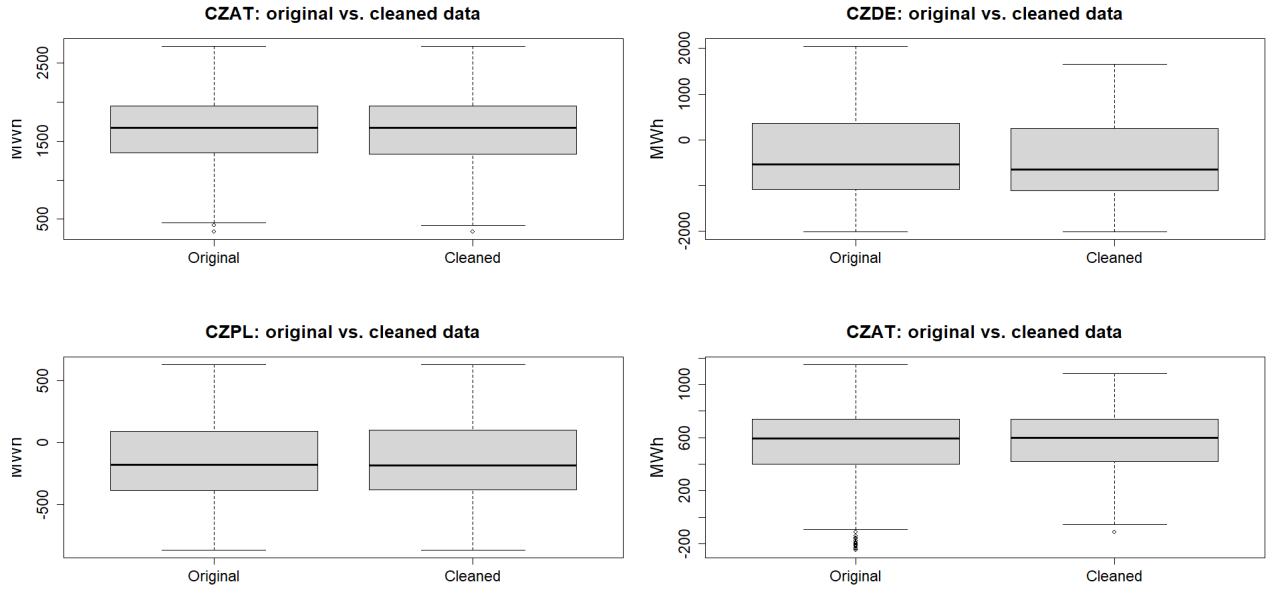


Source: Own calculation, RStudio

For the cross-border electricity flow balance variables, the minimum values remained the same for 'CZAT', 'CZDE', and 'CZPL'. However, in the case of 'CZSK', it increased (from approximately -250 to -114). The maximum values remained unchanged for 'CZAT' and 'CZPL' but decreased significantly for 'CZDE' (from 2052 to 1654) and 'CZPL' (from 1150 to 1083). The most notable change was observed in the median value of 'CZDE' (from -543 to -653).

Regarding the difference between the standard deviations of the original and cleaned datasets, the largest difference was seen in the 'CZDE' variable (56 MWh), followed by 'CZSK' (35 MWh). In contrast, the differences for 'CZAT' and 'CZPL' variables were much smaller (-5 MWh and 5 MWh, respectively).

Picture 3: Cross-border electricity flow balances by areas: original vs. cleaned data (MWh, 1/2023)



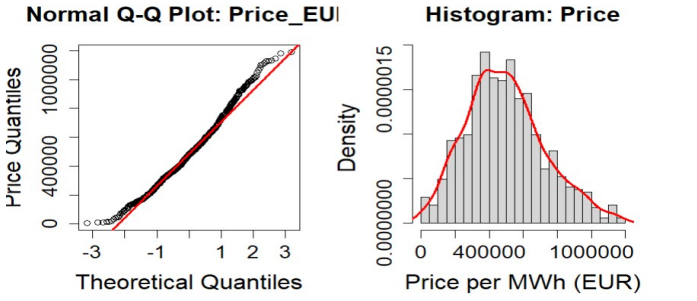
Source: Own calculation, RStudio

- Hypotheses Testing on Normal Distribution of Variables

The normality of variable distributions was assessed using graphical methods like Q-Q plots, which compare quantiles to those of a theoretical normal distribution, alongside histograms. Additionally, the Kolmogorov-Smirnov test was employed to assess normal distribution. The null hypothesis (H0) posits that the variable follows a normal distribution, while the alternative hypothesis (HA) suggests otherwise.

