Marquette University
e-Publications@Marquette

Dissertations (2009 -) | Dissertations, Theses, and Professional Projects

Data Cleaning in the Energy Domain

Hermine Nathalie Akouemo Kengmo Kenfack
Marquette University

Recommended Citation
Akouemo Kengmo Kenfack, Hermine Nathalie, "Data Cleaning in the Energy Domain" (2015). *Dissertations (2009 -)*. Paper 515.
http://epublications.marquette.edu/dissertations_mu/515

DATA CLEANING IN THE ENERGY DOMAIN

by

Hermine N. Akouemo Kengmo Kenfack, B.S, M.S.

A Dissertation Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for
the Doctor of Philosophy

Milwaukee, Wisconsin

May 2015

ABSTRACT
DATA CLEANING IN THE ENERGY DOMAIN

Hermine N. Akouemo Kengmo Kenfack, B.S, M.S.

Marquette University, 2015

This dissertation addresses the problem of data cleaning in the energy domain, especially for natural gas and electric time series. The detection and imputation of anomalies improves the performance of forecasting models necessary to lower purchasing and storage costs for utilities and plan for peak energy loads or distribution shortages.

There are various types of anomalies, each induced by diverse causes and sources depending on the field of study. The definition of false positives also depends on the context. The analysis is focused on energy data because of the availability of data and information to make a theoretical and practical contribution to the field. A probabilistic approach based on hypothesis testing is developed to decide if a data point is anomalous based on the level of significance. Furthermore, the probabilistic approach is combined with statistical regression models to handle time series data. Domain knowledge of energy data and the survey of causes and sources of anomalies in energy are incorporated into the data cleaning algorithm to improve the accuracy of the results.

The data cleaning method is evaluated on simulated data sets in which anomalies were artificially inserted and on natural gas and electric data sets. In the simulation study, the performance of the method is evaluated for both detection and imputation on all identified causes of anomalies in energy data. The testing on utilities' data evaluates the percentage of improvement brought to forecasting accuracy by data cleaning. A cross-validation study of the results is also performed to demonstrate the performance of the data cleaning algorithm on smaller data sets and to calculate an interval of confidence for the results.

The data cleaning algorithm is able to successfully identify energy time series anomalies. The replacement of those anomalies provides improvement to forecasting models accuracy. The process is automatic, which is important because many data cleaning processes require human input and become impractical for very large data sets. The techniques are also applicable to other fields such as econometrics and finance, but the exogenous factors of the time series data need to be well defined.

ACKNOWLEDGMENTS

Hermine N. Akouemo Kengmo Kenfack, B.S, M.S.

This research work would not have been possible without the financial support of the GasDay Laboratory at Marquette University and the assistance of my academic advisor Dr. Richard Povinelli.

I would like to express my gratitude to Drs. Richard Povinelli, George Corliss, and Ronald Brown for all their teaching and countless hours of help and guidance provided throughout the completion of this research work. The knowledge shared with me are not only academic but also life-long lessons for which I am grateful. Also, I must thank my committee members Drs. Monica Adya and James Richie for the ideas, insights, and contributions brought to improve the quality of this research work.

I would like to thank the Marquette University GasDay Laboratory, especially Thomas Quinn, for the financial support, but also for the great learning environment provided. Thank you to the GasDay laboratory graduate students for all their remarks, idea sharing, and constructive criticisms, but most importantly, for been an admirable and supportive community of peers.

This work is dedicated to my parents Jean and Helene Akouemo, my family Andre and Yohan Nguimfack, and all my brothers and sisters. Without their sacrifice, love, and support throughout this journey, this work would not have been possible. There are so many people I am thankful for, and while I cannot thank each one individually, I am blessed to have had so much support along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION TO DATA CLEANING

This dissertation addresses the problem of data cleaning in the energy domain, specifically the electric and natural gas industries. A detailed discussion of the problem is presented below. The contributions to be drawn from this research and an overview of the natural gas and electric industries also are presented in this chapter. This chapter concludes with an outline of the remainder of the dissertation.

1.1 Data Cleaning Problem Statement

Data cleaning is the process that consists of detecting, diagnosing, and imputing anomalous data [96]. Specifically, this dissertation focuses on cleaning time series data from the energy field. This data is used as input to energy demand forecasting models. Accurate forecasting is important because it helps the energy industry and their customers save energy and money. During severe weather such as extreme cold or heat waves, accurate forecasts can save lives by ensuring that the necessary energy is available for heating and cooling, respectively, and to support essential services such as hospitals. One of the important prerequisites for accurate forecasting is clean data. The objective of this work is to clean the data for the purpose of being used to train a forecasting model. Training a model on a time series containing anomalous data typically results in erroneous and biased

parameters [25, 27, 94]. There are numerous causes for anomalous data. Manually examining time series for all causes of anomalies is a tedious task and for large data sets an infeasible one; thus the need for an automated, repeatable, and accurate algorithm for data cleaning.

An essential component of data cleaning is domain knowledge. Domain knowledge consists of a predefined set of templates representing patterns in the data or a description of events that constitute anomalous behavior [2, 10]. For example, the delivered energy during a power outage may be accurate but still anomalous, because it does not depict the true demand as if the power outage had not occurred. For accurate results, the inputs of the forecasting models need to represent the historical demanded energy, as opposed to the supplied energy. A power outage is one example of an event that will be incorporated into the proposed data cleaning process. To make the data cleaning problem tractable, this dissertation will focus on the electric and natural gas domains, specifically detecting and imputing anomalies to predict accurately residential, commercial, and industrial energy demand.

The next sections of this chapter present the contribution of this work, provide an overview of the energy industry, and describe causes of anomalous data in the energy field.

1.2 Contributions

This thesis makes two major contributions. The first is an anomaly detection algorithm that consists of finding anomalies in energy time series and classifying the anomalies by looking at possible causes. The second contribution is a data imputation algorithm that estimates a replacement value for the anomalous data. Thus, the novel contributions of this thesis address the problem of cleaning time series energy data with minimal human input. The proposed algorithms are computationally tractable, repeatable, accurate, and automatic.

Anomaly detection and data imputation have been studied across various disciplines such as nursing, engineering, and economics. Chapter 2 of this dissertation presents existing outlier and anomaly detection techniques found in the literature. These techniques typically perform well on simulated data sets. However, on real data sets, these techniques do not yield accurate model parameters and estimates. A new probabilistic approach for anomaly detection is developed based on hypothesis testing that takes into account the number of samples in a data set and efficiently identifies anomalies. Given an energy time series, the data set is analyzed to extract domain knowledge features. Then the anomalies are found using the time series residuals. The data replacement is performed using multiple regression imputation. The contributions of this dissertation are both theoretical and practical.

The next section of this chapter presents an overview of the energy industry.

1.3 Energy Industry Overview

This section presents both the natural gas and electric industries. The categories of customers in each industry also are presented here.

1.3.1 Natural Gas Industry

Natural gas is a fossil fuel and nonrenewable source of energy that is extracted from the ground. Figure 1.1 illustrates the natural gas industry from production to end use.

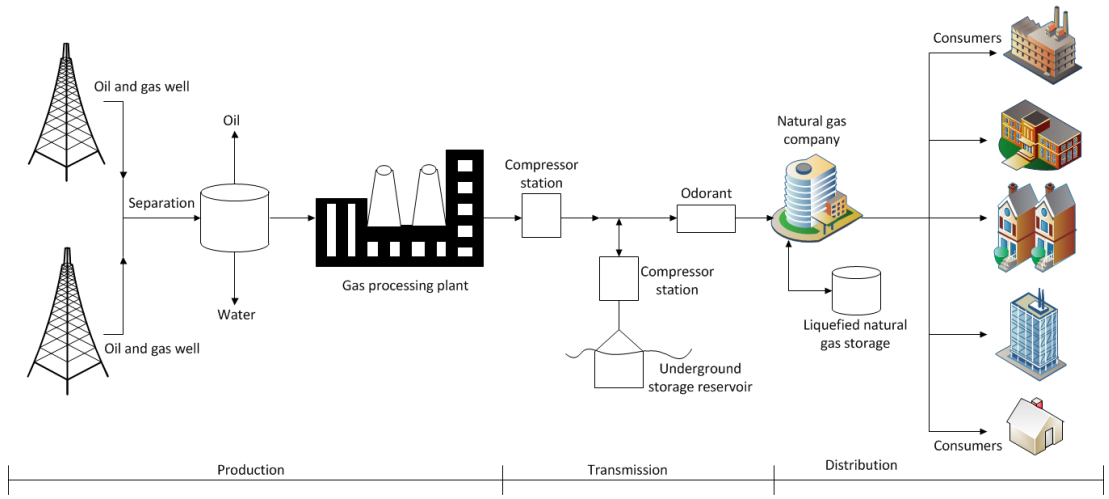

Figure 1.1: The natural gas production, transmission, and distribution system (adapted from [34])

In the production component of the system, the oil and gas extracted from wells are separated into oil, water, and natural gas. The natural gas is processed at

the plant to remove undesired hydrocarbons and other non-hydrocarbon gases. After its processing at the plant, the natural gas is transported to local distribution companies through pipelines. Natural gas local distribution companies (LDCs) are responsible for the distribution component of the system by supplying gas to their customers and storing gas for peak demand times. The four types of customers encountered in the natural gas industry and their uses of natural gas are [34, 73] :

- **Residential customers** use natural gas for cooking, space heating, space cooling, and running appliances such as clothes dryers and water heaters.
- **Commercial customers** use natural gas for space heating and cooling and water heating. Commercial customers are retail and service shops, administrative buildings, banks, schools, universities, health care buildings, and hotels.

  Consumption of natural gas by residential and commercial customers is weather dependent.
- **Industrial customers** use natural gas to run their manufacturing processes such as hydrogen and petroleum refining, and production of pulp, paper, metals, stone, clay, glass, and plastic.
- **Electric power plants** use natural gas to generate electricity.

The electric industry is presented in the next section.

1.3.2 Electric Industry

Electricity is a secondary energy source that is produced from the conversion of other energy sources such as coal, solar, and natural gas. Figure 1.2 illustrates the electric industry from the generation at the power plant to the delivery at the customers.

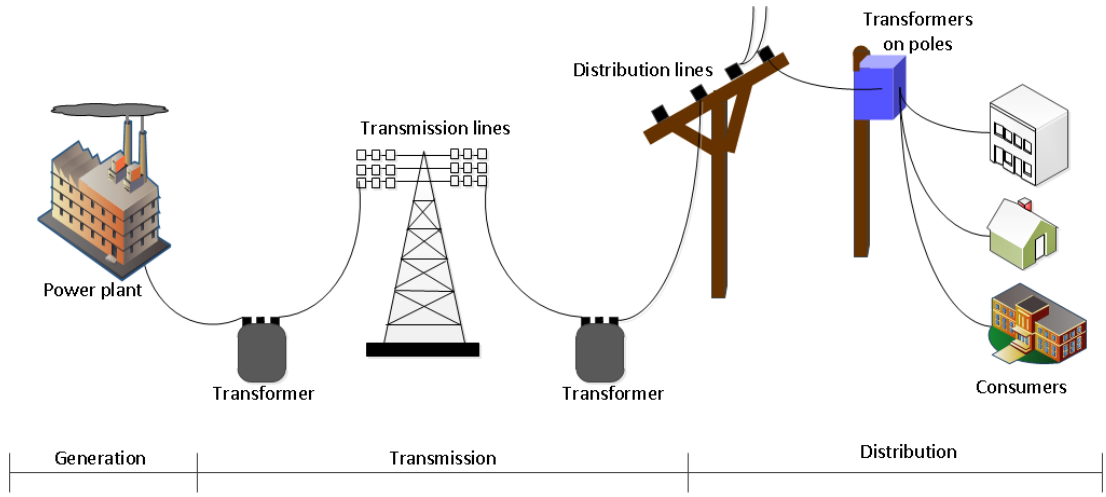

Figure 1.2: Electric power generation, transmission, and distribution system (adapted from [34])

Electric power is generated at power plants and transported to substations via large, high-voltage transmission power lines. Transformers are used to step up voltages for transmission and step down voltages for distribution. A local distribution system of smaller, lower-voltage distribution lines moves power from substations and transformers to customers. There are four types of customers in the electric industry [34].

- **Residential customers** use electricity for lighting, heating, cooling, cooking, and powering appliances and equipment.
- **Commercial customers** use electricity for lighting, heating, and cooling. Commercial customers are retail and service shops, administrative buildings, banks, schools, universities, health care buildings, and hotels.

  The short term consumption of electricity for residential and commercial customers is influenced by weather.
- **Industrial customers** use electricity to power and run equipment for their manufacturing processes, in addition to their daily usage for heating, cooling, and lighting.
- **Transportation customers** are electric cars and trains that use electricity as a power source.

The natural gas and electric consumption data are collected per customers type and per operating area. An operating area is a region comprised of a specific set of customers. The data is analyzed, and causes of anomalies are found. The next section of this chapter presents the types of anomalous data that have been encountered in the context of energy demand prediction and explains their causes.

1.4 Anomalous Data in the Energy Domain

Anomalous data can be missing data, unknown patterns, or data modified from its original value [69]. Understanding the sources of anomalies in energy time series data plays an important role in their detection because the definition of false positives depends on the context. This dissertation analyzes data sets representing the reported consumption for residential, commercial, and industrial customers. For those categories of customers, sources of anomalous data include:

- **Missing data or missing components of aggregated data** occur when there are no data values for a specific observation in a univariate data set or when there are no data values for a particular variable of a multivariate data set. Missing data primarily results from errors in data collection or data entry.
- **Electric power generation** is only relevant to the natural gas domain and occurs when the natural gas load used for the generation of electric power is included in the consumption of residential, commercial, or industrial customers. An example of power generation in a natural gas data set is presented in Figure 1.3. Figure 1.3 depicts an abnormally high consumption of natural gas during the summer of 2001, for an operating area where summer loads are typically flat.
- **Main breaks** are unplanned events that occur to the normal consumption of

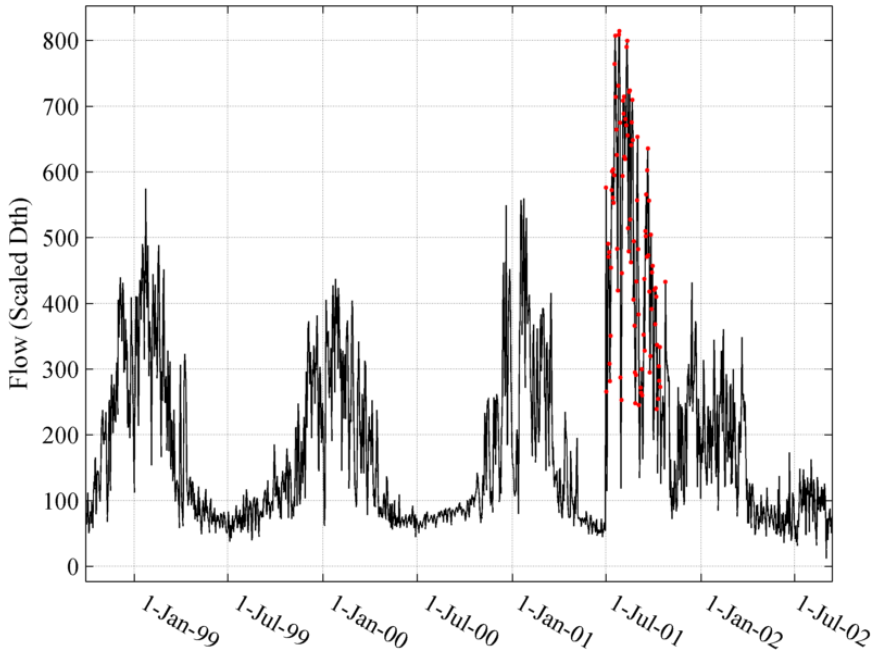

Figure 1.3: Daily natural gas reported consumption for operating area 1

energy, such as a backhoe hitting a pipeline, a tornado knocking down power lines, hurricane, heavy snow days, or service outages.

