National Research University Higher School of Economics

Faculty of Computer Science

Programme "Master of Data Science"

MASTER'S THESIS

Anomaly detection in Time series

Student: Izmalkin Aleksey Aleksandrovich

Supervisor: Kasianova Kseniya Alekseevna

Moscow, 2023

Abstract

Time series anomalies are defined as observations that exhibit a considerable deviation from the anticipated pattern or trend in the time series data. The occurrence of anomalies can be attributed to a multitude of factors, including unforeseen circumstances, inaccuracies in measurement, or technical malfunctions.

Anomalies possess the potential to adversely affect the precision of predictive modeling. The fundamental premise of time series models is predicated on the supposition that the intrinsic data conforms to a particular pattern or trajectory. In the event of aberrations in the data, the models may fail to effectively capture the fundamental patterns, resulting in suboptimal predictive efficacy. The issue can be especially challenging in sectors such as finance, healthcare, transportation, and energy, where precise forecasts are indispensable for informed decision-making.

The initial stage of this research is to disintegrate the time series data into its distinct constituents, comprising trend, seasonality, and residual, and subsequently scrutinize each constituent to acquire a more profound comprehension of the characteristics and dynamics of the time series. Subsequently, a variety of techniques for detecting and measuring anomalies in the data will be employed.

Upon detection of anomalies, the resultant data shall be utilized to construct predictive models for future pricing. The primary objective of this study is to assess the influence of anomalies on the precision of price forecasting. In order to accomplish this objective, we will conduct a comparative analysis of the efficacy of predictive models trained on datasets with and without anomalies. The aim of this analysis is to ascertain the impact of anomalies on the precision of price forecasts and the overall performance of predictive models.

The research aims to enhance the analysis and prediction of time series data, potentially leading to the creation of more precise and dependable energy pricing models. This could be advantageous for energy companies as it would enable them to make well-informed decisions based on accurate predictions and risk evaluations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and significance of the topic

Anomalies, commonly referred to as outliers, are data points that exhibit a substantial deviation from the remaining data. They can occur in any type of data, including time series data, and can have a significant impact on the accuracy of predictive models. There are many reasons why anomalies may exist in a dataset, including measurement errors, data entry errors, data processing errors, or simply because they represent unusual or rare events.

The presence of anomalies can pose a challenge as they have the potential to distort the outcomes of statistical evaluations, thereby resulting in imprecise forecasts and deductions. During the process of training predictive models, the presence of anomalies can result in overfitting. This phenomenon arises when the model becomes excessively tailored to the training data, thereby losing its ability to generalize effectively to novel data.

In order to address these concerns, it is imperative to identify and eliminate aberrations from the dataset prior to the development of prognostic models. There are several techniques for detecting anomalies, including statistical methods such

as Z-score and IQR, clustering methods, and machine learning algorithms. Once anomalies have been detected, they can be removed from the data using techniques such as data smoothing or interpolation.

By detecting and removing anomalies from the data, predictive models can be trained on more accurate and representative data, resulting in more accurate predictions and better performance. In brief, the identification and elimination of aberrations from temporal data sets constitute a pivotal phase in the training of prognostic models, and have the potential to significantly enhance the precision and dependability of such models.

The goal of this work is to gain a deeper understanding of how anomalies in time series data can affect the quality of price prediction models. Through conducting comprehensive analyses and assessing the resultant models, this study aims to offer valuable insights into enhancing the precision and dependability of price forecasting models in the presence of anomalies.

1.2 Structure of the work

The present thesis will be organized into four primary chapters, each with its own distinct structure. Chapter 2 will involve the processing and analysis of the daily electricity price, which will be decomposed into its constituent components. The residual component will be given special attention, with a focus on its stationarity. In order to ascertain the stationarity of the residual in the time series, we will conduct suitable testing.

Chapter 3 will examine diverse techniques for identifying anomalies in time series data. The efficacy of these techniques will be evaluated and the distinctions between them will be illustrated through visualization.

Chapter 4 will involve advancing to the prediction phase by utilizing the data

acquired from the preceding chapters. The investigation will also encompass an analysis of the influence of anomalies on the training procedure and the accuracy of our prognostications. The objective of this framework is to offer a thorough examination of anomaly detection in time series and its influence on the precision of predictions.

The research culminates in Chapter 5, wherein we will delineate the principal findings of our analysis and deliberate upon the ramifications of our results. Additionally, we will examine potential perspective for future investigation in this domain. The present chapter will function as a conclusive section of our study, furnishing a recapitulation of our discoveries and delineating the advancements that this investigation has brought to the domain of time series analysis and anomaly detection.

1.3 Related work

This study draws upon the seminal research conducted by Varun Chandola, Arindam Banerjee, and Vipin Kumar, titled "Anomaly Detection: A Survey" [1], which offers a comprehensive examination of the anomaly detection domain.

The present study examines diverse methodologies and strategies for detecting anomalies in time series data, encompassing statistical techniques, clustering-based techniques, classification-based techniques, and nearest neighbor-based techniques. The significance of preprocessing and feature extraction in the anomaly detection procedure is also emphasized by the authors.

Chapter 2

Time series analysis and its components

2.1 Data preparation and EDA

This section outlines the way employed for data preparation in the analysis of the equilibrium price index for electricity procurement, denominated in Russian rubles per megawatt-hour (RUB/MWh). The dataset was obtained from the official website of ATS Energo `https://www.atsenergo.ru/results/rsv/index` and comprises daily values of the aforementioned index spanning from April 2020 to April 2023. It consists of two separate datasets: one from the European part of Russia and the other from Siberia.

	Date	VTPC	EPIPE	PVURC	PVD	SVCRC	MaxEPI	MinEPI
0	05.04.2020	2011051.007000	1136.090000	400667.334000	1464802.233000	2756.988000	1399.900000	863.240000
1	06.04.2020	2053895.126000	1279.860000	400667.334000	1504613.045000	2682.265000	1597.000000	843.480000
2	07.04.2020	2041359.284000	1121.920000	400667.334000	1496706.986000	2951.107000	1417.780000	806.170000
3	08.04.2020	2033106.623000	1060.130000	400667.334000	1487836.262000	2922.426000	1391.270000	747.740000
4	09.04.2020	2023396.298000	1184.060000	400667.334000	1477808.105000	3190.166000	1427.300000	780.860000

Figure 2.1: European part

	Date	VTPC	EPIPE	PVURC	PVD	SVCRC	MaxEPI	MinEPI
0	05.04.2020	522796.483000	1030.620000	85518.334000	418946.716000	613.811000	1087.050000	954.870000
1	06.04.2020	526922.971000	1026.380000	85518.334000	423502.349000	527.945000	1170.340000	866.920000
2	07.04.2020	524266.260000	959.700000	85518.334000	420832.412000	816.114000	1064.310000	851.460000
3	08.04.2020	522847.708000	888.510000	85518.334000	419700.103000	611.128000	944.780000	841.710000
4	09.04.2020	533857.529000	887.760000	85518.334000	430573.229000	615.296000	955.370000	829.910000

Figure 2.2: Siberia

The dataset encompasses various variables, including but not limited to the Equilibrium price index (EPIPE), Volume of total planned consumption (VTPC)in MWh, Purchase volume under regulated contracts (PVURC) in MWh, Purchase volume at DAM (PVD(in MWh, Sales volume as collateral for RC (SVCRC) in MWh, Maximum equilibrium price index for the period (MaxEPI) in RUB/MWh, and Minimum equilibrium price index for the period (MinEPI) in RUB/MWh. For the purpose of this analysis, our focus will be solely on the Equilibrium price index (EPIPE).