In the Q-Q plot and in the histogram we can see that the 'price_EUR' variable has in general normal distribution, though left end tends to go off and even stronger deviation from the normal distribution can be seen on the right end. Nevertheless when applying Kolmogorov-Smirnov test on normal distribution we accept H0 as statistically significant at the five percent level of significance (p-value > 0.05). (Picture 4)

Picture 4: Variable 'Price_EUR' – normal distribution testing



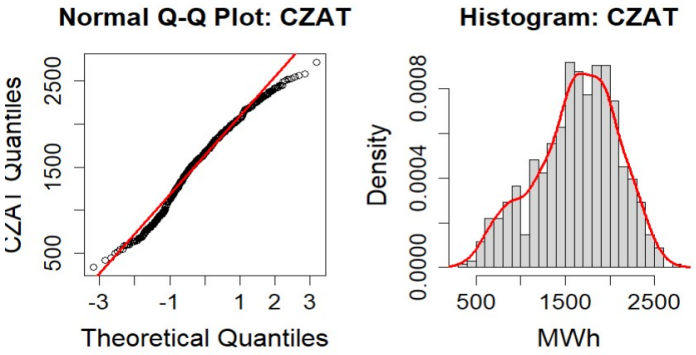

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data_pcbf_clean$Price
D = 0.042523, p-value = 0.1679
alternative hypothesis: two-sided
```

Source: Own calculation, RStudio

In regards of 'CZAT' variable, we can see greater deviation from the normal distribution on Q-Q plot and on histogram comparing to 'price_EU' variable. Still the Kolmogorov-Smirnov test on normal distribution is greater than 0.05 so we accept H0 as statistically significant at the five percent level of significance (p-value > 0.05). (see Picture 5)

Picture 5: Variable 'CZAT' – normal distribution testing

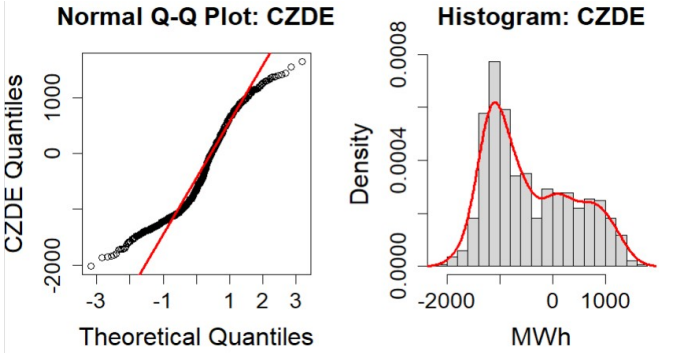



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data_pcbf_clean$CZAT
D = 0.050862, p-value = 0.05778
alternative hypothesis: two-sided
```

Source: Own calculations, RStudio

Picture 6: Variable 'CZDE' – normal distribution testing




```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data_pcbf_clean$CZDE
D = 0.12757, p-value = 4.146e-10
alternative hypothesis: two-sided
```

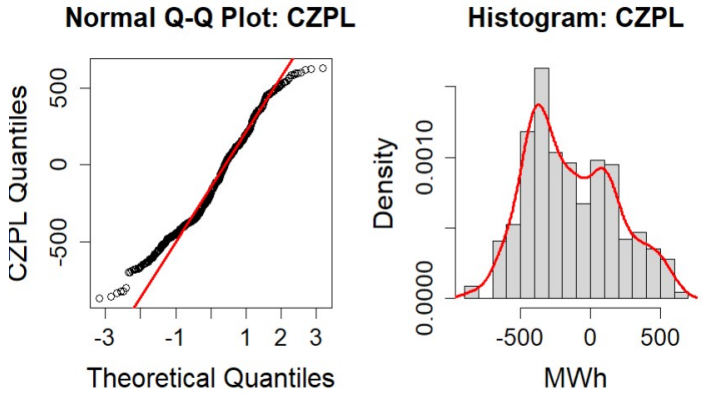
Source: Own calculations, RStudio

Both the Q-Q plot and histogram indicate that the 'CZDE' variable does not follow a normal distribution. Furthermore, this was corroborated by the Kolmogorov-Smirnov test on normal distribution, where the H0 was rejected due to statistical insignificance at both the five percent level of significance (p-value < 0.05) and the one percent level of significance (p-value > 0.01). Hence, the alternative hypothesis (HA) was accepted. (see Picture 6)

The situation observed with the 'CZPL' variable is similar to that of 'CZDE'. Graphs indicate a lack of normal distribution, which was confirmed by the Kolmogorov-Smirnov test on normal distribution. H0 was rejected as statistically insignificant at both the five percent level (p-value < 0.05) and the one percent level (p-value < 0.01), leading to the acceptance of the HA regarding the non-normal distribution of the variable. (see Picture 7).