- **Naïve disaggregation** or a **stuck meter** occur when a normally variable energy load does not vary across several meter reporting periods. An example of stuck meter is presented in Figure 1.4. Figure 1.4 shows a constant natural gas consumption load from 10 April 2011 to 24 May 2011.

- **Negative energy consumption** typically is the result of a system misconfiguration. Energy consumption can be zero but not negative. Negative energy readings may be reported because different pieces of the system (pipelines, types of customers, or corrections) have been mistakenly merged together. Figure 1.5 shows examples of negative flow values that occurred

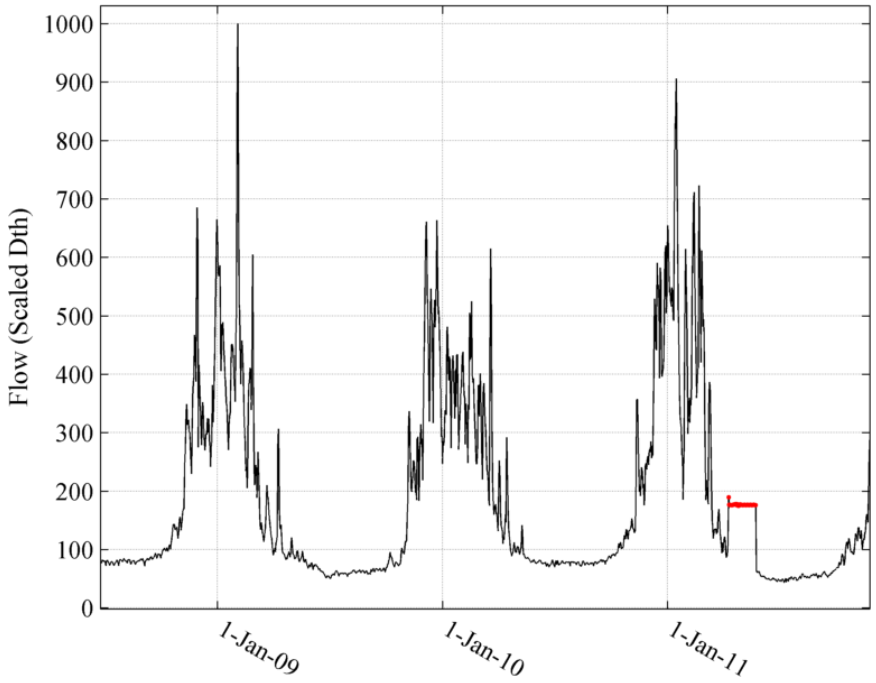

Figure 1.4: Daily natural gas reported consumption for operating area 2

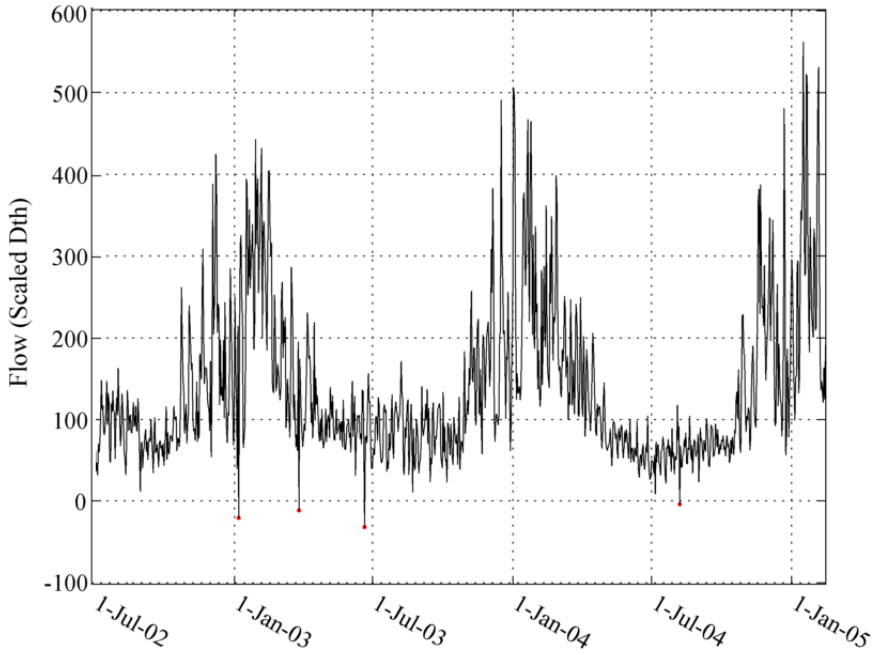

Figure 1.5: Daily natural gas reported consumption for operating area 3

apparently at random in the natural gas time series data of operating area 3.

- **Human error** yields unexpected data values resulting from a bad query, incorrect manual entry reporting, or meter misconfiguration.
- **Mismatched meter factor** or **mismatched units of aggregated data** occur when the meter factor is switched during data collection (usually, the energy load for an operating area is composed of energy loads from various territories) without applying the adjustment factor to previous data (for example kilowatts to watts). It also occurs when the units between subsets of the data are different, and the proper conversion is not applied when merging the data.
- **Outliers** are data points that are dissimilar to the remaining points in the data set [40]. If there is no correlation between energy consumption and the factors driving the consumption, the data point is considered an outlier if no other cause is identifiable.

After their detection, the anomalous data are imputed to obtain a clean signal. The next section presents an outline of the rest of the dissertation.

1.5 Outline of the Dissertation

Chapter 1 of this dissertation introduced the problem of data cleaning and its importance. It also presented the contribution made by this research and gave an

introduction to the energy domain. Chapter 2 provides a literature survey of the techniques used for outlier detection and data imputation in various domains. Probabilistic, machine learning, and statistical approaches have all been used to address the problem of data cleaning. These approaches are presented in Chapter 2 along with the advantages and disadvantages of each approach. An explanation of why some techniques work well in simulated data sets but do not yield good results in practice is also made. Chapter 3 and Chapter 4 present the algorithms developed for data cleaning. The role and contribution of each algorithm to the data cleaning problem is also explained. Chapter 5 presents the results obtained on simulated and real data sets. For the simulation case, the absolute percentage error (APE) between reported and imputed values are calculated. The APE evaluates the false positive rate of the anomaly detection algorithm and the performance of the imputation model. For the real data sets, original and clean data sets are used to forecast out-of-sample signals and compute root mean squared error (RMSE) and mean absolute percentage error (MAPE) measures. Chapter 6 presents a summary of the objectives, the solution proposed, and the contributions made. A conclusion and future research topics on the subject conclude this dissertation.

CHAPTER 2

ANOMALY DETECTION AND DATA IMPUTATION LITERATURE REVIEW

Data cleaning is the process that consists of detecting and imputing anomalous data [96]. The detection step consists of identifying different types of anomalous data while the imputation step consists of deciding on possible corrections for the anomalous values found. This chapter presents the categories of techniques that have been used for anomaly detection and data imputation. A summary of approaches in each category (probabilistic, statistical, and machine learning) is presented along with advantages and disadvantages of each approach. The novelty of the data cleaning algorithm with respect to existing approaches also is presented.

2.1 Anomaly Detection

The first step of the data cleaning process is to detect anomalous data. Anomaly detection refers to the problem of finding patterns in the data that do not conform to expected behavior [24].

Many authors have studied the problem of anomaly detection in various fields such as finance, health care, communication networks, and information technology [90]. The literature depicts two types of outliers in time series data:

additive outliers (AO) and innovative outliers (IO) [28]. An additive outlier is a single observation affected by an anomalous behavior, while an innovative outlier is induced by a random process that also affects the subsequent observations [25]. Typically, additive outliers need to be deleted or replaced because they induce biased variances and estimates [71]. Therefore, this dissertation focuses on additive outliers. The anomalies presented in Section 1.4 are particular cases of additive outliers. Innovative outliers occur as the result of a feedback system that induces an undesired process. Typically, innovative outliers do not require a correction of the measurements because they are noise and usually get corrected when the time series data are modeled [50].

Multiple outliers are especially difficult to detect because of masking, which occurs when one outlier is not detected because of the presence of others [50]. Grané and Veiga showed that masking can be reduced by sequentially correcting anomalies [27, 37].

Graphical approaches such as box-and-whisker plots also have been used for outlier detection, but such approaches are tedious for large data sets and require human input and expertise to obtain accurate results [47, 97]. Therefore, more automatic techniques usually are preferred. The three types of outlier detection approaches that are presented in this chapter are probabilistic, statistical, and

machine learning methods [43]. The next sections of this chapter present the three categories of outlier detection approaches.

2.1.1 Probabilistic Approaches

Probabilistic approaches use probability distribution functions (pdf) to fit the data and calculate parameters of the pdf. Probabilistic approaches identify outliers as data points whose probability is less than some chosen threshold, with respect to the estimated distribution of the data [22]. The anomalies in this case are the data points that deviate considerably from other members of the population [40].

There are two types of probabilistic approaches: parametric and nonparametric. Parametric methods use predefined distribution functions that can be described using a finite number of parameters, for example a Gaussian pdf, which is defined by the mean and variance. Nonparametric methods estimate the density function and the parameters of the model from the data [13].

2.1.1.1 Parametric Approaches

Parametric approaches assume that the data come from a family of known distributions. Buzzi-Ferraris and Manenti developed an approach that uses either Gaussian distribution functions or the median absolute deviation (MAD) for outlier detection [22]. The normal distribution and the MAD use the mean and median as measures of centrality, respectively, and the variance and MAD as measures of

variability, respectively, to fit the data. All data points with values above the threshold corresponding to a probability of 0.95 (5% error) are considered outliers [22]. These techniques do not take into account the number of samples in the data set, hence yielding many false positives for large data sets. Also, the mean and standard deviation are very sensitive to outliers [56].

The main disadvantage of parametric methods is that most distributions are univariate, and the underlying distribution of the observations needs to be known in advance. However, in real data sets, the underlying distribution of the data is not known [8, 67]. Also, there is not an optimal rule for choosing or calculating a rejection threshold.

2.1.1.2 Nonparametric Approaches

Distance-based and density-based methods are nonparametric approaches widely used for outlier detection found in the literature [11].

Distance-based Approaches

Distance-based approaches use the distance between a point and its neighbors to determine if the data point is anomalous. These approaches are efficient on multidimensional data sets [5, 52, 53] but are computationally expensive (usually $\mathcal{O}(n^2)$ time, where n is the number of samples in the data set) [90]. An example of a

distance-based algorithm is presented in Figure 2.1, where the radius r is calculated from the spacial distribution of the data.

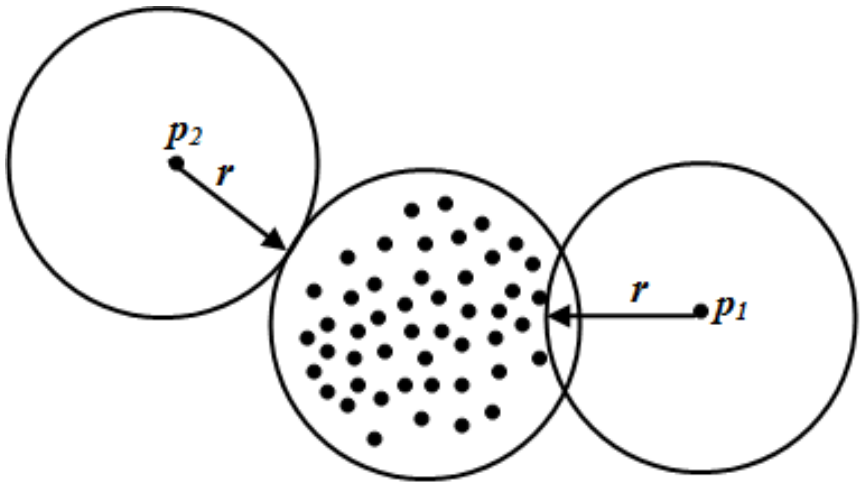

Figure 2.1: Example of a distance-based algorithm

A data point p is considered an outlier if at most α percent of all other points have a distance to p less than r [52], where the threshold α is a chosen parameter. Outliers are depicted by p_1 and p_2 in Figure 2.1, with a chosen threshold of 1% for 50 data points.

Distance-based approaches can be combined with clustering techniques such as the k-nearest neighbor to identify outliers [80], but identifying a good distance measure is difficult in real data sets.

Density-based Approaches

Distance-based approaches only take a global view of the data set. Density-based

approaches overcome this shortcoming by taking a local view of part of the data set [16]. Density-based approaches find anomalies by looking at the local density of the neighborhood of a point. The density of a data point is measured by the number of objects within a given area (or volume) [16]. Density-based techniques score outliers versus the remaining points of the distribution using different measures such as local outlier factors [17], kernel estimation, and Parzen window [42, 65, 91]. An example

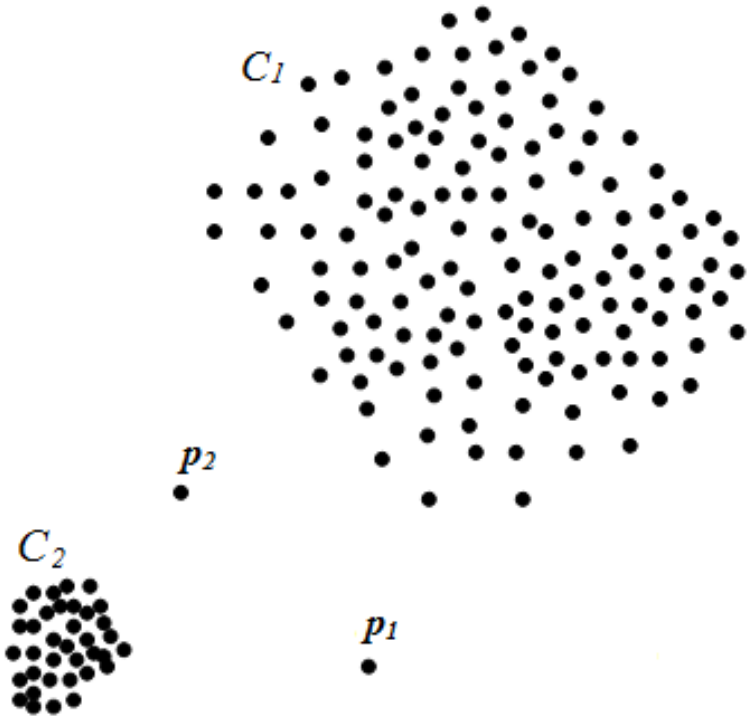

Figure 2.2: Example of a density-based algorithm using a local outlier factor

of a density-based algorithm using a local outlier factor is presented in Figure 2.2, with two clusters C_1 and C_2.