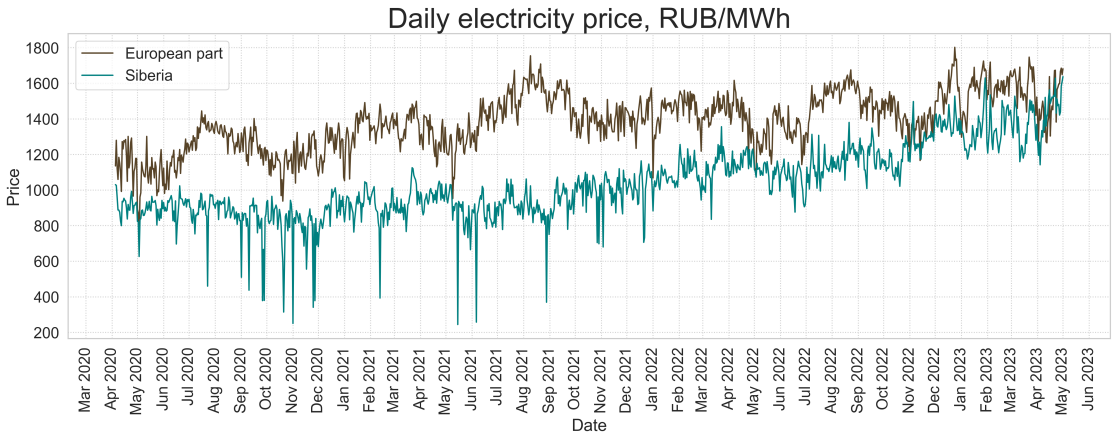

Figure 2.3: Equilibrium price index (EPIPE)

Prior to conducting data analysis, preliminary assessments were conducted to verify the integrity and coherence of the data. The data was thoroughly examined for missing values, outliers, and inconsistencies. The distribution of the data was

evaluated and basic exploratory data analysis was conducted to obtain insights into the data's characteristics.

After completing the necessary data cleaning and preparation procedures, we proceeded with the analysis of the Equilibrium price index (EPIPE). The ensuing section will explicate the preliminary analysis of the aforementioned variable and proffer some initial findings.

2.2 Time series decomposition

Time series data must be decomposed in order to analyze and comprehend the fundamental components that contribute to the overall patterns and trends. We can separate a time series into its trend, seasonality, and residual components by decomposing it. This decomposition procedure provides valuable insights into the individual characteristics and behaviors of these components, allowing for a more in-depth examination of the time series data.

The trend component in time series analysis represents the long-term behavior or general direction of the data. It enables us to identify any consistent upward or downward trends over time, thereby shedding light on the overall growth or decline patterns. By isolating the trend component, we can obtain a deeper comprehension of the time series' underlying dynamics and long-term changes.

The seasonal component depicts the recurring patterns that occur at regular intervals within the time series. These patterns can be daily, weekly, monthly, or any other cyclical variations that occur consistently. Analyzing the seasonality component enables us to identify and interpret periodic fluctuations, thereby giving insight on the data's periodic behaviors and patterns.

The residual component, also known as the error component, represents the variation in the time series that cannot be explained by the trend or seasonality. It

consists of random fluctuations, outliers, and any other irregularities not accounted for by the other factors. By analyzing the residual component, we can identify any remaining patterns, anomalies, or out-of-the-ordinary observations that may require additional investigation.

There are two common methods for decomposing time series: additive and multiplicative decomposition. In additive decomposition, the time series is expressed as the sum of its components, presuming that the magnitude of seasonal fluctuations remains constant over time [2].

$$y_t = S_t + T_t + R_t,$$

where y_t is the data, S_t is the seasonal component, T_t is the trend componennt and R_t is the reminder component, all at period t. Multiplicative decomposition, on the other hand, represents the time series as the product of its components, under the assumption that seasonal fluctuations are proportional to the trend.

$$y_t = S_t \times T_t \times R_t$$

The selection between additive and multiplicative decomposition is dependent on the characteristics and nature of the data. Several factors support the selection of additive decomposition for our daily electricity price dataset. As the observed data is expressed as the sum of the trend, seasonality, and residual components, additive decomposition provides a straightforward and intuitive interpretation of the individual components.

This additive relationship facilitates a clear comprehension of how each component contributes to the overarching behavior of the time series. In addition, additive decomposition implies that the magnitude of seasonal fluctuations remains constant over time, which is consistent with the expectation of relatively stable amplitudes in the seasonal patterns observed in the daily electricity pricing data.

For this research, we will utilize the "Statsmodels" library. The given library is a robust Python module that provides an extensive collection of tools and functions for statistical modeling, including time series analysis [3].

2.2.1 Trend component

The trend component of a time series is crucial to comprehending the long-term behavior and general direction of the data. In this study, the "seasonal decompose()" function and the "trend" method from the Statsmodels library [3] are used to extract the trend component.

The "seasonal decompose()" function permits the decomposition of a time series into its constituent components, such as trend, seasonality, and residual. By applying this function to our daily electricity price data, we are able to clearly distinguish between these components.

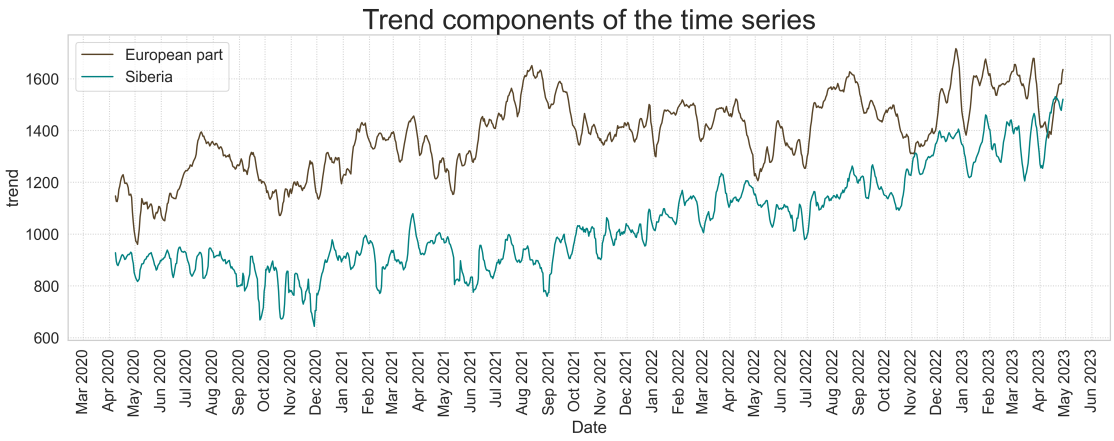

Figure 2.4: Trend components

Utilizing the "trend" method allows us to explicitly isolate the trend component. This technique permits the extraction of the fundamental trend pattern from decomposed time series. By utilizing this instrument, we can concentrate on

analyzing and comprehending the long-term trend of the daily electricity price, which is essential for making informed decisions and forecasts.

2.2.2 Seasonal component

The seasonal component of a time series refers to patterns or fluctuations that repeat over a particular interval, such as daily, weekly, monthly, or annually. In this research, the seasonal component is extracted using the "seasonal" method.

The seasonal component captures recurrent patterns that occur within a predetermined time frame. In the context of daily electricity prices, for instance, we may observe higher prices at specific times or days of the week. By extracting the seasonal component, it is possible to recognize and analyze these repetitive patterns, thereby enabling us to comprehend the seasonal variations in the electricity price data.