Conversely, the 'CZSK' variable demonstrates normal distribution based on the Q-Q plot and histogram, further validated by the Kolmogorov-Smirnov test on normal distribution, where H0 was accepted as statistically significant at the five percent level (p-value > 0.05). (see Picture 8)

Picture 7: Variable 'CZPL' – normal distribution testing

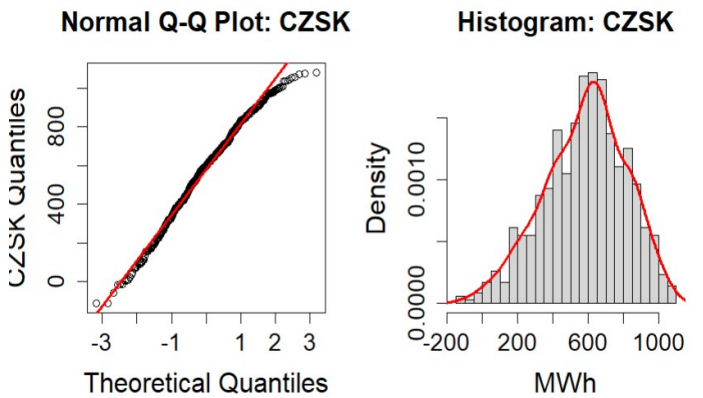



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data_pcbf_clean$CZPL
D = 0.083339, p-value = 0.0001475
alternative hypothesis: two-sided
```

Source: Own calculations, RStudio

Picture 8: Variable 'CZSK' – normal distribution testing