According to distance-based outlier detection approaches, all points in C_2, p_1

and p_2 are outliers because the cluster C_1 is predominant. The result is erroneous because there are two clusters. The data points are scored using the distance-based method. The points p_1 and p_2 are outliers because their respective distances to C_1 and C_2 are greater than the radius of the clusters.

Local outlier factor and kernel estimation techniques use a local distance-based approach for the computation of their scores. However, Parzen window methods interpolate the data to estimate the distribution from which the sample was derived. Density-based approaches are computationally expensive for large data sets and yield false positives because they focus on determining the top-n outliers, where n is a chosen parameter.

Other Approaches

Other nonparametric approaches include ranking or scoring all data according to similarities and differences to determine which ones are inconsistent [17, 24, 101]. Histogram analysis also is a widely used nonparametric technique, in which the frequency of occurrence by classes of data is studied instead of the data itself. The density estimation of the histograms becomes the main issue because the shape can vary significantly depending on the ordering of the classes [36].

Mixture models are another probabilistic approach used for anomaly detection and classification that represent subpopulations within an overall

population. Gaussian mixture models (GMM) estimate the density of the data using a weighted combination of normal distributions. Tarassenko, et al. studied the detection of masses in mammograms using Parzen windows and Gaussian mixture models [91]. The authors showed that GMMs do not work well when the number of training samples is very small, and that Parzen windows work much better on a small number of training samples, but yield false positives. Gaussian mixture models also were used by Tax and Duin to reject outliers based on the data density distribution [93]. They showed that the challenge in using GMMs is selecting the correct number of mixtures. Also, GMM approaches make the assumption that the abnormalities are uniformly distributed outside the boundaries of normality. GMMs work well for multivariate data and are a common descriptor of data, but the outliers need to be well defined.

Bouguessa proposed a probabilistic ensemble approach that uses scores from existing outlier detection algorithms to discriminate automatically between outliers and the remaining points in the data set. Gaussian mixture models, distance-based approaches such as the k-nearest neighbor, and density-based approaches such as the local outlier factor (LOF) are existing techniques that Bouguessa uses for the ensemble model [14]. Each technique provides a score to every data point, and the results are combined to decide which data points are outliers. The ensemble

approach developed by Bouguessa provides better accuracy compared to conventional techniques.

The probabilistic approaches presented above are not able to include domain knowledge. Probabilistic approaches also considered the data as a set of samples without being able to distinguish between features [77]. To incorporate domain knowledge in anomaly detection algorithms, statistical techniques such as auto-regressive moving average (ARMA) or linear regression have been studied and applied to the problem of anomaly detection. The next section presents statistical anomaly detection approaches.

2.1.2 Statistical Approaches

Auto-regressive moving average with exogenous inputs (ARMAX) models and linear regression have both been studied for outlier detection [35, 37, 100, 106]. In statistical approaches, anomalies are data points that deviate, relatively to a chosen threshold or a distance measure, considerably from their predicted values [90]. Anomalies are detected by analyzing the residuals (difference between actual and estimated values). The anomalies affect the structure, parameters, and variance of the models [12]. Therefore, the residuals expose the anomalies.

ARMA models provide a parsimonious description of a stochastic process, where the parameters and constants of the model are derived from the data [15].

The models have two polynomial parts, an auto-regressive function that is stationary and a moving average function that is invertible. ARMA models are efficient at detecting outliers, but the exact order of the polynomial functions for real time series data is difficult to identify [15, 89].

Regression models describe the relationship between a variable to be explained (dependent variable) and its explanatory variables [32]. If the relationship between the dependent and explanatory variables is linear, the models are called linear regression models. An example of a linear regression model for the electric consumption of operating area 4 is presented in Figure 2.3, where the dependent variable is the electric load consumption, and the independent variables are the weather inputs. HDD65 denotes the heating degree days at reference temperature 65°F, and CDD75 denotes the cooling degree days at reference temperature 75°F $\Big($HDD65 = max(0, 65 - Temperature) and CDD75 = max (0, Temperature - 75)$\Big)$. The red lines depict the linear trends found in the data, and the slopes of those lines are the coefficients of the linear regression model which is given by the equation $y_t = 550 + 5\text{HDD65} - 20\text{CDD75}$.

Linear regression models are efficient at using domain knowledge features, because they can be defined as explanatory variables [12, 31, 44]. The explanatory variables need to capture the dynamics of the system. The advantage of linear models is that computationally efficient algorithms exist to calculate the model

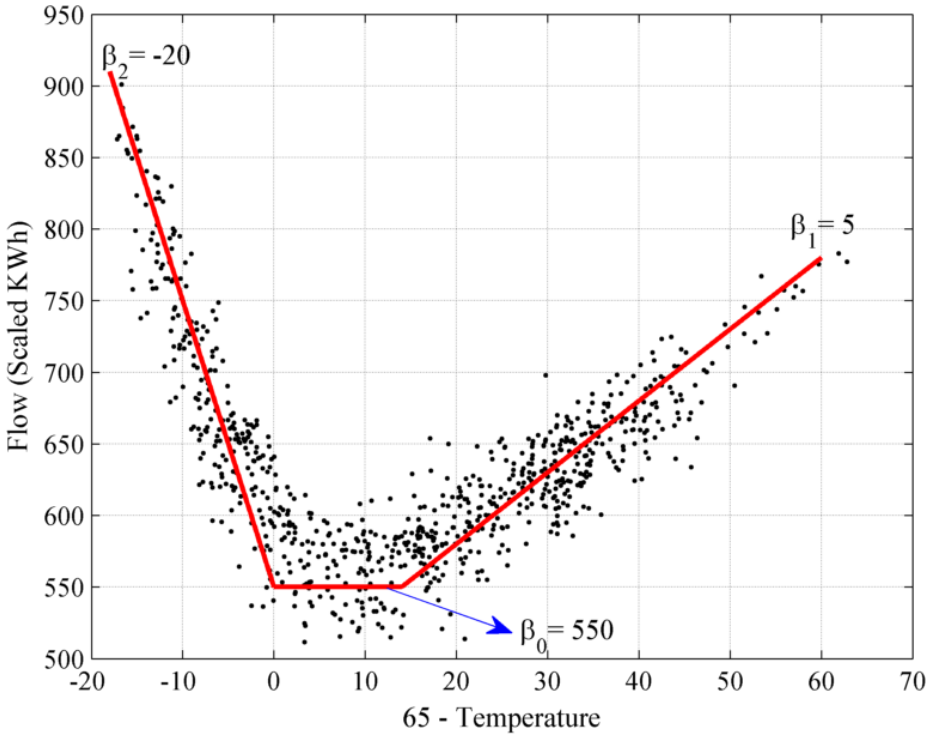

Figure 2.3: Scatter plot of electric load consumption vs. temperature for operating area 4. The red lines depict the trends of the linear regression model.

coefficients. If the dependent variables are explanatory, linear regression is able to identify the characteristics of the time series [66]. Wisnowski, et al. performed an analysis of linear regression models, showing that they perform well in low dimensions and for data sets containing few outliers [102]. Yuen and Mu proposed an approach to calculate the probability of a data point being an outlier by taking into account not only the optimal values of the parameters obtained by linear regression but also the prediction error variance uncertainties [105]. Zou, et al. proposed an approach that uses linear regression in combination with a penalty function to detect outliers [109]. The disadvantage of using a penalty function is

that the design of the tuning parameters needs to be precise. Therefore, penalty function strategies do not guarantee good results on real data sets. For multidimensional data sets, the linearity of the models becomes a limitation, and the definition of the independent variables also becomes complex. Lee and Fung showed that linear and nonlinear regressions can be used for outlier detection, but used a 5% upper and lower threshold limit to choose outliers after fitting, which yielded many false positives for large data sets [59]. Linear regression also has been combined with clustering techniques for the detection of outliers [1].

Hypothesis testing is another statistical technique that draws conclusions about a sample point by testing whether it comes from the same distribution as the training data [67]. Measures such as the t-test and the ANOVA table, which assesses whether the means of two groups of data are statistically different from each other, can be used on multiple subsets of the data to depict the variation of means in subsets that contain anomalies. Also, a level of significance, which corresponds to the probability of incorrectly rejecting the true null hypothesis, needs to be chosen. In statistics, the levels of significance are usually chosen to be 0.05 or 0.01. For smaller intervals where a larger error rate is necessary, 0.05 is selected [58]. For larger intervals in which there is more confidence and where a smaller error rate is necessary, 0.01 is selected [58].

Most statistical algorithms are designed for offline anomaly detection because

of performance. The parameters are usually calculated on historical data. For larger data sets, the estimation of the parameters at each step of the iteration for online anomaly detection becomes computationally expensive. However, statistical methods also have been used successfully for on-line anomaly detection, especially in wireless networks [64, 104]. Machine learning approaches have been developed to learn features from the data itself without prior assumptions of distributions or parameters. Machine learning approaches are presented in the next section.

2.1.3 Machine Learning Approaches

Machine learning approaches have been used for anomaly detection in network intrusion, fraud, medical health, and image processing [41, 91, 107]. Machine learning techniques learn anomalous features from a training set and use them to make predictions on an unseen testing set. There are two types of machine learning tasks: supervised learning, which consists of inferring a function from labeled data, and unsupervised learning, in which the algorithms are trained on unlabeled examples [70]. Supervised learning approaches are classification-based, while unsupervised learning approaches are clustering-based.

2.1.3.1 Clustering-based Approaches

Clustering-based approaches aim to partition the data into meaningful groups (clusters) based on the similarities and relationships between the groups found in

the data [38]. Clustering-based approaches have the objective to assign a score or label to each instance that reflects the degree to which the instance is anomalous [90]. Each data point is assigned a degree of membership for each of the clusters. Anomalies are data points whose cluster memberships are below a given threshold. The accuracy of the techniques depends on how well the structures of the clusters are represented.

The k-means algorithm is a simple iterative clustering approach used for outlier detection. It consists of partitioning the data into k clusters by assigning each data point to its closest cluster centroid and then choosing new centroids for the clusters by calculating their means [103]. The algorithm converges when the cluster assignments no longer change. An approach for outlier detection using the k-means algorithm is to select the top n points that are the furthest away from their nearest cluster centers as outliers [26]. This approach has a near-linear time complexity but yields false positives and negatives because outliers are masked by the clustering [26]. Another approach is to used the (k, n)-means algorithm which simultaneously find the k clusters and n outliers [108]. The problem is NP-Hard, but local optima can be found [49].

The problem with clustering approaches is that a set of anomalies can be considered a cluster rather than anomalies and vice versa, hence providing false

positives and negatives [56]. Also, the cluster degree of membership threshold is difficult to determine correctly.

2.1.3.2 Classification-based Approaches

Classification-based approaches find a concise model of the distribution of class labels in terms of predictor features [55]. The resulting classifier is used to assign class labels to the testing samples and to determine whether they are anomalous. A classifier is an algorithm with features as input that produces a label but also confidence values as outputs [74]. There are several machine learning classification techniques such as neural networks and support vector machines.

Neural networks have been used for outlier detection in diverse domains [41, 91, 107]. Neural networks select one model from a set of allowed models with the goal of minimizing a cost function. An outlier in this case is an observation that does not conform to the pattern of the selected model. The advantage of neural networks is that they can differentiate between anomalies from different classes. The drawback of neural network models is that the training examples and the cost function need to be well defined.

Support vector machines (SVM) are based on finding the optimum hyperplane that separates two data classes [20]. The distance between the separating hyperplanes is called the margin. An SVM classifier finds the maximum

margin necessary to separate two data classes. An example of a support vector machine algorithm result is presented in Figure 2.4. There are two classes of labeled data, represented by squares and dots, and named C_1 and C_2, respectively. The optimal and separating hyperplanes, along with the maximum margin m, also are depicted in Figure 2.4. The data point p_1 is an outlier because it is misclassified.

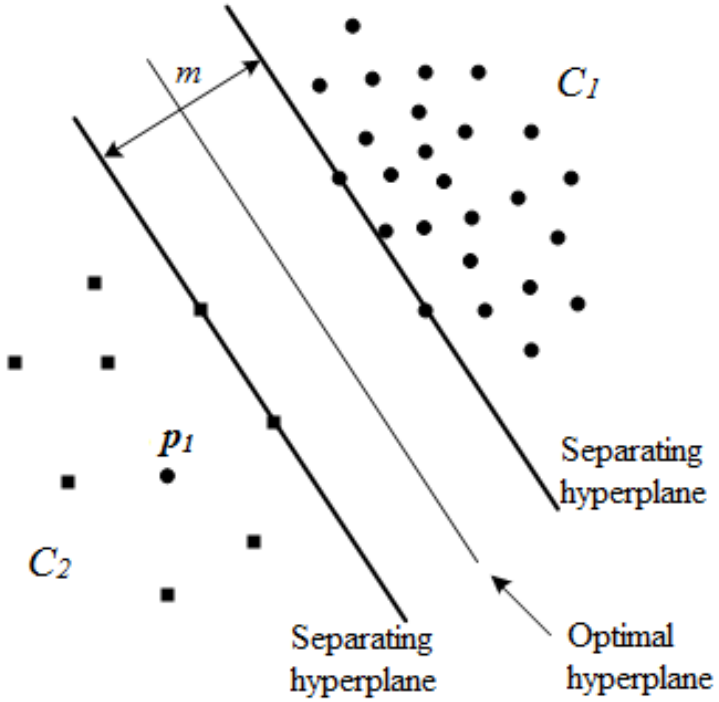

Figure 2.4: Example of a support vector machine algorithm

Another family of classifiers widely used because of their performance are Bayesian classifiers, which combine probabilistic and machine learning approaches. They apply Bayes' theorem with strong independence assumptions between the features given the classes [81]. The anomalous features are learned on a training set and used to classify any unseen data points as anomalous or not. Because time series

data are not the outcomes of a random process, Bayesian techniques are difficult to apply to time series data. Therefore, the data is transformed from the time domain to a phase space to extract the multidimensional features of the data [79, 85]. An approach developed by Sauer and Yorke demonstrated that transforming the signal from a time domain to a phase space improves the classifier [85]. The approach works well for small training samples and for multivariate data. Bayesian classifiers are robust because they model the underlying distribution of the data.