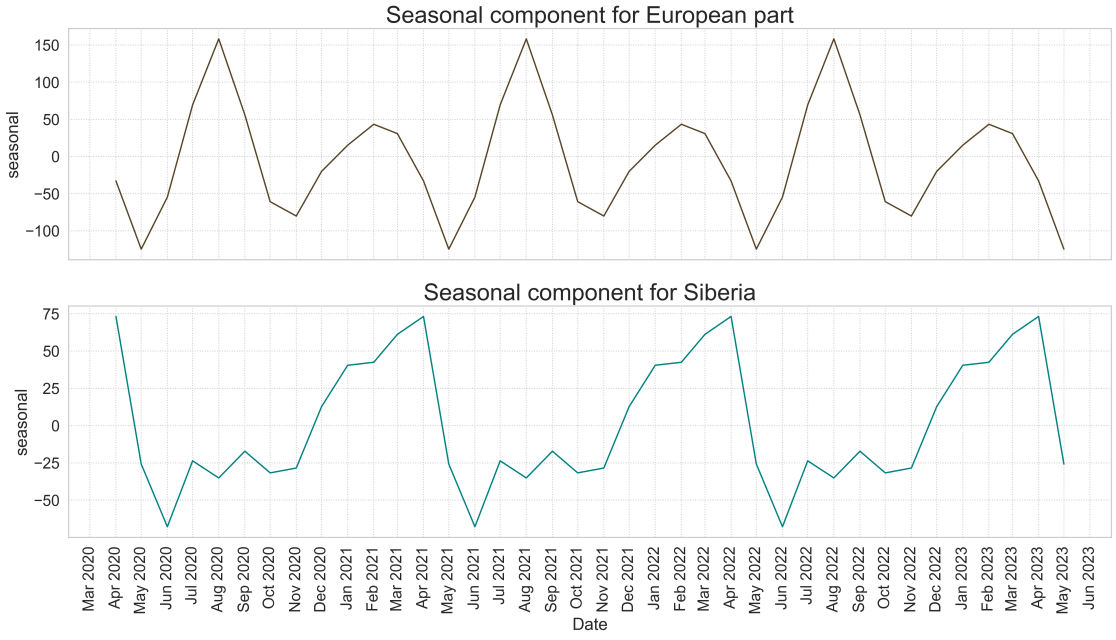

Figure 2.5: Seasonal components

Seasonal component analysis is advantageous for a number of factors. It allows to recognize patterns and trends that occur at regular intervals, allowing us to make informed decisions based on this information. In addition, knowledge of seasonal patterns can aid in anomaly detection by highlighting deviations from expected seasonal behavior.

By utilizing the "seasonal" method, we are able to obtain insight into the seasonal component of the daily electricity pricing data, which is crucial to our analysis. This information will contribute to a thorough comprehension of the data and provide a firm foundation for subsequent analysis, such as anomaly detection and prediction modeling.

2.2.3 Residual component

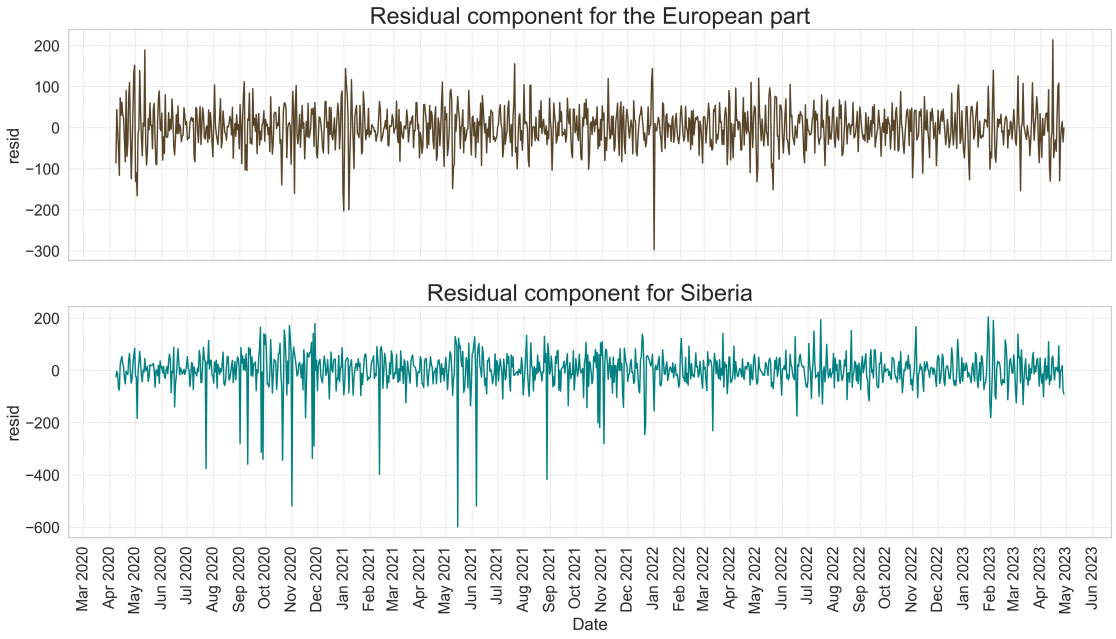

Figure 2.6: Residual components

The residual component of a time series represents the random variation that

remains after the trend and seasonal components have been extracted. Using the "resid" method, we focus on the residual component in our analysis.

The residual component is essential for comprehending the unpredictability or irregularity of data fluctuations. It represents the disturbance or error component that remains after the systematic components have been accounted for. Analyzing the residual component enables to evaluate the quality of our trend and seasonal models and to identify any remaining patterns or anomalies that may require additional investigation.

Analyzing the residual component also aides in anomaly detection by identifying any atypical or unexpected patterns that deviate from the expected trend and seasonality. These anomalies may provide valuable insight into the factors that influence the price of electricity or indicate potential data quality issues.

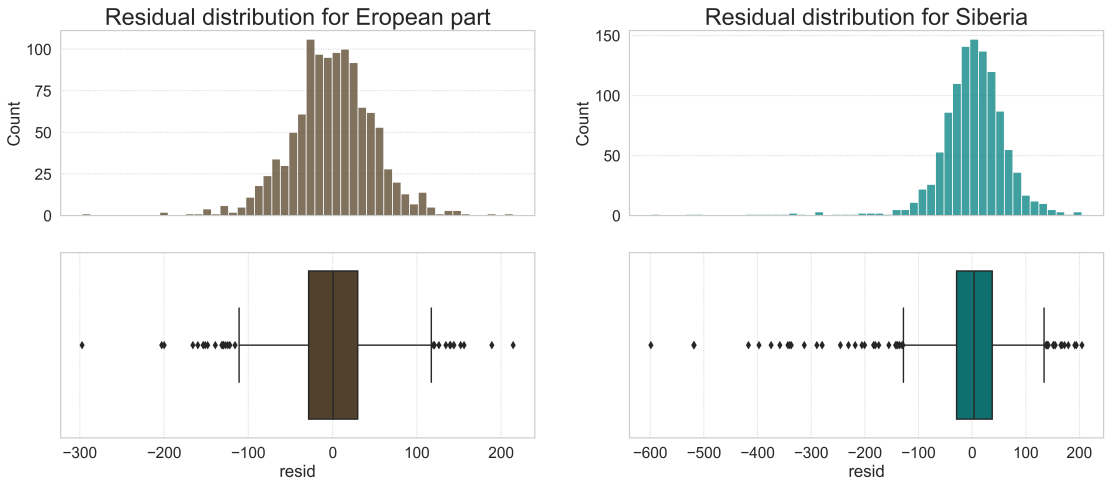

Figure 2.7: Residual distribution

2.3 Stationarity test

The stationarity of a series in time series analysis is a statistical characteristic that evaluates the constancy of its mean and variance over time. A series is considered stationary if both the mean and standard deviation remain constant over time. In contrast, the series is classified as a non-stationary process, such as a random walk with a unit root, if the mean or variance exhibits a trend or demonstrates significant changes over time.