```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  data_pcbf_clean$CZSK
D = 0.050449, p-value = 0.0612
alternative hypothesis: two-sided
```

Source: Own calculations, RStudio

Upon assessing normal distribution among the variables, it appears that 'Price_EUR', 'CZAT', and 'CZSK' exhibit normal distribution characteristics. However, for 'CZDE' and 'CZPL', the hypothesis of normal distribution was rejected.

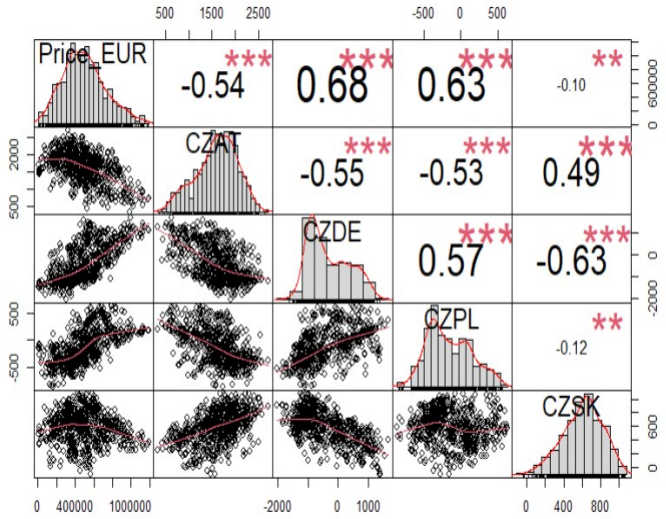
- Analysis of Relationships Between Variables.

The relationships between the analyzed variables are depicted in the correlogram shown in Graph 3. Below the diagonal, scatter plots illustrating the relationships between every pair of variables are presented. Above the diagonal, the coefficient of correlation (R^2) values for each pair of variables is provided. Additionally, Graph 4 displays detailed scatter plots specifically focusing on the relationship between 'Price_EUR' and 'CZAT', 'CZDE', 'CZPL', and 'CZSK' variables.

The results indicate a negative and relatively strong relationship (R^2 = -0.54) between the 'Price_EUR' variable and the 'CZAT' variable. Similarly, the relationships between 'Price_EUR' and 'CZDE', as well as 'CZPL', are strong (R^2 = 0.68 and 0.63, respectively), but positive. This suggests that an increase in 'CZDE' or 'CZPL' variables would result in an increase in the 'Price_EUR' variable. On the contrary, the relationship between 'Price_EUR' and 'CZSK' is negative and very weak (R^2 = -0.10), indicating that an increase in 'CZSK' would have a minimal impact on the growth of 'Price_EUR'.

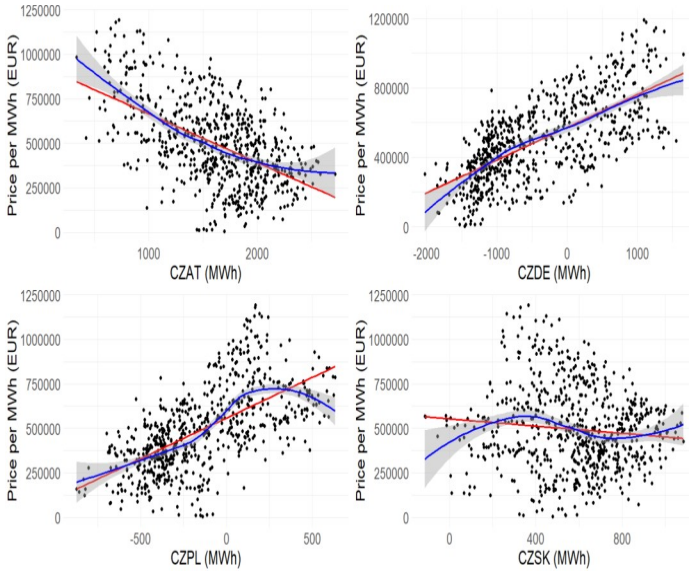
Regarding the relationships between the cross-border electricity flow balance variables ('CZAT', 'CZDE', 'CZPL', 'CZSK'), the coefficients of correlation are relatively high, ranging from -0.53 to -0.63 for negative relationships and from 0.12 to 0.57 for positive relationships. The weakest relationship is observed between 'CZPL' and 'CZSK' variables (R^2 = -0.12), as evident in the correlogram scatter plot. Following this is the relationship between 'CZAT' and 'CZSK' variables, with a strength (R^2) of 0.49 in this case.

Graph 3: Correlogram of variables



Source: Own calculations, RStudio

Graph 4: Scatter plots of relationships between variable 'Price_EUR' and 'CZAT', 'CZDE', 'CZPL' and 'CZSK' variables



Source: Own calculations, RStudio

4. REGRESSION ANALYSIS

Regression analysis helps understand the relationship between one or more independent variables and a dependent variable. The primary objective is to determine whether a relationship exists and, if so, to quantify its strength.

Subsequently, we employed our OTE dataset in a multilinear regression model using the ordinary least square method. The dependent variable was 'Price_EUR', and the independent variables were the cross-border electricity flow balance variables 'CZAT', 'CZDE', 'CZPL', and 'CZSK'.

The initial multilinear regression model (Model 1) was defined as follows:

lm(Price_EUR ~ CZAT + CZDE + CZPL + CZSK, data_pcbf_clean)

In the model validation (refer to the attached R code), it was observed that the variable 'CZPL' was statistically significant at a five percent level of significance (p-value > 0.05). Therefore, the null hypothesis (H0) on the dependent variable coefficient being zero was accepted, suggesting a weak or no relationship between 'Price_EUR' and 'CZPL'. Consequently, 'CZPL' was excluded from Model 1, and Model 2 was formulated:

lm(Price_EUR ~ CZAT + CZDE + CZSK, data_pcbf_clean)

Validation of Model 2, illustrated in Picture 9, revealed that the p-values (Pr(>|t|)) for all Model 2 variables were not statistically significant at the five percent level (p-value < 0.05). Thus, we rejected H0 (indicating multilinear regression coefficients are zero) and accepted the alternative hypothesis (HA): coefficients are not zero. Regarding the F-statistics, assessing the overall significance of the model, a larger F-statistic with a small p-value implies the model's significance, which is confirmed in our case. The multilinear R-squared value of 0.78 indicates a good fit, considering 1 as a perfect fit.

Picture 9: Summary of the linear Model 2

```
Call:
lm(formula = Price_EUR ~ CZAT + CZDE + CZSK, data = data_pcbf_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-404704  -77311    2688   77895  390691