The evaluation of classifiers are based on prediction accuracy. Because anomalous classes are usually much smaller than normal classes, measures such as precision, recall, and false positive percent are more appropriate to evaluate the results of the classifiers. The drawback of supervised anomaly detection is that they require the existence of a training set with both anomalous and clean data [90]. Therefore, they yield false positives if they are not accurately trained. Classification-based approaches are efficient on multidimensional data sets and are able to highlight features in a data set. They also have a low complexity and are able to classify any unseen data points according to the training features.

Machine learning approaches have been developed to learn features from the data itself without prior assumptions of distribution or parameters. However, modeling the underlying distribution of data sets using probabilistic approaches can be used to improve the accuracy of machine learning techniques.

To make valid and efficient inferences about the data, anomalous data needs to be imputed after their detection. The anomalous data are marked as missing, and missing data imputation techniques are used to find replacement values. The next section of this chapter presents approaches used in the literature for data imputation.

2.2 Data Imputation

Data collection is often costly. Usually, there are not enough data points to discard the anomalous and missing ones. In this case, data replacement becomes important. Ad hoc edits are avoided because they produce biased, inefficient, and unreliable results [86]. Therefore, an imputation of the missing data is necessary. Data imputation consists of discarding or replacing missing values or fields with suitable or substitute estimates. The problem is found in various fields such as the social sciences, medical fields, and engineering. Data imputation techniques are chosen based on the missingness of the data [48].

There are two types of missingness of the data: missing completely at random (MCAR) and missing at random (MAR). Data is MCAR if the cases with missing data are a random subset of the cases with complete data and MAR if the cases with missing data are related to the cases with complete data [82]. For example, let us consider the data collected and presented in Table 2.1 for an operating area. The

data is composed of flow values, date stamps (from 01 through 08 January), and actual temperature values and contains missing temperature and flow values.

Table 2.1: Example of multivariate data set with missing values

Date	01/01	02/01	03/01	04/01	05/01	06/01	07/01	08/01
Flow	500		300	450	375		425	325
Temperature	65	70	55		60	57		58

If there is no correlation between missing temperature values and missing flow values, the missing data is MCAR. If all missing temperature and flow values have the same date stamps, the missing data is MAR. In practice, data are assumed to be MCAR because the reasons for the data being missing are beyond the control of the researchers.

MCAR cases are either discarded or replaced. MAR cases use model-based methods for data imputation [39]. Little's test of MCAR is a statistical test that can be used to test the missingness of a multivariate data set [61]. For bi-variate data and with missing data confined into a single variable, Little's test is reduced to a standard t-test. The next sections of this chapter present the various imputation methods found in the literature depending on the data missingness.

2.2.1 Imputation Methods for Data Missing Completely At Random

There are three possible ways of imputing MCAR data. Missing data can be ignored, replaced by known values, or replaced by values estimated using the data features.

2.2.1.1 Imputation Using Only Valid Data

There are two methods for imputing missing data using valid or available data: complete case analysis or listwise deletion and available case analysis or pairwise deletion. Complete case analysis consists of excluding all cases with missing values. It is the easiest data imputation technique and is usually the default procedure in most statistical packages [39]. Available case analysis consists of using all available data to estimate parameters of the model. In listwise deletion, cases with partial data are not considered for parameters estimation, as opposed to pairwise deletion [78]. In multivariate data sets, available case analysis uses different sets of samples for different parameter estimation. Because parameters are estimated from different sample sets, it is difficult to compute standard errors [86], and the estimates of the parameters are biased.

Let us consider again the data set of Table 2.1. The data set has two missing flow values (corresponding to 02/01 and 06/01) and two missing temperature values (corresponding to 04/01 and 07/01). The result obtained using listwise deletion

imputation is presented in Table 2.2, which has all four days of either missing flow or temperature deleted.

Table 2.2: Listwise deletion result for the data set of Table 2.1

Date	01/01	03/01	05/01	08/01
Flow	500	300	375	325
Temperature	65	55	60	58

The resulting pairwise deletion imputation performed on the same data set of Table 2.1 is presented in Tables 2.3 and 2.4. To estimate the temperature and flow variables, the data sets used are the ones from Tables 2.3 and 2.4, respectively. The two samples sets resulting from available case analysis have different time stamps and are therefore two distinct sample sets.

Table 2.3: Pairwise deletion for the data set of Table 2.1 on the variable "Temperature"

Date	01/01	02/01	03/01	05/01	06/01	08/01
Temperature	65	70	55	60	57	58

Table 2.4: Pairwise deletion for the data set of Table 2.1 on the variable "Flow"

Date	01/01	03/01	04/01	05/01	07/01	08/01
Flow	500	300	450	375	425	325

The main virtue of listwise and pairwise deletion is their simplicity. They are effective when the data set contains few anomalies. However, they reduce the sample size, which decreases the statistical power and the precision of parameters and values estimation [84].

2.2.1.2 Imputation Using Known Replacement Values

Imputing missing data on a variable using known replacement values consists of replacing the missing data by a value that is chosen from an estimate of the distribution of the variable [33]. The advantage of an imputation using known replacement values compared to imputation using only valid data is that it retains all the data and their features, hence improving the estimation of model parameters. Data imputation methods using known replacement values are case substitution, hot deck imputation, and cold deck imputation.

Case substitution consists of replacing the entire case containing missing values with another similar non-sampled case [39]. Hot and cold deck imputation methods replace missing values of one or more variables for a recipient with observed values from a donor [4]. In cold deck imputation, the donor is an external source, whereas in hot deck imputation, the donors are similar cases in the same data set.

For imputation techniques using known replacement values, a similarity measure needs to be defined. Also, additional cases not in the training set and external values are necessary to find good replacements. These imputation methods are applicable only when the number of missing variables is limited.

2.2.1.3 Imputation by Calculating Replacement Values

Imputation using only valid data and imputation using known replacement values have the pitfall of requiring special formulas for standard errors and produce biased estimates. Therefore, imputing missing data by calculating replacement values was developed to overcome those limitations. The advantage of data imputation using calculated replacement values is that it retains all the data and their features [63]. Mean substitution and regression are two methods used to compute replacement values for missing data.

Mean substitution replaces missing values of a variable with the mean of the observed values of that particular variable [84]. Mean substitution preserves the mean of the distribution but distorts estimates of variance and covariance [62]. It is best used for data sets containing few missing observations and when there is a strong relationship between variables.

Regression methods impute missing values of a variable based upon its relationship to other observed variable values (predictors). Single regression imputation is the particular case of using only one predictor to estimate the replacement value. The approach is univariate and does not suffice for multi-features and multidimensional data sets. Also, the standard errors are too small, and the imputed values do not reflect uncertainty because they are found from only one predictor [83].

In multiple regression imputation, multiple predictors are defined to model the distribution of the data. They take into account the imprecision of estimating the distribution of the variables with missing values [33]. Multiple regression is robust at estimating missing values and efficiently handles patterns of missing data [63]. It is also used to estimate the conditional distribution of an outcome given specific inputs [99]. However, there should be a substantial correlation between the variables with missing data and other variables. Also, the predictors variables need to be well defined.

2.2.2 Imputation Methods for Data Missing At Random

Model-based approaches are suitable for the imputation of data missing at random because they provide the best representation of original distribution of values with the least bias [39]. Model-based approaches estimate replacement values for missing data based on all non-missing data for a given variable. One of the model-based approaches widely used for the imputation of MAR data is maximum likelihood. Maximum likelihood estimation consists of drawing inferences from a likelihood function. In maximum likelihood, parameter values with the highest probability are assigned using a likelihood function. Then, the parameters are estimated based on all available data, including the incomplete cases. The maximum likelihood estimation can be done with algorithms such as expectation maximization. Expectation maximization is a computational method for estimating likelihood from

incomplete data sets in which the observed values provide indirect evidence about the likely values of the unobserved ones [87]. The expectation step computes the expected value of the complete-data log-likelihood and the maximization step maximizes the resulting function to provide new parameter estimates [30]. Other methods that can be used to estimate maximum likelihood include Bayesian methods and Gaussian mixture models.

Maximum likelihood has the advantage of being able to handle high levels of missing data. However, it is a large sample tool, and the sample should be large enough for the estimates to be approximately unbiased and normally distributed [86]. Only rarely does real data conform to normality, but many tools are available to help preserve distributional shape [87]. Model-based approaches can also be used to calculate replacement values for data missing completely at random (MCAR).

One of the main disadvantages of data imputation is that the imputed values are treated as real data, which overstates their precision [63]. However, many statistical methods used to study time series require samples with complete values, or they yield biased and inadequate parameters estimation in case of large missing data.

2.3 Importance of our Data Cleaning Algorithm

The approach proposed in this dissertation for anomaly detection is a hybrid model that combines the advantages of statistical and probabilistic methods to provide accurate theoretical and practical results. Time series are not distribution sets. Therefore, statistical approaches are used to extract domain knowledge of the data sets and calculate the residuals. The residuals form a data set where anomalies can be found in the tails of the distribution. Probabilistic methods are used to model the residuals and calculate the probability of each data point to discriminate between data belonging to the underlying distribution and anomalies.

After their detection, identified anomalies are marked missing and are imputed. The process is done repetitively to avoid masking. Data imputation on data sets still containing anomalies leads to higher than necessary prediction error due to bias in the estimates. However, Nahi presented the case of recursive estimation to correct the problem of biased estimates [72]. The recursion of the process avoids biased parameter estimates in the analysis.

Because the data are time series in this dissertation, the missing data is not ignorable unless it is located at the beginning or at the end of the data set. Usually, the missing data will be missing completely at random because the anomalies are additive outliers. Therefore, replacement values are calculated for data imputation using multiple regression imputation. Multiple regression imputation is chosen

because it represents a good balance between quality of the results and ease of use. Multiple regression imputation also results in correctly estimated standard errors and confidence intervals [33]. The trends modeled for anomaly detection using statistical methods give a good basis of the general relationships among the variables in energy data sets.

Often, data imputation depends on domain knowledge of the problem studied. For example, imputation models for electric and natural gas demand prediction are different. In this dissertation, the general regression model defined for anomaly detection can also be used for energy time series imputation. However, the model only provides a naïve imputation in case no other values are available. Forecasting models for natural gas or electric time series can be substituted in the algorithm to improve the data cleaning process. Kaynar, et al. presented a study of forecasting natural gas demand using various models such as auto-regressive moving average, artificial neural networks, and ensemble models [51]. Shen, et al. presented a combination of various machine learning techniques to forecast electricity demand time series [88]. Carmona, et al. also presented an electric demand forecast model using neural networks [23]. Since 1993, the GasDay laboratory at Marquette University has developed a variety of energy demand forecasting techniques and analytical tools that are also presented in data cleaning examples in this dissertation.

This chapter presented a literature survey on anomaly detection and data imputation. It also presented the motivation behind the methods chosen for data cleaning in this dissertation. The next chapter of this dissertation presents the hypothesis-driven anomaly detection algorithm that is the underlying method for the data cleaning process.

CHAPTER 3

HYPOTHESIS-DRIVEN ANOMALY DETECTION ALGORITHM

This chapter presents the first major contribution of this dissertation, an anomaly detection algorithm based on probability distribution functions. The hypothesis-driven anomaly detection (HDAD) algorithm is a probabilistic approach that detects anomalies in a data set where the underlying distribution of the data points is known or assumed. An example is presented to illustrate the algorithm. The complexity of the algorithm is analyzed to conclude this chapter.

3.1 Algorithm

The hypothesis-driven anomaly detection algorithm is based on statistical hypothesis testing. Let X be a set of observations assumed to be drawn from a probability distribution function, and let x be an element of X. The extrema of the data set X are identified as potential anomalies. A statistical hypothesis test determines if an extremum is anomalous. The null and alternative hypotheses are H_0: (extremum is not anomalous) and H_1: (extremum is anomalous), respectively.

The null hypothesis is rejected in favor of the alternative hypothesis with a level of significance α, which is the probability of incorrectly rejecting the true null hypothesis or committing a type I error. An α of 0.01 is used in this dissertation.

Let the experiment $E = \{\text{Classifying an extremum}\}$. The possible outcomes of the experiment E are "anomaly" or "not an anomaly". Let p be the probability that the chosen extremum is drawn from the underlying distribution of the remaining elements in the data set.

$$p = P\{x \sim \text{Distribution}(X \setminus \{x\})\}, \tag{3.1}$$

where x is an extremum, and $\{x\}$ is the set of all elements whose values are the same as x. If the probability of "anomaly" in the experiment E is p, the probability of "not an anomaly" is (1 - p). Each classification of an extremum is an independent experiment. Therefore, the experiment E is a Bernoulli trial. The problem is reduced to finding the number of Bernoulli trials needed to find an anomaly in at least n trials and supported by the set of n samples. This corresponds to the cumulative distribution function of a geometric distribution [76]. The cumulative distribution function of a potential anomaly should be less than the level of significance α for the data point to be considered anomalous.

$$1 - (1 - p)^n < \alpha. \tag{3.2}$$

When calculating the cumulative distribution function of the geometric distribution, finite precision becomes a limitation. For a large sample size and a low probability p, the value of the cumulative distribution function is truncated to zero. In this

case, an approximation of the geometric cumulative distribution function in the limit of the neighborhood of zero is used instead of Equation 3.2. The geometric cumulative distribution function is $1-(1-p)^n$. Rewriting (3.2) yields

$$1-(1-p)^n = 1-e^{n\ln(1-p)}. \tag{3.3}$$

$$\lim_{p\to 0} \ln(1-p) = -p. \tag{3.4}$$

Therefore, for $p \ll 1$, (3.3) reduces to

$$1-e^{-np}. \tag{3.5}$$

Furthermore, if $p \ll 1$ and $np \ll 1$,

$$\lim_{np\to 0} 1-e^{-np} = np. \tag{3.6}$$

The probability p depends on the extremum value and the underlying distribution of the remaining points in the data set. By taking into account the number of samples and the probability of each point belonging to the underlying distribution, the hypothesis-driven algorithm sets an effective bound on the number of potential anomalies in the data set. Most importantly, the algorithm detects

Algorithm 1 HYPOTHESIS-DRIVEN-ANOMALY-DETECTION

Require: data set X, level of significance α, assumed distribution Dist(X, β)

potentialAnomalies $\leftarrow$ true
indices $\leftarrow \emptyset$

while potentialAnomalies **do**

% The minimum ($\underline{x}$) and the maximum ($\overline{x}$) values of X are chosen as potential
% anomalies, and the parameters of their corresponding distributions are found
$X_{min} \leftarrow X \setminus \{\underline{x}\}$
$X_{max} \leftarrow X \setminus \{\overline{x}\}$
$\text{Dist}_{min} \leftarrow$ estimate parameters β_{min} from X_{min}
$\text{Dist}_{max} \leftarrow$ estimate parameters β_{max} from X_{max}

% Compute the probability that the potential anomalies belong to the
% underlying distribution of the remaining data points
$p_{min} \leftarrow \text{Probability}\big(x_{min} \sim \text{Dist}_{min}\big)$
$p_{max} \leftarrow \text{Probability}\big(x_{max} \sim \text{Dist}_{max}\big)$

% Determine if $\underline{x}$ or $\overline{x}$ are anomalous based on the level of significance α
$g_{min} \leftarrow 1 - (1 - p_{min})^n$
$g_{max} \leftarrow 1 - (1 - p_{max})^n$
if $(g_{max} < \alpha) \vee (g_{min} < \alpha)$ **then**

% The extremum which has the lowest probability is considered anomalous
if $(g_{min} < g_{max})$ **then**
$X \leftarrow X_{min}$
indices $\leftarrow$ indices $\cup$ index($\{\underline{x}\}$)
else
$X \leftarrow X_{max}$
indices $\leftarrow$ indices $\cup$ index($\{\overline{x}\}$)
end if
else

% Exit condition: $\underline{x}$ and $\overline{x}$ are not anomalies at the level of significance α
potentialAnomalies $\leftarrow$ false
end if
end while
return X, indices

points that are most unlikely to be drawn from the assumed underlying distribution. Therefore, the technique can be used with any assumed distribution.