In the context of the Augmented Dickey-Fuller (ADF) test, the presence of a unit root process in a time series is referred to as a unit root. A unit root process is a stochastic process whose series lacks stationarity. It suggests that the series tends to deviate from its mean over time and lacks a consistent mean or trend.

The Augmented Dickey-Fuller (ADF) test is utilized to assess the stationarity of a time series. The ADF test statistic is an essential component of the ADF test for unit roots and stationarity in a time series [4]. Following is the formula for the ADF test statistic:

$$ADF(t) = (Y(t) - Y(t-1)) - \sum_{i=1}^{p}(\alpha(i) \times \Delta Y(t-i)) + \xi(t),$$

where:

- $ADF(t)$ is the ADF test statistic at time t.
- $Y(t)$ is the value of the time series at time t.
- $Y(t-1)$ is the lagged value of the time series at time t-1.
- $\alpha(i)$ represents the estimated coefficients of the autoregressive terms.
- $\Delta Y(t-i)$ is the differenced value of the time series at time t-i.
- $\sum_{i=1}^{p}(\alpha(i) \times \Delta Y(t-i))$ is the sum of the autoregressive terms multiplied by their corresponding differenced values.
- $xi(t)$ is the residual or error term.

By regressing the differenced time series on its lagged values, the ADF test

statistic is calculated. The significance of the evidence against the existence of a unit root in the time series is represented by the test statistic. If the ADF test statistic is substantially smaller (more negative) than the critical values, it suggests that the null hypothesis of a unit root has been rejected, indicating that the time series is stationary.

In order to determine whether a time series is stationary, the calculated test statistic is contrasted with critical values. Critical values define the range of values beyond which we can reject the null hypothesis of a unit root and conclude that the series is stationary. The calculated critical values are provided in statistical tables. They are specific to various sample sizes and levels of statistical significance. For us to reject the null hypothesis of a unit root and conclude stationarity, the ADF test statistic must be smaller (more negative) than the critical values.

Notably, the ADF test statistic is supplemented by p-values that indicate the statistical significance of the test. Under the assumption that the null hypothesis is true, the p-value represents the probability of observing a test statistic as extreme as or more extreme than the calculated value. Lower p-values indicate that the evidence against the null hypothesis is stronger.

Using the ADF test to evaluate stationarity is advantageous because it provides a formal statistical framework for evaluating the stationarity of a time series, allowing us to draw trustworthy conclusions based on empirical evidence.

In the given research ADF test was performed using the "adfuller" function from the "Statsmodels" library [3]. The results of the ADF test for the residual component are presented in the following table:

Residual Statistics		
	European part	Siberia
ADF stattistic	-13.624	-13.219
p-value	0.0	0.0
Critical values		
1%	-3.436	-3.436
5%	-2.864	-2.864
10%	-2.568	2.568

Table 2.1: Table of residual statistics.

The ADF Statistic gauges the evidence against the unit root in the residual component. In both instances, the ADF Statistic is extremely negative, indicating substantial evidence against the unit root null hypothesis. This indicates that the component is stationary.

Also reported are the p-values associated with the ADF Statistic. In both instances, the p-value is 0.000, which is considerably lower than the commonly employed significance level of 0.05. This provides additional evidence against the presence of a unit root and bolsters the conclusion that the residual component is stationary.

In addition, Critical Values are the threshold values against which the ADF Statistic is compared to determine statistical significance. In both instances, the ADF Statistic is well below the critical values at all levels of significance (1%, 5%, and 10%), confirming the stationarity of the residual component and reinforcing the rejection of the null hypothesis.

Based on these results, we can conclude that the residual component exhibits stationarity, signifying that it lacks a unit root and has a constant mean and standard deviation over time.

Chapter 3

Anomalies detection

3.1 Introduction

This chapter focuses on the critical task of detecting anomalies in time series analysis. Methods for detecting anomalies are crucial for identifying and comprehending data points or patterns that significantly deviate from the expected behavior. We investigate the statistical method based on the three sigma rule, the Isolation Forest algorithm, and the k-means clustering algorithm for detecting anomalies.

The statistical method based on the three sigma rule is an uncomplicated technique that implies the residual component of the time series follows a Gaussian distribution. Positive values that exceed a threshold of three standard deviations (sigma) are categorized as anomalies. Likewise, negative values below the negative threshold are regarded as anomalies. This technique offers simplicity and interpretability, but it may not be able to detect all forms of anomalies. It may result in a limited number of detected anomalies.

We introduce the Isolation Forest algorithm in order to surmount the limitations of the statistical method. Instead of modeling the typical data points, this

algorithm isolates anomalies. Utilizing the concepts of randomness and isolation, it detects anomalies in time series with efficiency. The Isolation Forest algorithm constructs isolation trees by recursively partitioning the data, and anomalies are identified as data points for which fewer partitions are required to isolate them. We utilize the implementation provided by the scikit-learn library for our analysis.

We also investigate the use of the k-means clustering algorithm for anomaly detection, in addition to the Isolation Forest algorithm. This algorithm divides the data into a predetermined number of clusters and assigns each data point to the cluster containing the nearest mean. Anomalies are designated as data points that do not conform to any cluster or that fell into clusters with sparse populations. The k-means clustering algorithm offers an alternate method for detecting anomalies based on dissimilarity to the predominant clusters.

We compare the efficacy and characteristics of these three anomaly detection methods throughout this chapter. We assess the number of detected anomalies, their distribution, and their cumulative impact on the analysis. Notably, while the statistical method and the Isolation Forest algorithm may produce a comparable number of anomalies, the k-means clustering algorithm detects anomalies in different locations. Numerous anomalies detected by the k-means clustering algorithm are close to zero, indicating a distinct pattern.

By analyzing and contrasting these various anomaly detection methods, we hope to obtain a better understanding of their merits, limitations, and applicability within the context of time series analysis. Understanding the characteristics of each method will enable us to select the most appropriate approach for detecting anomalies in various scenarios based on informed judgment. This chapter ultimately contributes to the improvement of our ability to recognize and interpret anomalies, resulting in enhanced forecasting and informed decision-making.

3.2 Statistical approach for anomalies detection

Utilizing the residual component, which is considered to have a distribution close to a Gaussian distribution, is the basis for a statistical technique commonly used to detect anomalies. Any residual component values that transcend a predetermined threshold are considered anomalous. In this instance, the threshold is set to three standard deviations, denoted by the three sigma [5].

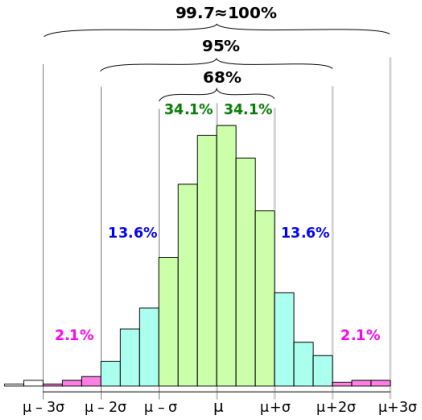

Figure 3.1: Sigma rule

To identify anomalies, we focus on positive values exceeding three sigma and negative values below minus three sigma. Using the three sigma rule, values that deviate substantially from the mean are categorized as anomalies. This threshold is chosen based on the assumption that the probability of obtaining values within this distribution beyond three standard deviations is comparatively low, approximately 0.3%.