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 539595.42   20852.92   25.88   <2e-16 ***
CZAT          -189.01      12.38  -15.27   <2e-16 ***
CZDE           241.36       7.63   31.63   <2e-16 ***
CZSK           623.29      26.20   23.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 121000 on 681 degrees of freedom
Multiple R-squared:  0.728,     Adjusted R-squared:  0.7268
F-statistic: 607.5 on 3 and 681 DF,  p-value: < 2.2e-16
```

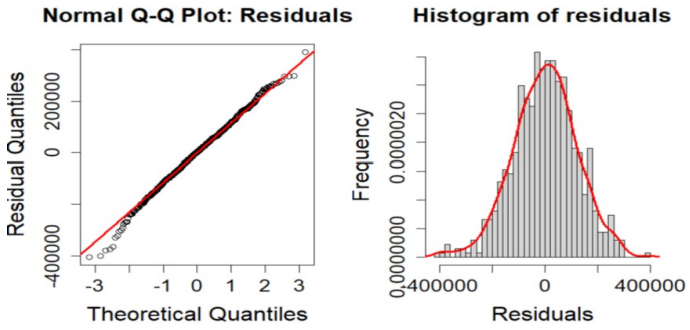
Source: Own calculation, RStudio

Residual analysis is a crucial step following the creation of a regression model. It is employed to assess whether the residuals, representing the differences between actual and predicted values, adhere to the assumptions of linear regression. These assumptions encompass the normal distribution of residuals, homoscedasticity, and lack of autocorrelation. The last examined assumption is on the presence of multicollinearity among the variables. Each assumption is examined below:

- Normal Distribution

The normal distribution of residuals for Model 2 was scrutinized in a manner similar to the examination of variables in the exploratory analysis above. Graphs presented in Picture 10 illustrate that the residuals exhibit a normal distribution. This observation was further corroborated as statistically significant at the five percent level of significance (p-value > 0.05) by the Shapiro-Wilk normality test (p-value = 0.1437). (see Picture 10)

Picture 10: Tests on normal distribution of residuals

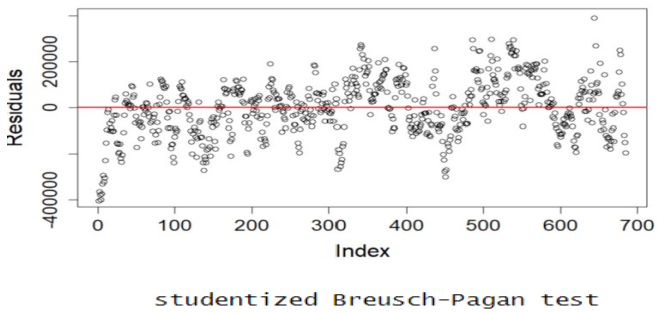



```
        Shapiro-Wilk normality test

data:  model_2$residuals
W = 0.99654, p-value = 0.1437
```

Souce: On calculations, RStudio

Picture 11: Tests on heteroscedasticity of residuals




```
        studentized Breusch-Pagan test

data:  model_2
BP = 24.295, df = 3, p-value = 0.00002168
```

Source: Own calculations, RStudio

- Heteroscedasticity

Homoskedasticity was assessed both graphically through a scatter plot and by applying the Breusch-Pagan test. The scatter plot revealed heteroskedasticity, indicating a lack of constant variance in the residuals. Subsequently, the Breusch-Pagan test[4] confirmed the presence of heteroskedasticity. The rejection of the null hypothesis (H0), which posits homoscedasticity (constant variance of residuals), was statistically significant at the five percent level (p-value < 0.05) (see Picture 11).

- Autocorrelation

Autocorrelation, relevant in time series regression analysis, refers to the presence of correlation between the error terms of a regression model, which may occur when data points are not independent. The Durbin-Watson test was conducted with the null hypothesis (H0) stating there is no first-order autocorrelation in the residuals.. The alternative hypothesis suggests there is first-order positive or negative autocorrelation in the residuals.. The test results led to the rejection of H0, with statistical significance at the five percent level (p-value < 0.05), supporting the acceptance of the alternative hypothesis that residuals exhibit positive or negative autocorrelation. (see Picture 12)

Picture 12: Durbin-Watson test on autocorrelation