The values of the two potential anomalies, $\overline{x}$ (maximum value of the data set) and $\underline{x}$ (minimum value of the data set), are examined simultaneously at each iteration of the algorithm. The extremum that has the lowest probability of belonging to the assumed distribution of the remaining points in the data set, and which the geometric cumulative distribution function value is less than α, is considered anomalous [3]. The HDAD algorithm is presented in Algorithm 1. The disadvantage of the HDAD algorithm is that the data set is assumed to be samples drawn from a distribution. The next section presents an illustrative example to explain the HDAD algorithm.

3.2 Hypothesis-Driven Anomaly Detection Algorithm Example

The residuals obtained from the time series of operating area 6 using a 6-parameter linear regression model is used as example (see Subsection 4.3) and is presented in Figure 3.1. The data set has 2,192 samples and the level of significance is 0.01. Only the first iteration of the algorithm is presented here.

The maximum and minimum values of the data sets are potential anomalies, presented in Figure 3.2. For this example, a normal distribution function is used to calculate the probability of an extrema belonging to the remaining points in the

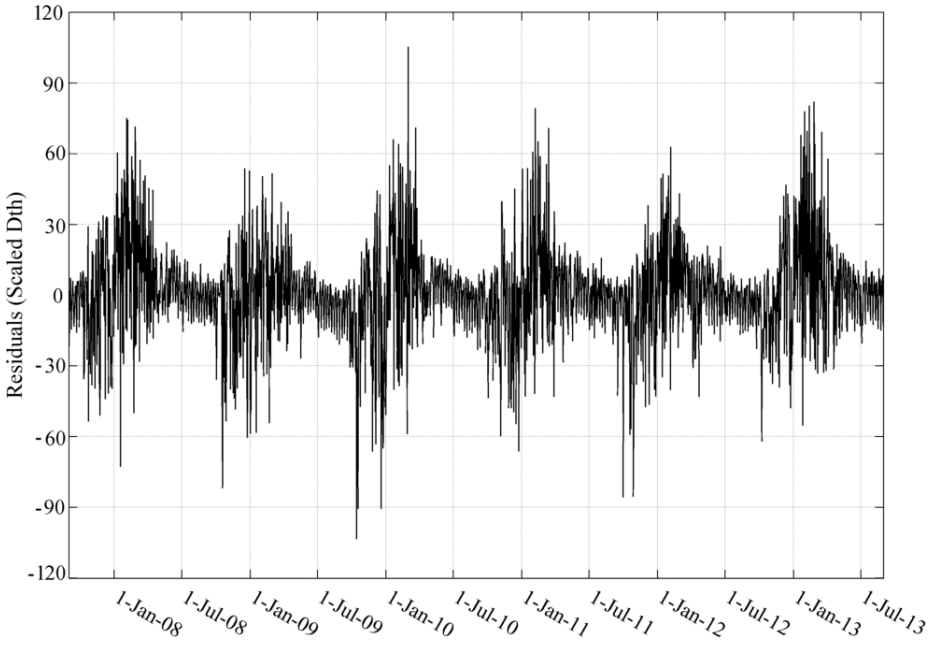

Figure 3.1: Residuals calculated using the time series data set of operating area 6 and a 6-parameter linear regression model

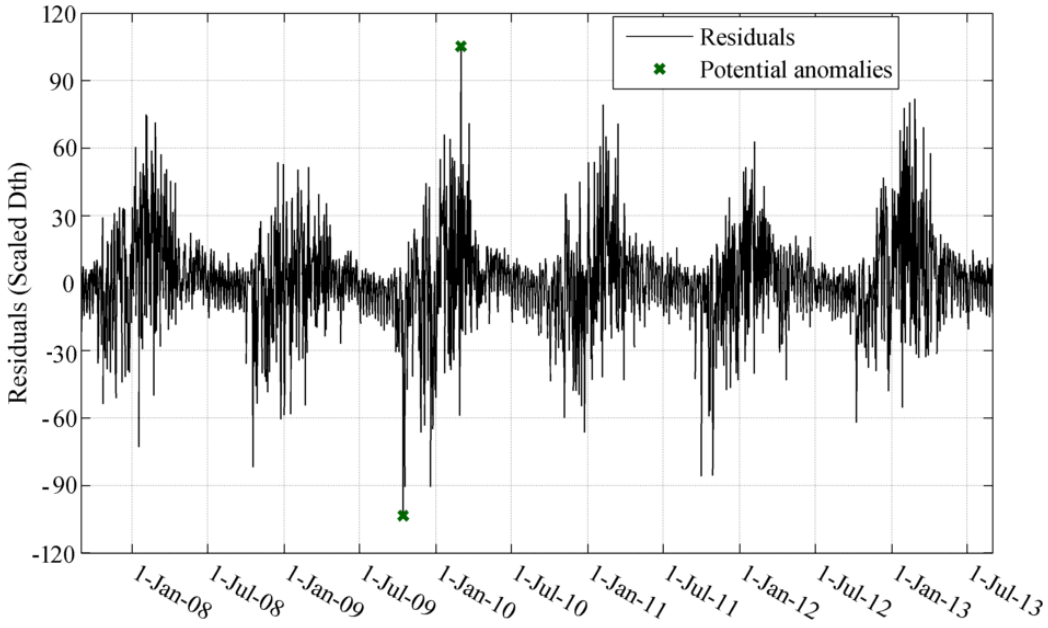

Figure 3.2: Residuals data set with potential anomalies highlighted

data set. The Gaussian probability distribution function is fit to the normalized frequency of the residuals, presented in Figure 3.3.

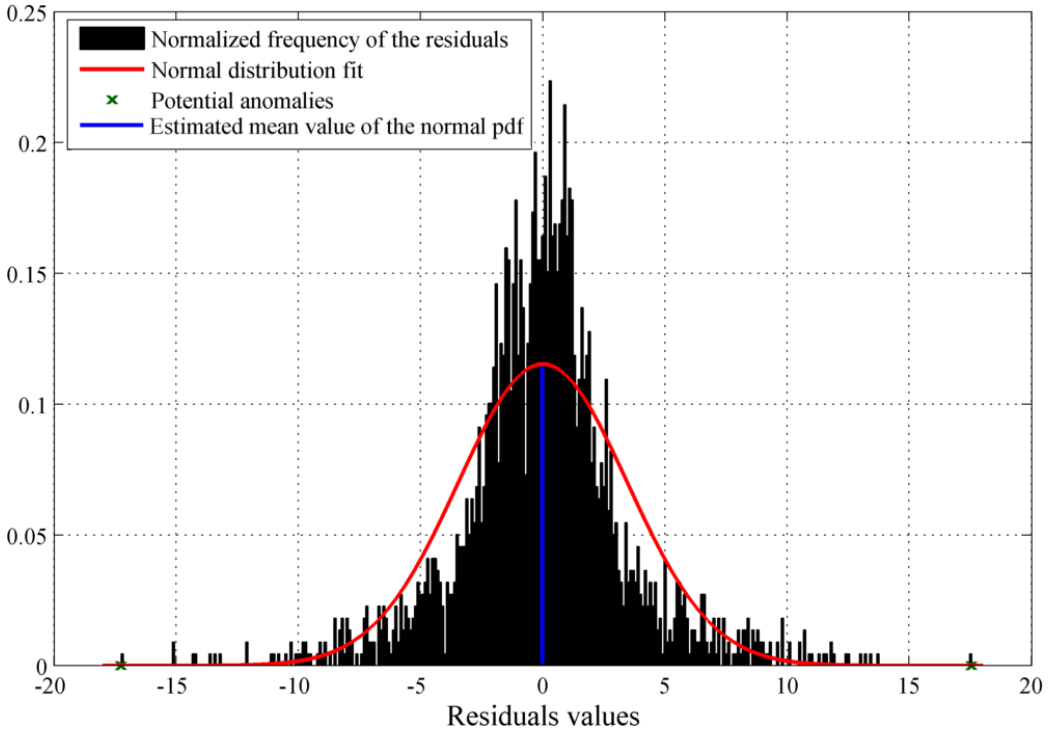

Figure 3.3: Normalized frequency of the residuals fit with a Gaussian pdf, potential anomalies, and mean value of the distribution

The probabilities of the extrema are calculated using the parameters of the fitted normal distribution. The probabilities found ($p_{min} = 2.65\times10^{-7}$ and $p_{max} = 1.68 \times 10^{-7}$) are then used in Equation 3.2 to determine whether the extrema are anomalous. $g_{min} = 1 - (1 - p_{min})^{2192} = 5.8\times10^{-4}$, and $g_{max} = 1 - (1 - p_{max})^{2192} = 3.7\times10^{-4}$. g_{min} and g_{max} are both less than the level of significance, but g_{max} is smaller than g_{min}. We conclude that the maximum value

of the residuals set is an anomaly at the level of significance of 0.01. The residual plot with the anomaly depicted is presented in figure 3.4.

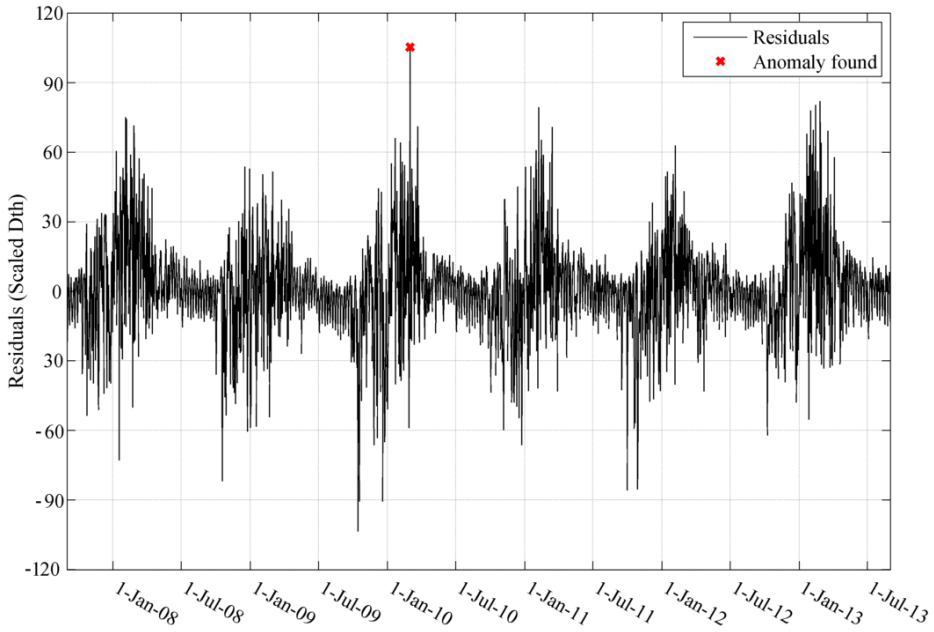

Figure 3.4: Residuals plot with the first anomaly found depicted by a red cross

The next section presents a brief analysis of the computational complexity of the HDAD algorithm and discusses the impact for very large data sets or for a disk-based implementation.

3.3 Complexity Analysis of the HDAD Algorithm

The HDAD algorithm is able to handle very large data sets. Even for a disk-based implementation, the run time is still reasonable. There are two options that can be considered for the implementation of the HDAD algorithm, assuming a Gaussian

probability distribution function. The two options are presented and explained below.

3.3.1 Option 1: with sorting

Option 1 consists of first sorting a data set X. The sort operation requires $\mathcal{O}(n \log n)$ operations. After the sorting operation, the extrema are located at the beginning and at the end of the data set. Therefore, selecting a potential anomaly is done in $\mathcal{O}(1)$. Assuming a normal distribution, sum(X) and sum(X^2) are calculated only once at the beginning of the algorithm in this case, which requires $\mathcal{O}(n)$ operations. To compute the mean and the variance at each iteration of the algorithm, the value represented by each potential anomaly is removed from the sums accordingly, and the number of samples is modified, which requires $\mathcal{O}(1)$ operations. If m is the number of potential anomalies found, option 1 requires

- Sort: $\mathcal{O}(n \log n)$,
- Initial calculation for μ and σ^2: $\mathcal{O}(n)$,
- Adjust μ and σ^2: $\mathcal{O}(m)$.

In summary, option 1, which uses sorting, requires $\mathcal{O}(n \log n)$ to detect anomalous points in a data set.

3.3.2 Option 2: with pointers

Option 2 consists of tracking the position and value of a potential anomaly at each iteration using pointers. The operation of choosing an extremum requires $\mathcal{O}(n)$ operations in this case. The initial computation of the mean and the variance requires $\mathcal{O}(n)$ operations. The sums are also accumulated here to perform the adjustment of the mean and the variance in $\mathcal{O}(1)$ as in option 1. Each adjustment requires $\mathcal{O}(n)$ operations to remove an anomaly from the data set. If m is the number of potential anomalies found, option 2 requires

- Find an extrema: $\mathcal{O}(n)$,
- Initial calculation for μ and σ^2: $\mathcal{O}(n)$,
- Adjust μ and σ^2: $\mathcal{O}(mn)$.

In summary, option 2, which uses pointers requires $\mathcal{O}(mn)$ to detect anomalous points in a data set.