Using this statistical technique, we can identify data points in the residual component that exhibit atypical behavior relative to the expected distribution. These anomalies may be indicative of possible outliers, abnormalities, or unanticipated patterns in the time series data. Detecting such anomalies is vital because they can provide valuable insight into the underlying factors or events that may have influenced the observed values.

It is essential to note that while the three sigma rule is a widely employed method for detecting anomalies, it implies that the residual component follows a Gaussian distribution. If the distribution substantially deviates from a Gaussian distribution, it is essential to verify this assumption and investigate alternative approaches. In addition, the selection of the threshold and the interpretation of

anomalies must be modified based on the particular characteristics and requirements of the dataset and the analysis.

Having completed all calculations described above let's superpose all anomalies detected on the chart of residual component.

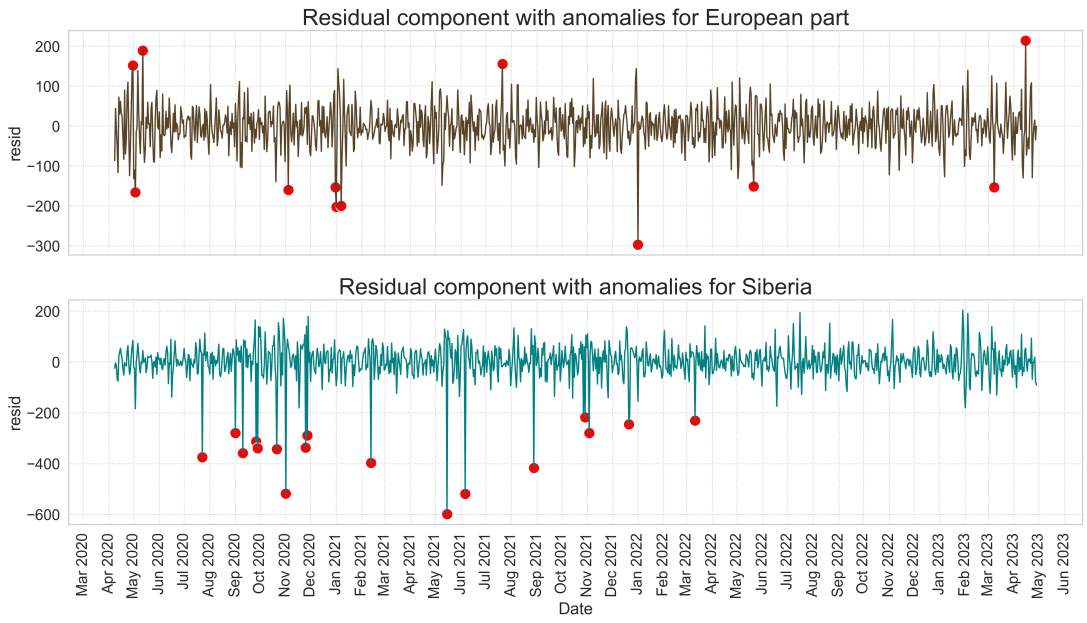

Figure 3.2: Combination of residual component and anomalies detected by statistical method

Using the statistical procedure based on the three sigma rule for anomaly detection, we found that only a small number of anomalies were detected in the time series data. Several factors can account for this small number of detected anomalies.

First, the three sigma rule is a conservative method for detecting outliers based on the assumption of a Gaussian distribution. Considered anomalous are data values that deviate more than three standard deviations from the mean. However, this method may be less sensitive to subtle deviations or outliers that fall below the three-sigma threshold. As a result, it may overlook anomalies that fall between

two and three standard deviations.

The three sigma rule also implies that the data follow a normal distribution. This method's efficacy may be compromised if the underlying data deviate significantly from a normal distribution. Non-Gaussian distributions with hefty tails or skewness may contain anomalies that fall within the range of three standard deviations, but nevertheless display anomalous behavior.

In addition, the nature of the time series itself can impact the number of anomalies detected by the three sigma rule. The number of detected anomalies may be low if the time series exhibits relatively stable and predictable patterns with few fluctuations or extreme values. In contrast, the three sigma rule may identify more anomalies if the time series is highly volatile or subject to frequent abrupt changes.

The small number of anomalies detected by the statistical method based on the three sigma rule can be attributed to its conservative nature, the assumption of normality, and the particular characteristics of the analyzed time series. While this method provides a starting point for identifying extreme values, it may not capture all types of anomalies, especially those that deviate moderately from the mean or do not follow a normal distribution. Therefore, alternative techniques, such as the Isolation Forest algorithm, may be required to capture a broader range of anomalies and provide a more thorough analysis of the time series data.

3.3 Classification methods for anomalies detection

For the detection of anomalies in time series, the Isolation Forest classification method has been selected. This method was implemented utilizing the well-known machine learning library scikit-learn [6].

Isolation Forest is a potent algorithm specifically designed for anomaly detection. It is based on the idea that anomalies are uncommon instances that can be isolated more readily than typical instances within a dataset. To partition the data recursively, the algorithm constructs a collection of binary trees, also known as isolation trees [7]. To separate instances, the procedure involves randomly identifying a feature and a dividing point within the range of that feature. The procedure is repeated until all instances are isolated or a maximal tree depth is reached.

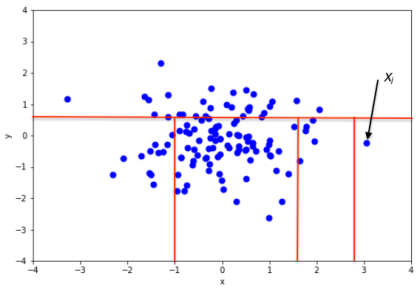

Figure 3.3: Isolating an anomalous point

Isolation Forest is based on the principle that anomalous instances require fewer partitions to be isolated than normal instances. Consequently, the algorithm allocates higher scores (anomaly scores) during training to instances that require fewer divisions for isolation [8]. These scores can be used to rank instances and identify those with the highest scores as possible anomalies.

The selection of Isolation Forest as the classification method for anomaly detection is supported by its capacity to deal with high-dimensional datasets, its efficacy in dealing with large datasets, and its resistance to outliers. In addition, Isolation Forest requires no assumptions regarding the data distribution and is insensitive to the presence of irrelevant features.

Using the algorithm Isolation Forest from the scikit-learn library, we can effectively identify anomalies in the time series data. The ability of the algorithm to isolate and classify anomalies according to their anomaly scores enables accurate and efficient detection of peculiar patterns or outliers in the dataset.

After completing all the calculations described above, we can now superpose all the detected anomalies on the residual component's chart. This visualiza-

tion enables us to observe the locations and patterns of the identified time series anomalies.