```
        Durbin-Watson test

data:  model_2
DW = 0.29275, p-value < 0.00000000000000022
alternative hypothesis: true autocorrelation is greater than 0
```

Source: Own calculation, RStudio

- Multicollinearity

Multicollinearity of variables was previously analyzed in section 3.2, Exploratory Analysis above. The data suggest that the absolute value of the coefficient of correlation remains below 0.63, which is deemed acceptable for considering the model to be free of multicollinearity.

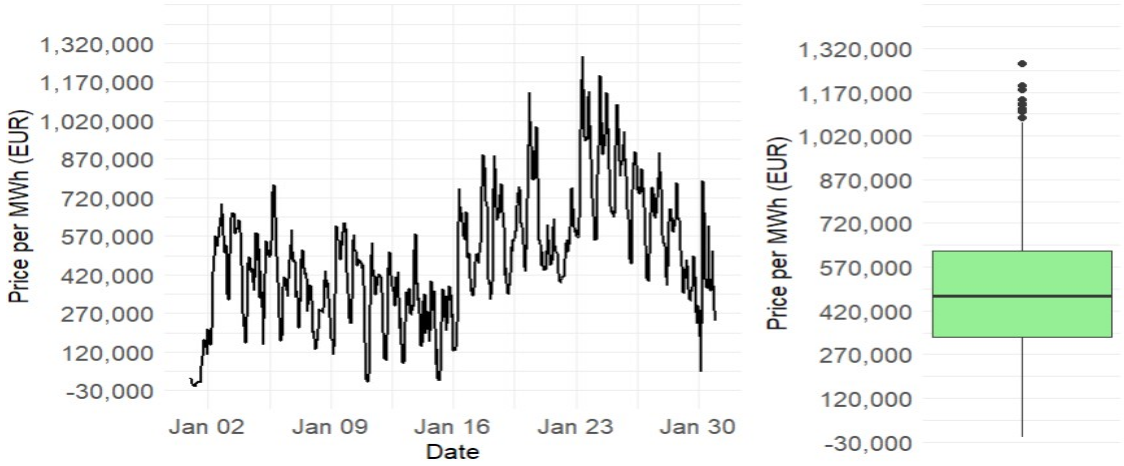
In response to assumptions outputs above, a reasonable approach would be to continue refining the model. It's important to note that addressing heteroscedasticity and autocorrelation may require more advanced methods, and while acknowledging their presence, addressing these issues is beyond the scope of the current work.

5. TIME SERIES ANALYSIS

In this section, we analyzed the time series 'Price_EUR', representing electricity market price results per MWh available by days and hours for January 2023. This dataset has been utilized consistently throughout all seminar work. The time series was split into the training set (from 1st January 00:00:00 untill 30st January 23:00:00) and the test time series (from 31st January 00:00:00 untill 31st January 23:00:00). The structure of the training time series is illustrated in Graph 5.

4 Null hypothesis (H0): The variance of the residuals is constant (there is homoskedasticity of residuals) Alternative hypothesis (HA): The variance is not constant (there is heteroscedasticity of residuals).

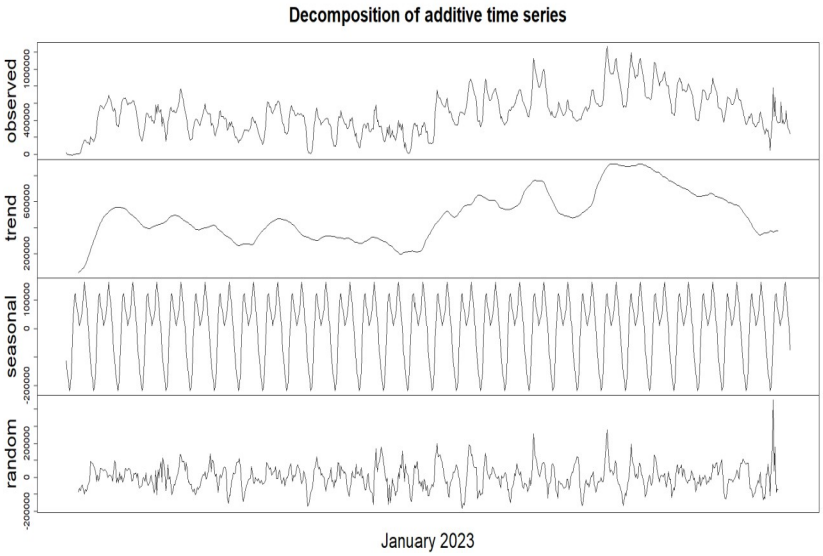
Graph 5: Structure of 'Price_EUR' variable time series – line graph and boxplot



Source: Own calculation, RStudio

The decomposition of a time series into additive components can aid in comprehending and modeling the underlying patterns, trends, and irregularities within the data, thereby facilitating forecasting and analysis. The decomposition shown in Graph 6 reveals a decreasing trend in the first half of the month and growth in the other half. Additionally, clear seasonality can be observed with an estimated frequency of 1 day.

Graph 6: Decomposition of additive time series, variable Price_EUR



Source: Own calculations, RStudio

Picture 13: Dickey-Fuller test on stationarity

```
        Augmented Dickey-Fuller Test

data:  train_data_ts
Dickey-Fuller = -4.4819, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

Source: Own calculations, RStudio

The Dickey-Fuller test determines the presence of a unit root in a time series dataset, indicating non-stationarity, where the mean and variance change over time. The null hypothesis (H0) posits the existence of a unit root, signifying non-stationarity, while the alternative hypothesis (HA) suggests the absence of a unit root, implying stationarity. (see Picture 13)