Overall, option 2 has a lower computational complexity than option 1 if m is less than log$n$. The problem with a disk-based implementation is that a disk access is many orders of magnitude slower than memory access, and the bandwidth of a disk is about 50 times less than a single random access memory subsystem [57]. To find a potential anomaly with option 2, the entire data set on the hard drive is scanned. The adjust step in option 2 has a large hidden constant, in the case of a

disk-based implementation. Thus, option 1 becomes the best option for a data set that is too large to fit in memory. The cost of scanning the hard drive to find an extrema is far greater than the cost of initially sorting the data set and knowing the position of the extrema. Option 2 is used in this dissertation because the data sets used do not exceed 12 years of data, and the position of the anomalies are necessary for data imputation.

This chapter explained the HDAD algorithm along with an illustrative example and discussed the complexity of the algorithm. However, probabilistic approaches do not perform well on time series because of their variability. Therefore, the HDAD algorithm is combined with statistical methods to perform data cleaning on time series data. Chapter 4 of this dissertation presents the linear regression data cleaning algorithm along with an illustrative example.

CHAPTER 4

LINEAR REGRESSION DATA CLEANING ALGORITHM

This chapter presents the linear regression data cleaning algorithm. The linear regression data cleaning algorithm models time series features using a linear regression model and applies the HDAD algorithm on residuals to find the largest anomaly at each iteration. The linear regression models are used to fit the historical data set, and the process is done iteratively to reduce masking. The algorithm is improved with a rule-based anomaly detection, which is an ensemble of energy domain knowledge rules used to improve the data cleaning process. A description of the inputs to the algorithm, which are energy information, weather, and level of significance, is first provided. Then, the rule-based anomaly detection is described. The linear regression data cleaning algorithm, which is a combination of rule-based anomaly detection, HDAD algorithm, and linear regression models, concludes this chapter along with an illustrative example.

4.1 Inputs to the Algorithm

The inputs to the data cleaning algorithm are energy information, weather, and a level of significance. The energy information is the signal to be cleaned. The temperature and wind are the weather data used as exogenous inputs to the algorithms. They are used to model the amounts of energy necessary for heating

and cooling. The level of significance is the threshold at which a residual data point is considered anomalous. A detailed explanation for each of these inputs is presented in the next sections.

4.1.1 Energy Information

An energy time series is a sequence of N observations sampled uniformly in time [15]. It represents the average daily or hourly consumption of natural gas or electricity by operating area (a region comprised of a specific set of customers).

$$y = \{y_t,\ t = 1, ..., N\}. \tag{4.1}$$

An example of natural gas reported consumption from 01 September 2007 through 31 August 2013, for an example operating area is depicted in Figure 4.1. All data sets presented in this dissertation are natural gas and electricity data from utilities. However, the data sets are scaled to maintain confidentiality.

The measure of natural gas is the British thermal unit (Btu), which corresponds to the amount of energy required to raise the temperature of one pound of water by one degree Fahrenheit at the temperature at which water has its greatest density (39°F) [34]. The natural gas industry generally expresses natural gas in Decatherms (Dth), where one Dth is equivalent to one million Btu. The

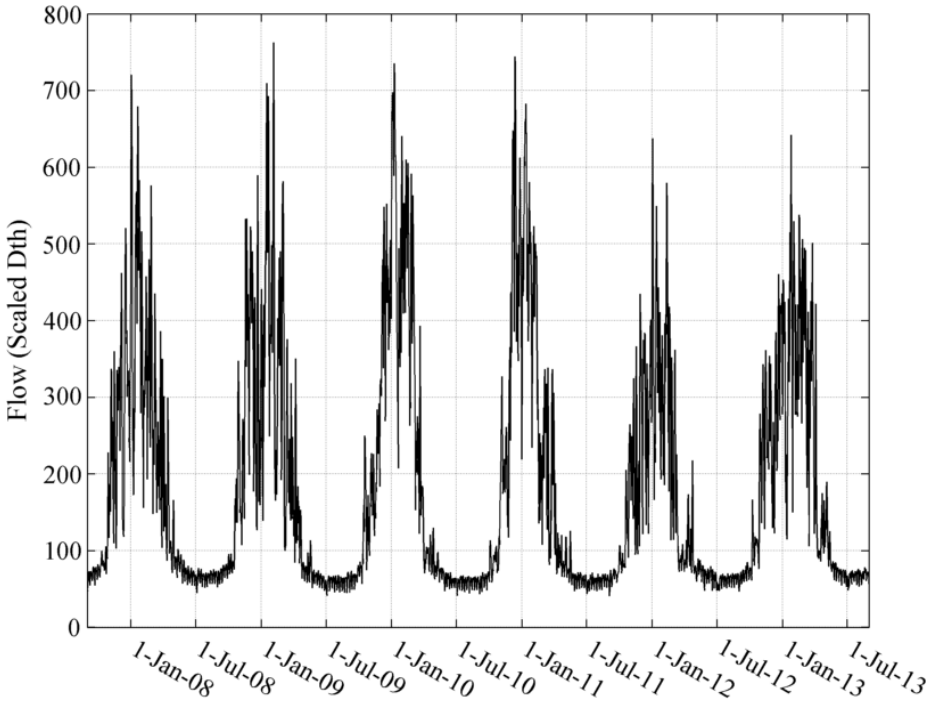

Figure 4.1: Natural gas reported consumption of operating area 5 from 01 September 2007 to 31 August 2013

measure of electric power is in watts (W), which is the amount of power defined as one joule in one second [46].

4.1.2 Weather

Both natural gas and electricity consumptions are influenced by weather conditions. Temperature and wind affect the consumption of energy [98]. The relationship between temperature and energy is nonlinear. This nonlinearity is caused by human behavior. A degree day or degree hour is used to model this nonlinearity. The heating degree days (HDD), heating degree hours (HDH), cooling degree days (CDD), and cooling degree hours (CDH) are the differences between the mean daily

or hourly temperatures and a base temperature [21]. The base or reference temperature is the temperature below or above which heating or cooling is needed, respectively [9]. If T is the daily or hourly average temperature for an operating area and the heating reference temperature is T_{ref_H},

$$\text{HDD}T_{ref_H} \text{ or } \text{HDH}T_{ref_H} = \max(0, T_{ref_H} - T). \tag{4.2}$$

Similarly, cooling degree days or cooling degree hours are defined as

$$\text{CDD}T_{ref_C} \text{ or } \text{CDH}T_{ref_C} = \max(0, T - T_{ref_C}), \tag{4.3}$$

where T_{ref_C} is the cooling reference temperature.

Energy usage is also affected by wind. The wind speed influences how quickly heat loss is conducted through buildings' walls [29]. Therefore, the HDD and HDH are usually wind-adjusted. There is no influence of wind on warmer days. Let w represent the wind speed in miles per hour (mph). The wind-adjusted HDD is

$$\text{HDDW}T_{ref_H} = \max\left(\frac{72+w}{80}, \frac{152+w}{160}\right) \text{HDD}T_{ref_H}. \tag{4.4}$$

The wind-adjusted HDH ($\text{HDHW}T_{ref_H}$) is also calculated using the same formula. The wind-adjustment has been found in practice to not affect heating degree days at 8 mph [98]. Below 8 mph, the wind speed effect is propagated slowly, and the energy

transfer is slower. Above 8 mph, the energy transfer occurs more quickly. The temperature values used in this dissertation are expressed in degree Fahrenheit (°F).

4.1.3 Level of Significance

The level of significance corresponds to the smallest probability value at which a data point is considered anomalous. The level of significance used for data cleaning in this dissertation is 0.01 because the data sets have enough sample points to support this level of significance.

The next section of this chapter presents the rule-based anomaly detection. The rule-based anomaly detection is a way of pre-processing the data, used to improve the data cleaning process.

4.2 Rule-based Anomaly Detection

The rule-based anomaly detection is a set of energy domain knowledge rules used to improve the data cleaning process. In this dissertation, the only rule used is missing data replacement. The linear regression data cleaning algorithm is auto-regressive. Therefore, all missing data points are imputed to avoid distorted estimates of the linear regression model coefficients. The rule is to find all missing data in the set and perform an imputation using calculated values at the beginning of the data cleaning process. The new signal improves the data cleaning process, reduces masking, and provides better estimates.

The next section of this chapter presents and describes the linear regression data cleaning algorithm. The role and use of each input also is explained in the description of the algorithm.

4.3 Linear Regression Data Cleaning Algorithm

The hypothesis-driven anomaly detection (HDAD) algorithm is a probabilistic approach that detects anomalies in a data set. However, energy time series are not stationary and are not the outcomes of random processes. Therefore, the HDAD algorithm is applied on residuals of time series, and the time series features are extracted using other techniques.

A time series can be decomposed into four elements: trend, seasonal effects, cycles, and residuals [6]. Therefore, the idea behind this approach is that fitting the data with linear regression models extract the trend, seasonal effects, and cyclical characteristics of the data set. The residuals found form a distribution of points in which anomalies are detected using hypothesis testing. The algorithm uses an n-parameter linear regression model for anomaly detection, where (n - 1) is the number of inputs. The general form of the model is

$$\widehat{y} = \beta_0 + \beta_1 \mathrm{HDDW}T_{ref_H} + \beta_2 \mathrm{CDD}T_{ref_C} + \beta_3 y_{-1} \tag{4.5}$$

for a daily set, and

$$\widehat{y} = \beta_4 + \beta_5 \text{HDHW} T_{ref_H} + \beta_6 \text{CDH} T_{ref_C} + \beta_7 y_{-1} \tag{4.6}$$

for an hourly set. Heating and cooling degree days can be calculated at multiple reference temperatures for an operating area to capture the climate of the region. Therefore, the number of model parameters varies according to the number of reference temperatures. The y and $\widehat{y}$ represent the reported and estimated energy values for a time t, respectively. The β_0 and β_4 are the baseloads or the minimum amounts of non-varying load of energy. The reference temperatures, T_{ref_H} and T_{ref_C}, are the reference temperatures for heating and cooling, respectively. β_1 and β_5 are the amounts of energy used per heating degree day or heating degree hour, respectively, at reference temperature T_{ref_H}. The β_2 and β_6 are the amounts of energy used per cooling degree day or cooling degree hour, respectively, at reference temperature T_{ref_C}. The energy consumption for a particular day usually depends also on the energy consumption and temperature of the previous day [44]. Therefore, the β_3 and β_7 are the changes in energy consumption between two consecutive days or hours, respectively. The flow diagram of the linear regression data cleaning approach is summarized in Figure 4.2.

At the first iteration, missing values and negative flow values are found and imputed. Then, the linear regression model coefficients are re-calculated on cleaner

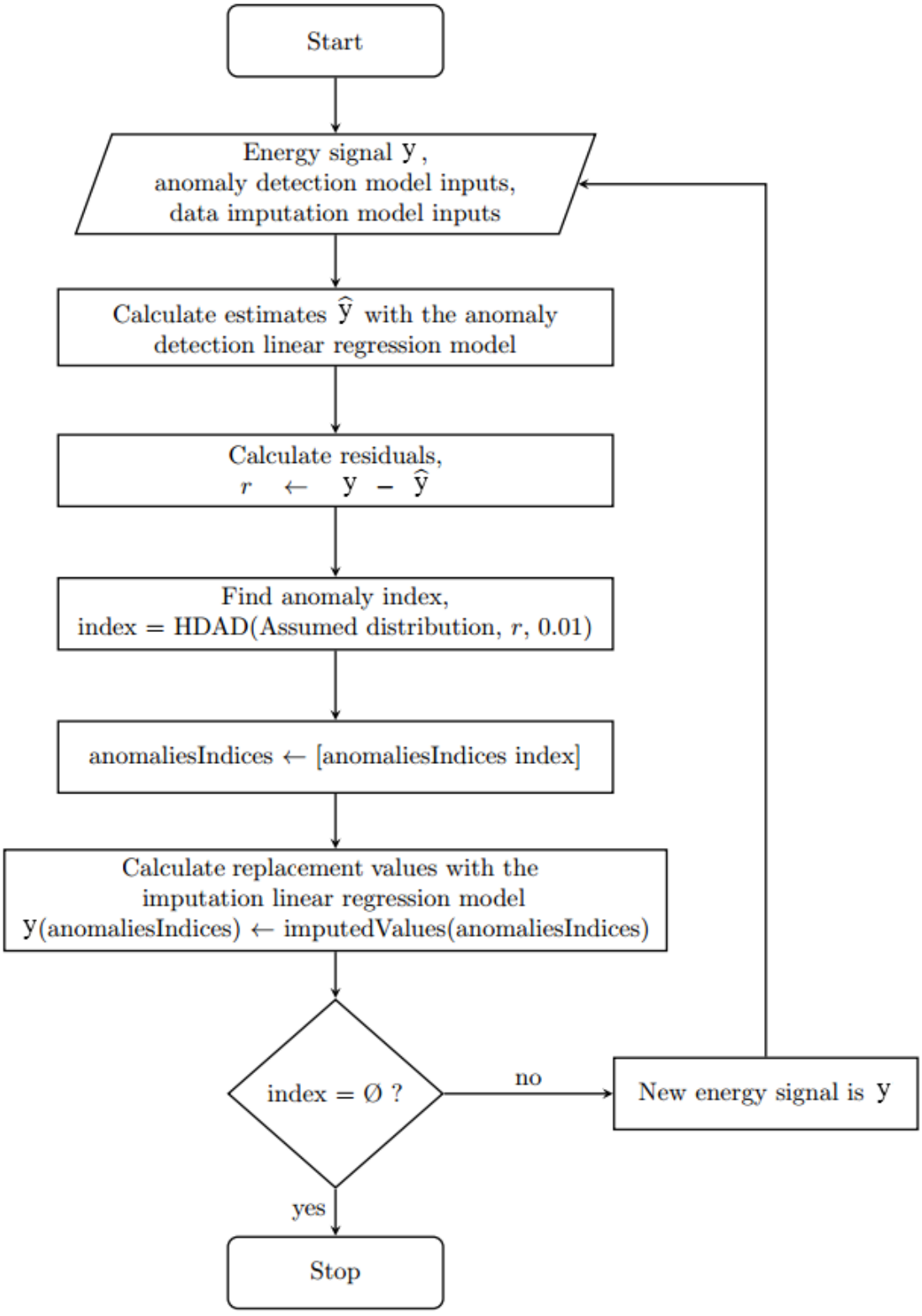

Figure 4.2: Flow diagram of the linear regression data cleaning algorithm

data at each iteration of the algorithm. The algorithm is assumed to extract all time series properties without distorting anomalies. The estimated coefficients are possibly erroneous at the beginning of the process because it is uncertain whether the data set contains only one anomaly. An extremum is an anomaly if its probability of belonging to the same distribution as the other residual values is less than the probability of committing a type I error at the specified level of significance α.

After an anomaly is removed, the process continues with the anomalous value replaced by an imputed value. This step improves the data cleaning process by removing masking. The model parameter estimation for anomaly detection improves after each anomaly is replaced. The algorithm stops when no more anomalies are identified. The final imputation values are recalculated on the clean signal after all anomalies are found.