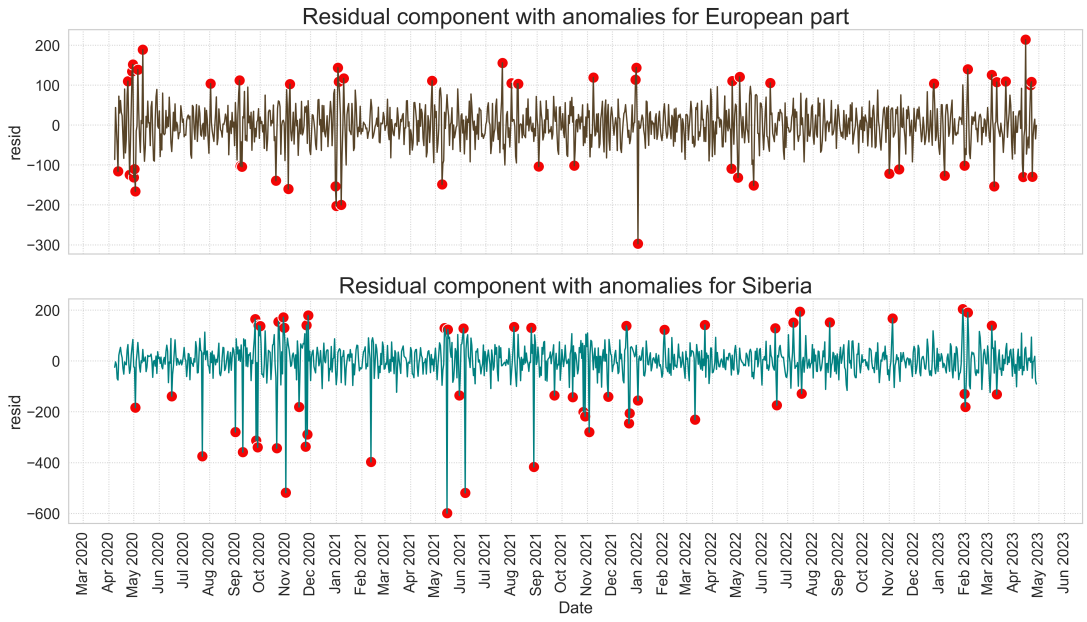

Figure 3.4: Combination of residual component and anomalies detected by classification method

When using the Isolation Forest algorithm for anomaly detection, we observed a greater number of anomalies than when using the statistical method. This indicates that the Isolation Forest algorithm is more sensitive and efficient at detecting anomalous instances or patterns in time series data.

Visualizing the detected anomalies on the plot of the residual component provides a clear representation of their positions relative to the overall data distribution. This visual inspection can assist in validating the efficacy of the selected anomaly detection method and shedding light on the nature of the anomalies.

By contrasting the results of the statistical method and the Isolation Forest algorithm, we can acquire a better comprehension of the advantages and disadvantages of each method. While the statistical method may be more cautious in

detecting anomalies, the Isolation Forest algorithm is typically more sensitive and capable of identifying subtle deviations from the norm.

Ultimately, the combined analysis of the detected anomalies and their visualization on the residual component chart provides valuable insights into the presence of aberrant behaviors or events in the time series data. These insights can support decision-making processes in various domains and contribute to a better comprehension of the fundamental factors that influence anomalies.

3.4 Clustering methods for anomalies detection

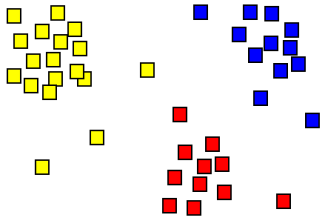

Figure 3.5: Cluster analysis

The k-means clustering algorithm [9] is an additional effective technique for detecting anomalies in time series analysis. In this method, we add a new feature to our analysis, namely the residual component altered by 30 days. By incorporating this shifted residual, we hope to identify any potential long-term patterns or trends in the data.

To apply the k-means clustering algorithm, we must first determine the optimal cluster size. We employ the elbow method [10], a widely used technique for determining the optimal number of data clusters. By evaluating the sum of squared distances between data points and their cluster centroids for various numbers of clusters, we determine the "elbow point" at which the rate of improvement declines considerably. This point represents the optimal cluster count for our analysis.

According to the calculation and information on the diagram, the optimal number of clusters for European data is five, and for Siberian data, it is seven, as

the variance score does not change significantly after these values.

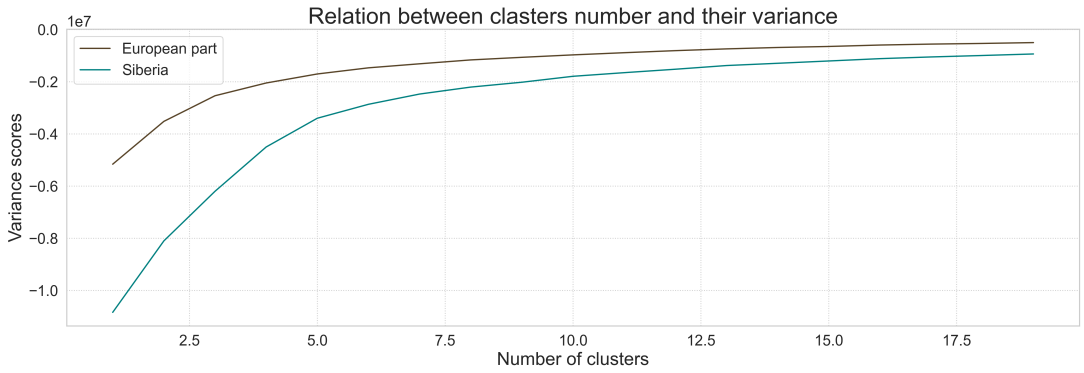

Figure 3.6: Dependence of variance on the number of clusters

After determining the optimal number of clusters, the clustering procedure is initiated. The k-means algorithm divides the data into the specified number of clusters by designating each data point iteratively to the cluster containing the nearest mean centroid. The centroids are modified based on the allotted data points, and the process is repeated until convergence is reached.

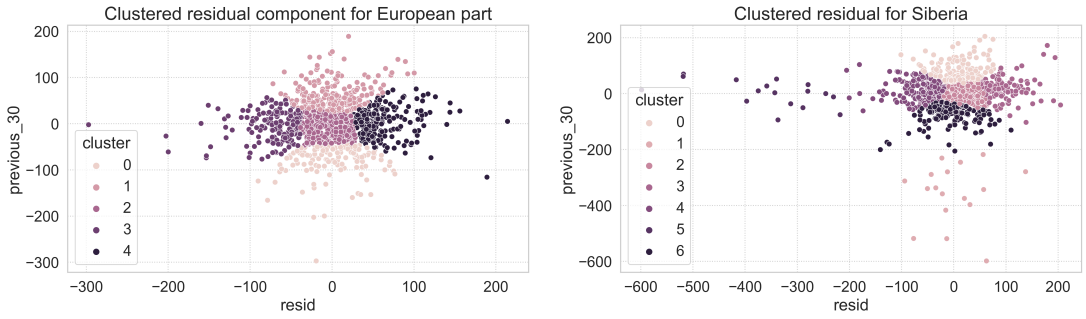

Figure 3.7: Clustered residual component

We define a threshold based on the distance between data points and their respective cluster centers in order to identify anomalies within each cluster. In particular, we identify as anomalous distances that exceed the 0.95 percentile of all distances within each cluster. This method enables us to concentrate on data

points that deviate significantly from the cluster's center and are regarded as potential anomalies.

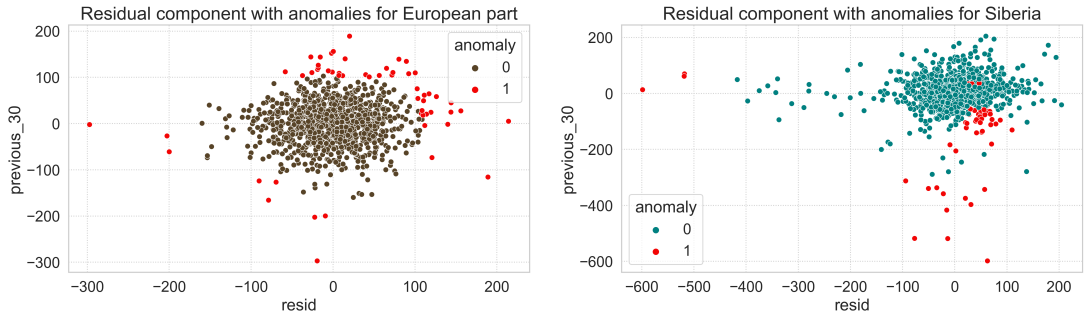

Figure 3.8: Combination of residual component and anomalies detected by classification method

By employing the k-means clustering method with the shifted residual feature, we intend to identify any distinct patterns or behaviors in the data that may be indicative of outliers. The combination of clustering and anomaly detection provides a thorough method for identifying anomalies based on dissimilarity to the prevalent clusters.