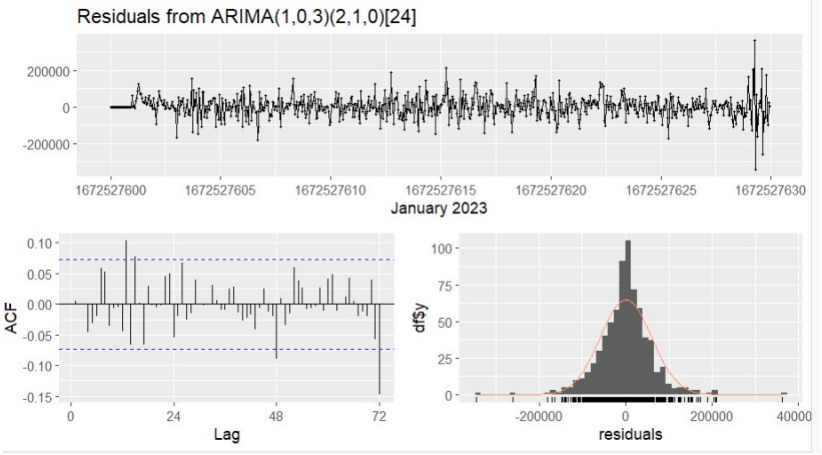
Based on the test results, the null hypothesis (H0) was rejected as statistically insignificant at the five percent level of significance (p-value < 0.05), and the alternative hypothesis (HA) was accepted. Consequently, the time series is deemed stationary.

The Auto.arima function automatically determined the optimal ARIMA model for our time series dataset. The resulting model was identified as ARIMA(1,0,3)(2,1,0)[24], indicating a combination of a non-seasonal ARIMA(1,0,3) model and a seasonal ARIMA(2,1,0) model, with a seasonal model difference (D parameter) set to 1.

The primary objective of diagnosing time series residuals is to ensure randomness, constant variance, normal distribution, and absence of autocorrelation. Deviations from these characteristics may indicate issues with model assumptions or suggest areas for improvement.

Residual diagnostics are presented in Graph 7. Picture 14 and Picture 15. The top plot in the Graph 7 assesses whether residuals exhibit a pattern. Ideally, no clear pattern should be present, but there seem to be regular spikes in the time series of residuals. The ACF plot in Graph 7 displays the blue line representing the 95% confidence interval. Spikes outside these intervals in the ACF plot may suggest non-randomness of residuals. The histogram checks the distribution of residuals, which appears to resemble a normal distribution, aligning with our time series model.

Graph 7: Residual diagnostics



Source: Own calculations, RStudio

Picture 14: Jarque-Bera test on normal distribution of residuals

```
        Jarque Bera Test

data:  model_aa_train_output1$residuals
X-squared = 587.47, df = 2, p-value < 0.00000000000000022
```

***S**ource: Own calculations, RStudio*

Picture 15: Ljung-Box test on autocorrelation of residuals

```
        Box-Ljung test

data:  model_aa_train_output1$residuals
X-squared = 0.018785, df = 1, p-value = 0.891
```

Source: Own calculations, RStudio

We also conducted the Jarque-Bera test of normality (Picture 14), which resulted in the rejection of the null hypothesis (H0) on the normal distribution of residuals and accepting alternative hypothesis (HA) on the residuals not having normal distribution being statistically insignificant at the five percent level (p-value < 0.05).

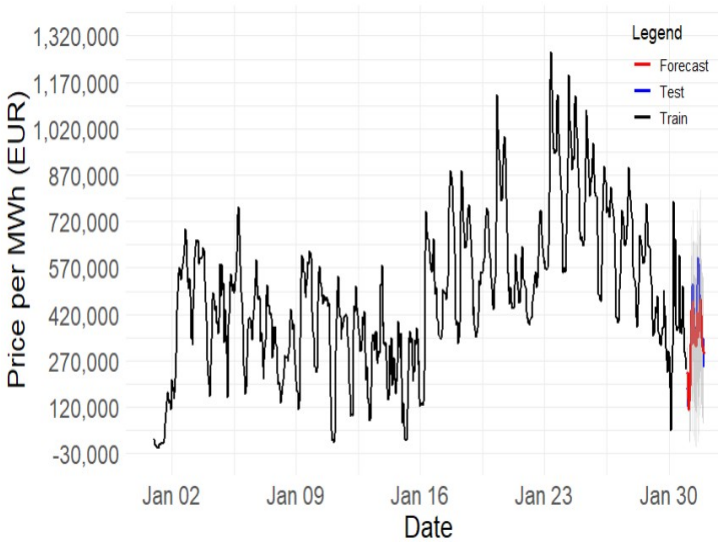
Conversely, the Ljung-Box test (Picture 15) rejected the null hypothesis (H0): The residuals in the time series are independently distributed; there is no autocorrelation. The results were not statistically significant at the five percent level (p-value > 0.05). Therefore, the alternative hypothesis (HA): The residuals in the time series are not independently distributed; there is autocorrelation, was accepted.

Despite the diagnostics of residuals that indicated the model not to fully meet assumptions on randomness, constant variance or normal distribution, we have chosen to proceed with the model for forecasting.