Because the data cleaning process is implemented iteratively to reduce masking, at each iteration of the algorithm, the HDAD algorithm is modified to report only the largest anomalous value in the residuals. The set of all elements in the residuals data set, whose values are equal to the largest anomalous value, are all considered anomalous and their positions are also returned by the HDAD algorithm. The algorithm used to detect the largest anomaly value and the linear regression data cleaning algorithm pseudo-codes are presented in Algorithms 2 and 3.

Algorithm 2 DETECT-LARGEST-ANOMALY

Require: data set X, level of significance α, assumed distribution Dist(X, β)

indices $\leftarrow \emptyset$

% The minimum ($\underline{x}$) and the maximum ($\overline{x}$) values of X are chosen as potential
% anomalies, and the parameters of their corresponding distributions are found
$X_{min} \leftarrow X \setminus \{\underline{x}\}$
$X_{max} \leftarrow X \setminus \{\overline{x}\}$
$\text{Dist}_{min} \leftarrow$ estimate parameters β_{min} from X_{min}
$\text{Dist}_{max} \leftarrow$ estimate parameters β_{max} from X_{max}

% Calculate the probability that the potential anomalies belong to the underlying
% distribution of the remaining data points
$p_{min} \leftarrow \text{Probability}\big(x_{min} \sim \text{Dist}_{min}\big)$
$p_{max} \leftarrow \text{Probability}\big(x_{max} \sim \text{Dist}_{max}\big)$

% Determine if $\underline{x}$ or $\overline{x}$ are anomalous based on the level of significance α
$g_{min} \leftarrow 1 - (1 - p_{min})^n$
$g_{max} \leftarrow 1 - (1 - p_{max})^n$

% The extremum which has the lowest probability is considered anomalous
% Otherwise, the algorithm returns the empty set
if $(g_{max} < \alpha) \vee (g_{min} < \alpha)$ **then**
    if $(g_{min} < g_{max})$ **then**
        indices $\leftarrow$ index($\{\underline{x}\}$)
    else
        indices $\leftarrow$ index($\{\overline{x}\}$)
    end if
end if
return indices

Usually, more complex models that include domain knowledge are suitable for imputation because they model the particularities of each data set or utility system. However, the linear regression model used for anomaly detection also can be used for data imputation, in cases were no other model is available. The model is simple enough to provide best-guess estimated values, but it is not complex enough

Algorithm 3 ENERGY-TS-LINEAR-REGRESSION-DATA-CLEANING

Require: energy time series y, temperature T, wind w, α, T_{ref_H}, T_{ref_C}, assumed distribution Dist(X, β), imputationModel(A, γ), imputationInputs A

potentialAnomalies $\leftarrow$ true
anomalies $\leftarrow \emptyset$

% Find and impute all missing flow values
anomalies $\leftarrow$ find all missing elements of y

% Impute all anomalies found
replacementValues $\leftarrow$ imputationModel(A, y, γ)
y(anomalies) $\leftarrow$ replacementValues(anomalies)

% Calculate non-varying inputs to the anomaly detection linear regression model
detectionInputs $\leftarrow \begin{bmatrix} 1 & \text{HDDW}T_{ref_H} & \text{CDD}T_{ref_C} \end{bmatrix}$

while (potentialAnomalies) **do**

    % Include the first lag of y as input and calculate the estimated values $\widehat{y}$
    % and the residuals
    detectionInputs $\leftarrow \begin{bmatrix} \text{detectionInputs} & y_{-1} \end{bmatrix}$
    $\beta \leftarrow$ Coefficients of the linear regression model LR(y, detectionInputs)
    $\widehat{y} \leftarrow \beta \times$ detectionInputs
    residuals $\leftarrow y - \widehat{y}$

    % Find the largest anomalies in the residuals at the level of significance α
    indices $\leftarrow$ DETECT-LARGEST-ANOMALY$\big($residuals, α, Dist(residuals, β)$\big)$

    if indices == $\emptyset$ **then**
        % Exit condition: No more anomalies found
        potentialAnomalies $\leftarrow$ false
    else
        % Impute all anomalies found and continue iterating
        replacementValues $\leftarrow$ imputationModel(A, y, γ)
        anomalies $\leftarrow$ anomalies $\cup$ indices
        y(anomalies) $\leftarrow$ replacementValues(anomalies)
    end if
end while
return anomalies, y

to model all the particularities of a data set. The forecasting models also can be composed of an ensemble of various techniques and can vary depending on the type of energy data (natural gas and electricity). Various imputation models can be substituted easily into the linear regression data cleaning algorithm. In this dissertation, the imputation model used for energy time series is a forecasting model using weather, data trends, and seasonality components.

The linear regression data cleaning algorithm illustrates both the anomaly detection and imputation process. It also shows the iterative nature of the process. Algorithm 3 also is applicable to the case of an hourly data set with daily inputs replaced by hourly inputs. The next section presents an illustrative example of the linear regression data cleaning algorithm.

4.4 Linear Regression Data Cleaning Algorithm Example

An illustrative example is presented in this section to clarify and explain the linear regression data cleaning algorithm. The data set presented in Figures 4.3 and 4.4 is the daily natural gas reported consumption of operating area 6, and ranges from 01 September 2007 to 31 August 2013. Figure 4.3 and Figure 4.4 show the time series and scatter plots of the data set, respectively. The scatter plot shows the relationship between the daily natural gas flow and the daily average temperature.

Coefficients of a 6-parameter linear regression model, calculated using least

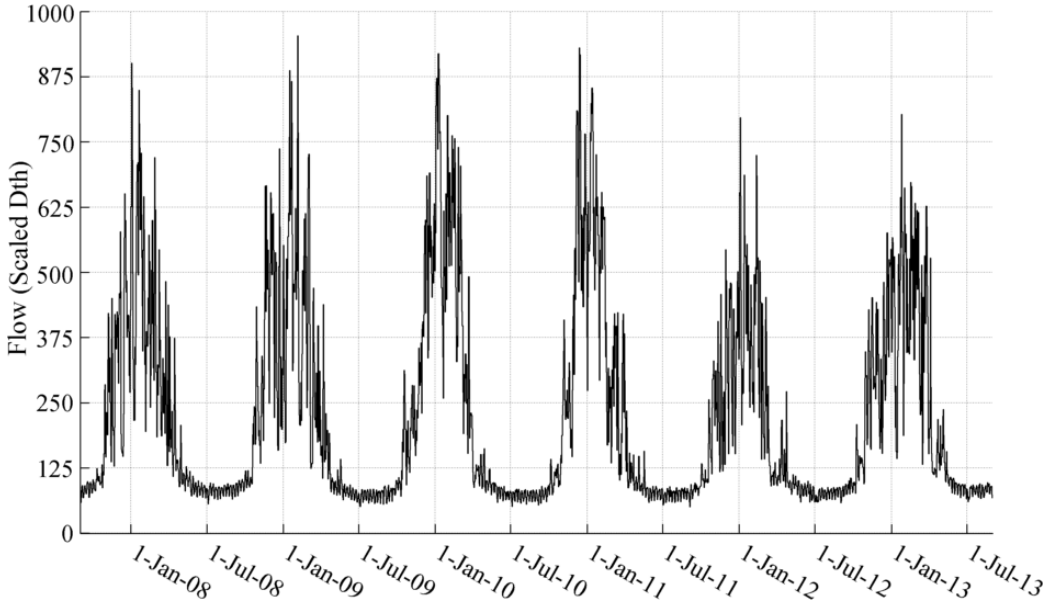

Figure 4.3: Time series plot of the natural gas reported consumption of operating area 6

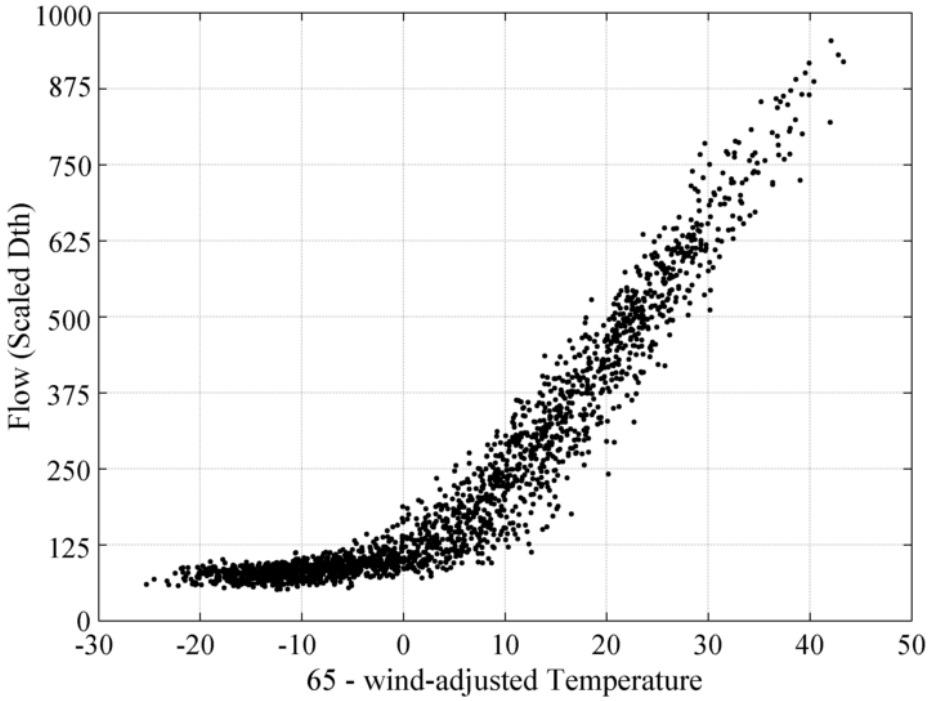

Figure 4.4: Scatter plot of the natural gas reported consumption of operating area 6

squares fitting, are obtained with the HDDW evaluated at reference temperatures 65°F and 55°F and the CDD evaluated at reference temperatures 65°F and 75°F.

$$\widehat{y} = \beta_0 + \beta_1 \text{HDDW55} + \beta_2 \text{HDDW65} + \beta_3 \text{CDD65} + \beta_4 \text{CDD75} + \beta_5 y_{-1}. \qquad (4.7)$$

The estimated values and the residuals, which are the difference between the reported consumption and the estimated values, are calculated using the coefficients of the linear regression model.

- ***First iteration***

At the first iteration, the DETECT-LARGEST-ANOMALY algorithm is used to find the largest anomaly in the residuals. The reported energy consumption, the estimated values, the residuals, and the first anomaly found on the residuals are depicted in Figure 4.5. The first anomaly found is the maximum value of the set of residuals.

A 21-parameter linear regression model (inputs are historical load, trends, seasonality components (day of week, week of the month, month of the year, and holidays), weather information (HDDW, CDD, and the change in HDDW between two consecutive days), and cross terms between weather and seasonality) is used to calculate replacement values for the anomalies. The time series plot, with the

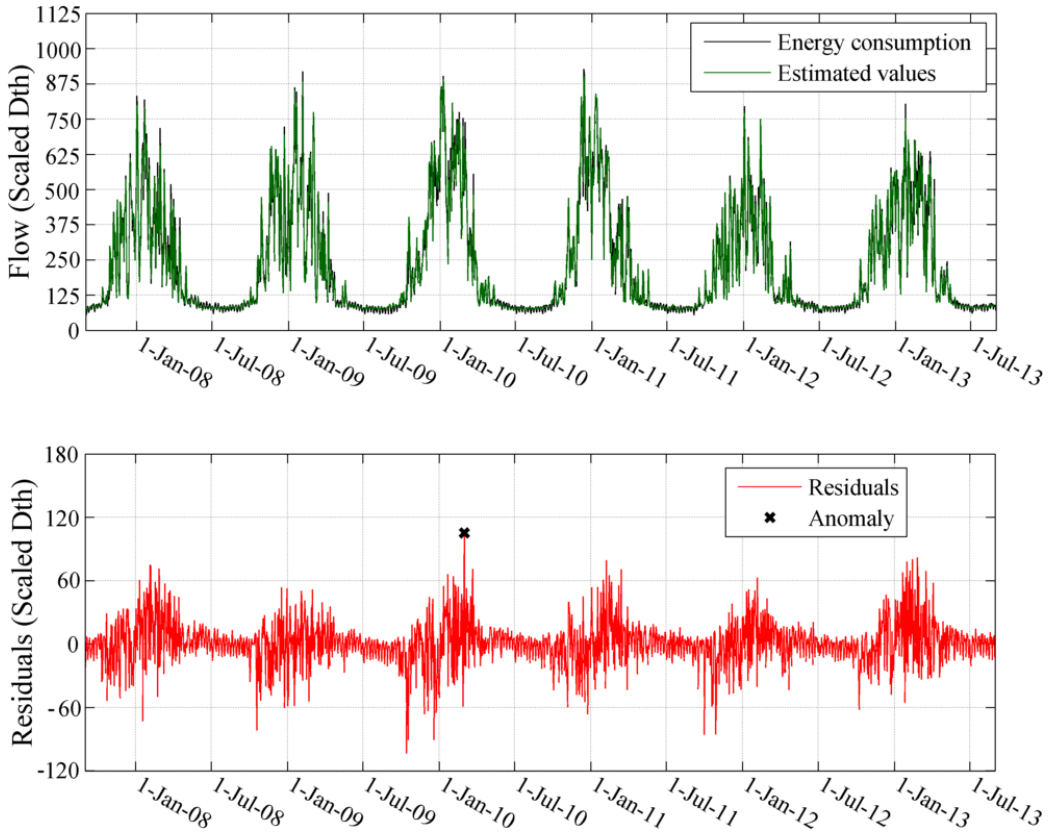

Figure 4.5: Time series plot of the energy versus estimated values and plot of the residuals with the first anomaly found depicted by a black cross

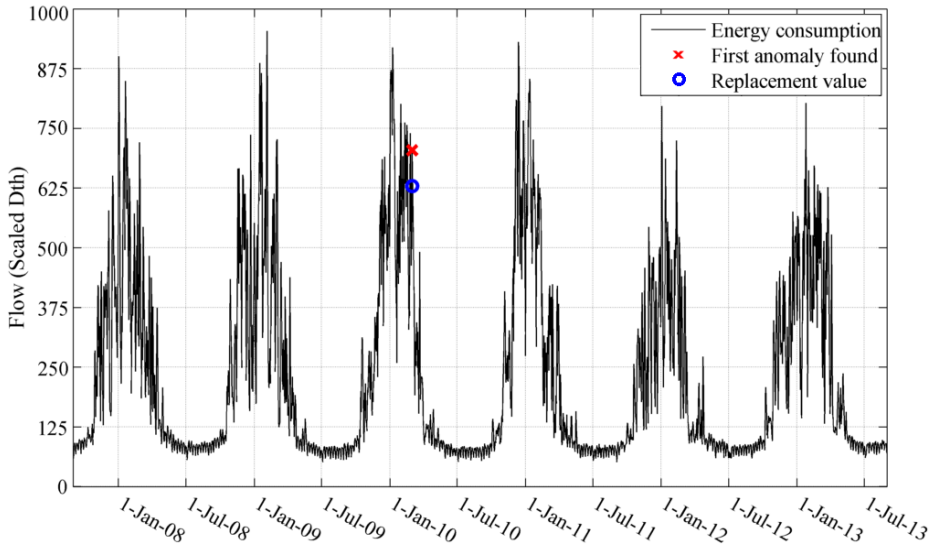

Figure 4.6: Time series plot showing the first anomaly and its replacement value depicted with a red cross and a blue circle, respectively

anomaly found and its corresponding replacement value depicted with a red cross and blue circle, respectively, is shown in Figure 4.6.