Utilizing the concepts of clustering and dissimilarity, the k-means clustering method provides a potent tool for detecting anomalies. By incorporating the shifted residual and using the elbow method to determine the optimal number of clusters, we hope to improve the precision and dependability of anomaly detection in our analysis.

Using the k-means clustering algorithm for anomaly detection, we discovered that the number of detected anomalies did not differ substantially from the number detected by the previous method. Nonetheless, there was a discernible shift in the location of the anomalies, with a significant number of them appearing close to zero.

The shift in the distribution of anomalies toward zero can be attributed to

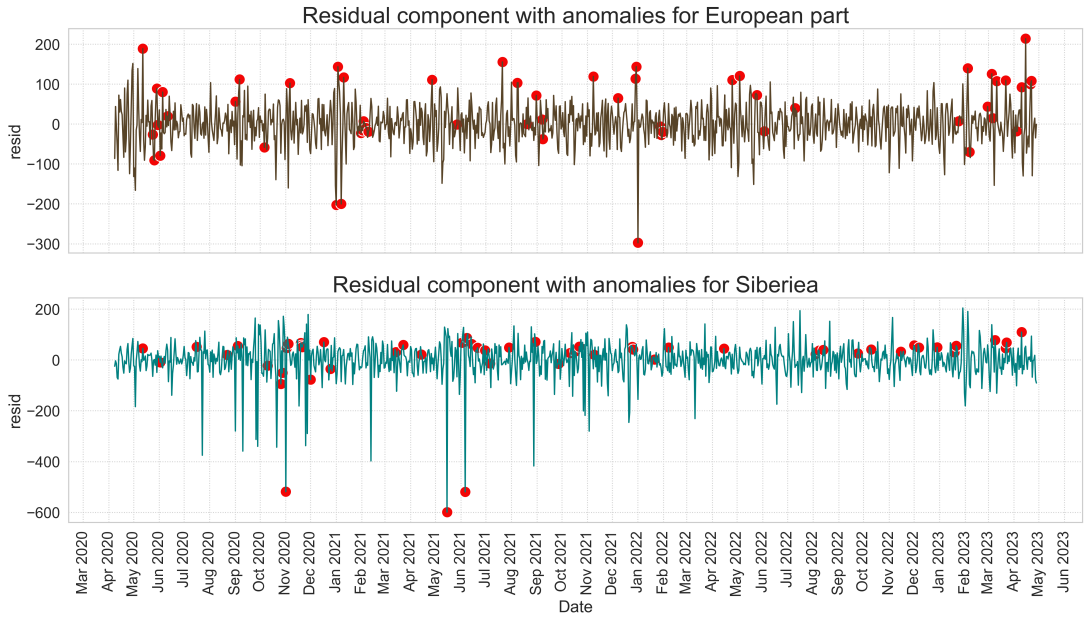

Figure 3.9: Combination of residual component and anomalies detected by clustering method

the algorithm's inherent characteristics. This algorithm divides the data into a specified number of clusters, assigning each data point to the cluster with the closest mean. In the context of anomaly detection, data points that do not conform to any of the clusters or fall into clusters with sparse populations are regarded as potential anomalies.

The preponderance of anomalies closer to zero may indicate that the behavior of these data points deviates considerably from that of the majority of the data. These anomalies may represent instances of exceedingly low or near-zero values that deviate from the expected patterns observed in the time series. These data points are identified as anomalies by the clustering algorithm due to their dissimilarity to the dominant clusters.

It is important to note that the k-means clustering algorithm operates under the assumption that clusters have similar mean values. In situations where anoma-

lies exhibit behavior that is distinct from the preponderance of data, the clustering algorithm may not effectively capture them. This can result in a relatively constant number of anomalies being detected across a time series, but with varying distribution and location based on their dissimilarity to the dominant clusters.

Chapter 4

Time series forecast

4.1 Introduction

In this chapter, we examine the effect of anomalies on the accuracy of electricity pricing forecasts using the Triple Exponential Smoothing model. The prediction dataset is comprised of daily electricity prices collected from two distinct regions: the European part of Russia and Siberia. We hope to obtain insight into the behavior and dynamics of electricity prices in these regions by analyzing these datasets.

We employ the Triple Exponential Smoothing model [11] from the statsmodels Python library [3] for the prediction model. This model is an extensively employed time series forecasting method that depicts the data's trends, seasonality, and level. This model was chosen because it is appropriate for capturing the underlying patterns and dynamics in the electricity price time series, taking seasonality and trend into account.

To investigate the effect of anomalies on the accuracy of predictions, we train the models using two distinct data sets: one with anomalies included and one without. By comparing the prediction errors of these two models, we can determine

the effect of anomalies on the accuracy of predictions. This analysis provides insightful information regarding the robustness and dependability of the prediction model in the presence of anomalies.

As a metric for evaluating the accuracy of the predictions, we employ the mean squared error (MSE) [12] from scikit-learn Python library [6]. The MSE is a commonly employed metric for evaluating the accuracy of predictions. This metric was chosen because it provides a comprehensive evaluation of the prediction errors, considering both the magnitude and direction of the deviations into consideration.

By employing the Triple Exponential Smoothing model and evaluating the prediction errors using the MSE metric, we hope to provide a thorough analysis of the effect of anomalies on the accuracy of electricity price forecasts. This analysis enables us to draw conclusions about the performance of the prediction model in the presence of outliers and to gain insight into the overall accuracy of the forecasting procedure.

4.2 Results of prediction

In this section, we present the results of the Triple Exponential Smoothing model's electricity price prediction and analyze the influence of anomalies on the accuracy of the predictions. Comparing the efficacy of prediction models trained on datasets with and without anomalies allows us to evaluate the impact of various anomaly detection techniques.

Mean squared error (MSE) for the dataset acquired from the European part of Russia was 3633 for the model trained on the dataset with anomalies. In contrast, the model produced an MSE of 4,398 when the Statistical method for anomaly detection was applied to the dataset devoid of anomalies. The MSE for the Isolation Forest method was 2216, while the MSE for the K-means algorithm was 21610.

Similarly, the model trained on the dataset with anomalies produced an MSE of 4982 for the dataset acquired in Siberia. In contrast, the model trained on the dataset devoid of anomalies using the Statistical method for anomaly detection demonstrated an MSE of 4937. The MSE for the Isolation Forest method was 4881, while the MSE for the K-means algorithm was 4550.

Chapter 5

Conclusions and future work

5.1 Limitations

This section examines the limitations of the thesis, concentrating on three essential aspects: Anomaly Detection Methods, Assumptions and Simplifications, and Evaluation Metrics.

Anomaly Detection Methods: This thesis investigates three distinct anomaly detection methods: the Statistical method, the Isolation Forest, and the K-means algorithm. Even though these methods have been shown to be effective at detecting anomalies in the electricity price dataset, it is essential to recognize that alternative or more sophisticated methods may exist. The limitations of the selected methodologies and their applicability to various datasets should be closely considered.

The thesis makes certain assertions about the distribution of anomalies and employs fixed thresholds for anomaly detection. It is crucial to recognize that these assumptions may not hold true in all circumstances, thereby limiting the generalizability of the findings. In addition, the analysis considers anomalies as separate data points without considering their potential influence on neighboring

points.