The Mean Absolute Percentage Error (MAPE) assesses the accuracy of a forecasting model, especially in the context of time series forecasting. It quantifies the average percentage difference between the predicted values and the actual values in a dataset. Our forecast's MAPE is 12.8, indicating that, on average, the forecast deviates by almost 13% from the actual values.

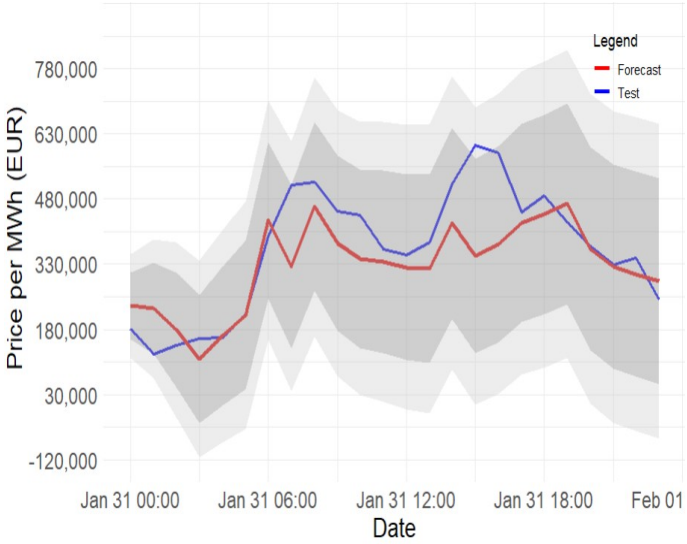
In Graph 8, the line graph illustrates both the training and test datasets, along with the forecast for January 31, 2023, accompanied by confidence intervals of 80% (light grey) and 95% (dark grey). Additional details about the forecast, including the test data and 80% and 95% confidence intervals, are provided in Graph 9. The data reveals that real values closely correspond to the forecast values and generally fall within the predicted 95% confidence interval throughout the day, except for brief deviations around 6 a.m. and 3 p.m.

Graph 8: Electricity prices results and their forecast (EUR, 1/2023)



Source: Own calculations, RStudio

Graph 9: Detailed graph of electricity prices results forecast (EUR, 31/1/2023)



Source: Own calculations, RStudio

6. CONCLUSION

The dataset utilized in this seminar work includes a time series variable reflecting daily electricity market price results, along with four variables showcasing daily cross-border electricity flow balances ('CZAT', 'CZDE', 'CZPL', 'CZSK'). These variables, also time series in nature, provide a geographical context that is integral to the focus of this seminar work.

Visualization of data was performed throughout the work using various graph types such as linear graphs, scatter plots, boxplots, Q-Q plots, histograms, and correlograms. Hypothesis testing was conducted using several tests, including the Kolmogorov-Smirnov test for normal distribution, Shapiro-Wilk test for normality, Breusch-Pagan test for heteroscedasticity, Jarque-Bera test for normal distribution of residuals, Durbin-Watson test for autocorrelation, and Dickey-Fuller test for stationarity.

The second version of the regression model, utilizing the dependent variable of daily electricity market price results and independent variables of cross-border electricity flow balances 'CZAT', 'CZDE', 'CZSK', was discussed, and residual testing was conducted. The analysis revealed that the model has its limitations, and further refinement in model setup is warranted.

Time series analysis was performed on the daily electricity market price results time series variable from Source file. An ARIMA model was obtained using the auto.arima function. Subsequent residual testing

indicated that not all assumptions were met. Nevertheless, the forecast function was applied, and the results were visualized in graphs

In this seminar work, we had the opportunity to practice various data visualization techniques in the R programming language, as well as regression and time series analysis. I find this seminar work useful, interesting, practical and enjoyable.

LITERATURE

Lectures from MDAV – Analýza dat a jejich vizualizace (Winter 2022)
Lectures from MECM – Ekonometrické modely (Summer 2023)
Lectures from MENE – Energetika, data a IT prostředí přenosových soustav (Winter 2023)
Lectures from MMDA – Statistické metody v analýze dat (Winter 2022)
E-CITACE. [Trh s elektřinou] In: OTE [online]. Cit. 2023-12-09. Available from: https://www.ote-cr.cz/cs/kratkodobe-trhy/elektrina/files-informace-vdt-vt/trh_s_elektrinou.pdf

DATA

OTE, „Rocni_zprava_o_trhu_2023_V0.xls“. Downloaded 2023-12-09 from https://www.ote-cr.cz/cs/statistika/rocni-zprava?date=2023-01-01

LIST OF TABLES, GRAPHS AND PICTURES