Metrics for Evaluation: The evaluation of prediction performance in this thesis is based solely on the mean squared error (MSE). MSE is a measure of prediction accuracy; however, it may not capture other essential aspects, such as the bias-variance trade-off or the ability to capture extreme events. Incorporating additional evaluation metrics, such as mean absolute error or quantile loss, could provide a more thorough evaluation of the performance of prediction models.

Recognizing these limitations in the selection of anomaly detection methods, assumptions and simplifications, and evaluation metrics enables a more nuanced comprehension of the thesis's findings. In addition, it provides valuable directions for future research to surmount these limitations and improve the accuracy and applicability of anomaly detection techniques in the field of electricity pricing forecasting.

5.2 Conclusions

It is evident from the results of the prediction analysis that the presence of anomalies in the datasets has a substantial impact on the performance of the prediction models.

For data acquired from the European part of Russia, the model obtained an MSE of 3633 when trained on the dataset containing anomalies. However, when there were no anomalies in the dataset, the efficacy of the model varied contingent on the anomaly detection method employed. The MSE for the model trained on the dataset without anomalies using the Statistical method was 4398, while the MSE for the model trained using the Isolation Forest method was 2216. Surprisingly, the MSE for the K-means algorithm was 21610, indicating that it may not be suitable for anomaly detection in this context.

In the case of Siberian data, the MSE of the model trained on the dataset containing anomalies was 4982. In contrast, when trained on a dataset devoid of anomalies, the performance of all anomaly detection methods improved. The MSE for the Statistical method was 4937, the MSE for the Isolation Forest method was 4881, and the MSE for the K-means algorithm was 4550. These results suggest that removing anomalies from the dataset improves the accuracy of electricity price predictions in Siberia.

	Statistical method	Isolation Forest	K-means
European part with anomalies	3633	3633	3633
European part without anomalies	4398	2216	21610
Siberia with anomalies	4982	4982	4982
Siberia without anomalies	4937	4881	4550

Table 5.1: Table of MSE in relation to anomaly detection method.

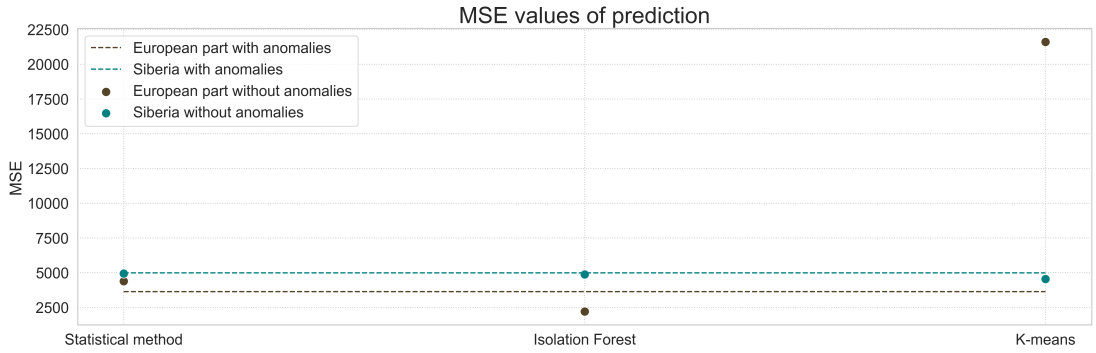

Figure 5.1: MSE values in relation to anomaly detection method

The significance of effectively detecting and managing anomalies in the prediction process is highlighted by these results. The presence of anomalies can introduce substantial disturbance and disrupt data patterns, resulting in less precise predictions. Improving prediction performance requires the implementation of

effective anomaly detection techniques, such as the Isolation Forest and Statistical methods. In addition, the selection of an appropriate evaluation metric, such as MSE, provides a quantitative evaluation of the accuracy of the models.

This analysis concludes by emphasizing the influence of anomalies on prediction outcomes and the necessity of incorporating reliable anomaly detection techniques into electricity price forecasting. By addressing anomalies effectively, researchers and practitioners can improve the accuracy and reliability of prediction models, leading to improved energy sector decision-making.

5.3 Future work

In the research, we have investigated various methods for anomaly detection and their influence on electricity price forecasting. However, there are still several avenues for future research that can improve our understanding and the precision of our forecasts.

First, broadening the scope of anomaly detection techniques beyond those investigated in this study could yield valuable insights. Exploring machine learning algorithms such as Support Vector Machines, Random Forests, and Neural Networks could provide alternative approaches to effectively detect anomalies. It is possible to conduct comparative studies to evaluate the performance of various anomaly detection methods in various scenarios.

Investigating the connection between the detected anomalies and external factors or occurrences could provide a deeper understanding of their causes. By incorporating additional data sources, such as weather patterns, holidays, and economic indicators, we are able to analyze the contextual factors that contribute to electricity prices' atypical behavior. This analysis can aid in identifying the underlying causes of anomalies and, by incorporating these external variables, enhance the

accuracy of prediction models.

Additionally, investigating ensemble methods that incorporate multiple anomaly detection algorithms could potentially improve the process's robustness and dependability. By exploiting the assets of various methods and combining their outputs, we can construct anomaly detection systems that are more capable of handling diverse datasets and intricate patterns.

Expanding the scope of prediction models to include multivariate analysis can provide a more comprehensive understanding of the dynamics of electricity price. By incorporating additional relevant characteristics, such as electricity demand, supply, and market conditions, we can create more sophisticated prediction models that account for the interdependencies and interactions between various factors.

The study concludes by laying the foundation for anomaly detection in electricity pricing forecasting. However, additional research is required to investigate advanced anomaly detection methods, investigate the relationship between anomalies and external factors, analyze the temporal patterns of anomalies, employ ensemble methods, and expand prediction models to incorporate multivariate analysis. These efforts will result in more precise and reliable forecasts, allowing for improved energy sector decision-making.

Bibliography

[1] Varun Chandola and Arindam Banerjee and Vipin Kumar. " Anomaly detection: A survey ". In: *ACM Computing Surveys* 41 (2009), pp. 15–.

[2] Hyndman, R.J. and Athanasopoulos, G. *Forecasting: principles and practice*. Vol. 382. OTexts, 2018, pp. 157–161.

[3] Seabold, Skipper and Perktold, Josef. " statsmodels: Econometric and statistical modeling with python ". In: *9th Python in Science Conference*. 2010.

[4] " Time Series Analysis Using SAS Part I The Augmented Dickey-Fuller (ADF) Test ". In: 6 (2008).

[5] Wikipedia contributors. *68–95–99.7 rule — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=68%E2%80%9395%E2%80%9399.7_rule&oldid=1153583638.

[6] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. " Scikit-learn: Machine Learning in Python ". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[7] " Extended Isolation Forest ". In: *IEEE Transactions on Knowledge and Data Engineering* 33 (2021), pp. 1479–1489.

[8] Wikipedia contributors. *Isolation forest — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=Isolation_forest&oldid=1154885602.

[9] Wikipedia contributors. *Cluster analysis — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=Cluster_analysis&oldid=1152294323.

[10] " Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster ". In: 336 (2018).

[11] Wikipedia contributors. *Exponential smoothing — Wikipedia, The Free Encyclopedia*. 2023. URL: https://en.wikipedia.org/w/index.php?title=Exponential_smoothing&oldid=1151784266.

[12] Wikipedia contributors. *Mean squared error — Wikipedia, The Free Encyclopedia*. 2022. URL: https://en.wikipedia.org/w/index.php?title=Mean_squared_error&oldid=1127519968.