- ***Second iteration***

After the beginning of the second iteration, the time series signal is the original data set with the first anomaly replaced. The 6-parameter linear regression model is used to re-calculate a new set of residuals, in which the second anomaly is found at the level of significance of 0.01. The new time series plot, with the first anomaly replaced and the second anomaly found are shown in Figure 4.7.

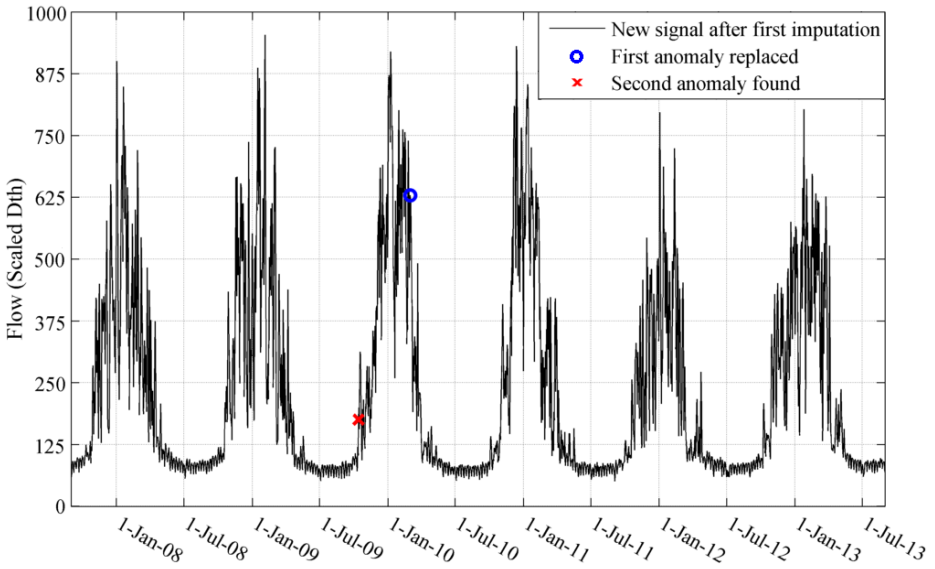

Figure 4.7: Time series plot of the energy signal at the beginning of the second iteration with the second anomaly found depicted by a red cross

The set of residuals found at the first and the second iteration, along with all the anomalies found so far, are presented in Figure 4.8.

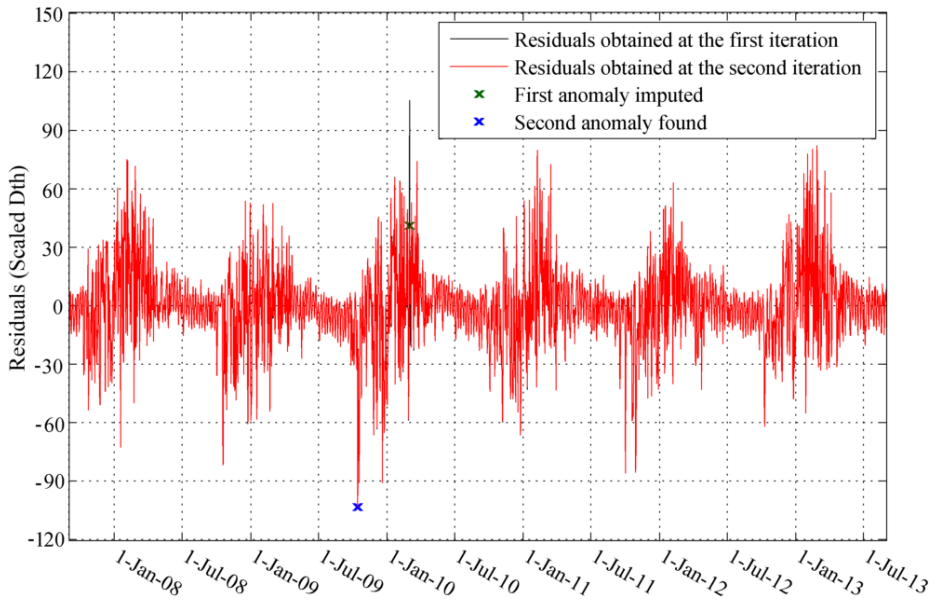

Figure 4.8: Change in residuals from the first to the second iteration, with anomalies depicted

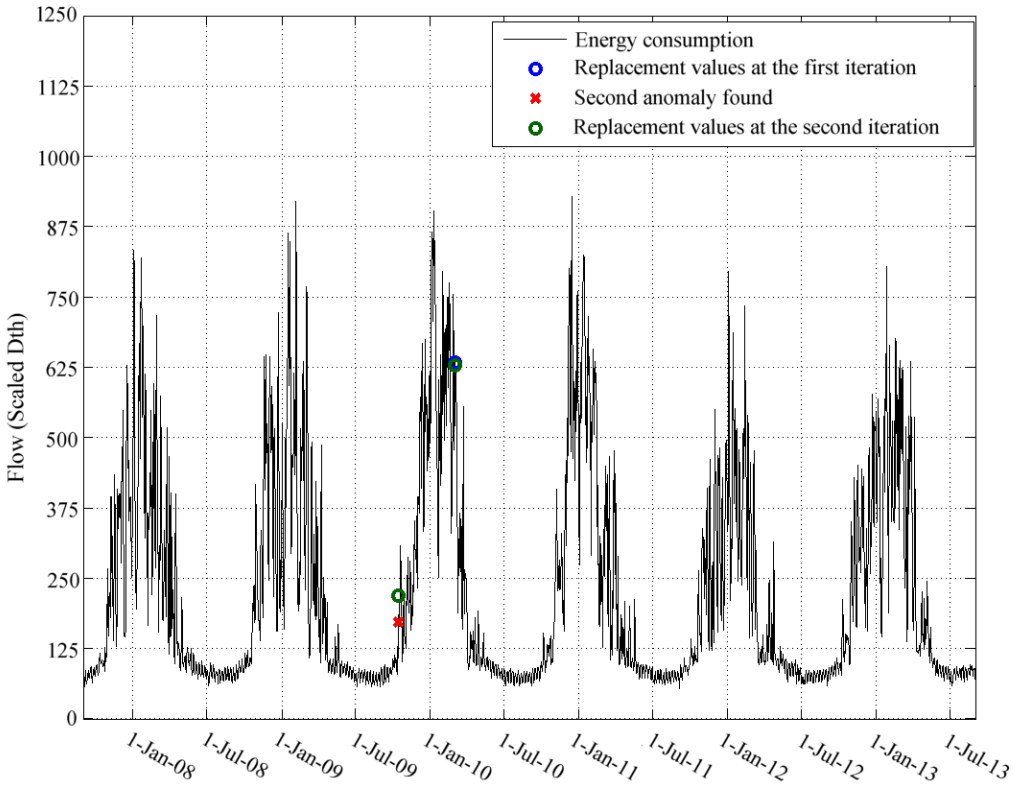

Figure 4.9: Time series plot with the second anomaly found and the new replacement values depicted

The anomaly found at the second iteration is the minimum of the set of residuals as opposed to being the maximum as in the first iteration. The replacement values are re-calculated for all anomalies with the same model used at the first iteration. The second anomaly found, along with the new replacement values for all anomalies, are depicted on the time series plot in Figure 4.9.

New replacement values are re-calculated at each iteration because the data gets cleaner as each anomaly is identified and replaced.

- ***Final iteration***

The algorithm stops when no more anomalies are found. Final replacement values are calculated for all the anomalies found. The result of the linear regression data cleaning algorithm is summarized in Figures 4.10 and 4.11. Figures 4.10 and 4.11 depict all three anomalies identified and their corresponding replacement values with red crosses and blue circles, respectively.

This chapter explained the techniques developed for data cleaning, along with an example to clarify the algorithm. Chapter 5 of this dissertation presents the evaluation of the data cleaning algorithm. A simulation study that evaluates the false positive and negative rates of the algorithm is made. Utilities' data sets also are used to test the algorithm and evaluate the improvement of forecasting accuracy obtained by cleaning the data.

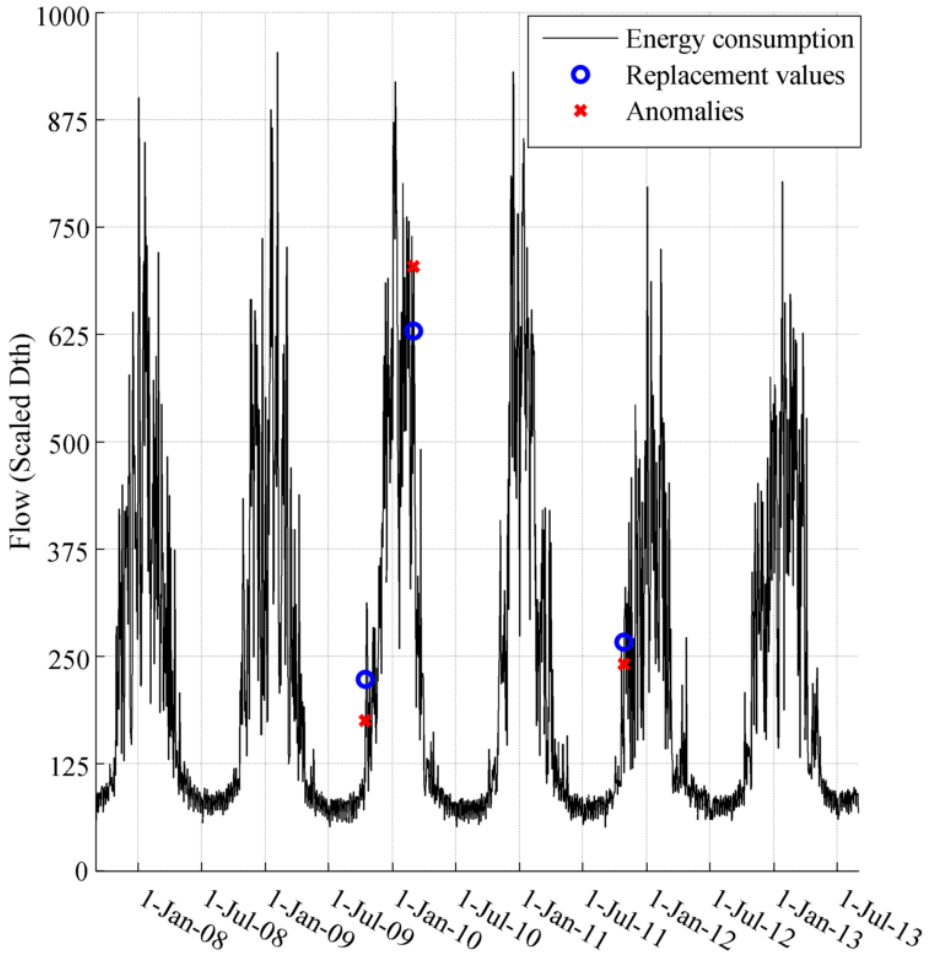

Figure 4.10: Time series plot depicting all the anomalies identified and their corresponding replacement values

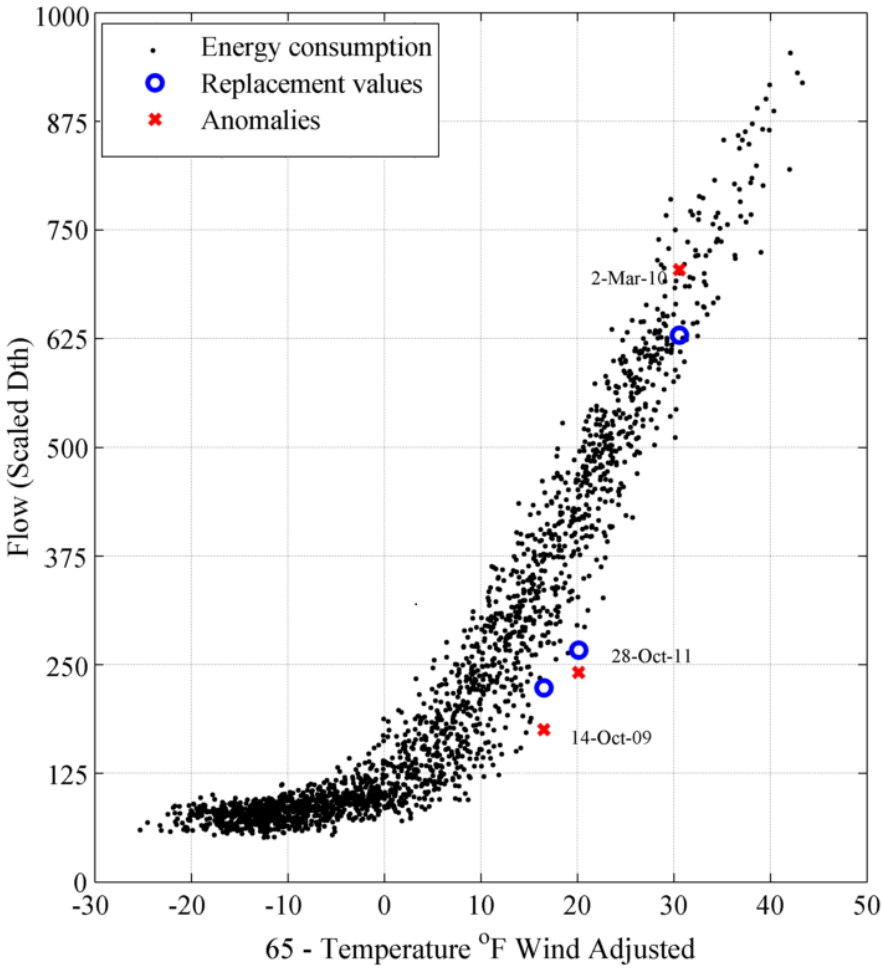

Figure 4.11: Scatter plot depicting all the anomalies identified and their corresponding replacement values

CHAPTER 5

EVALUATION AND ANALYSIS OF THE DATA CLEANING METHODS

This chapter presents the evaluation procedure for the data cleaning algorithm, the results on various data sets, and an analysis of the results. First, data sets and their pre-processing are described. Then, two sets of tests are performed; a simulation study and utilities' data study. The algorithm is evaluated on simulated data sets. The anomalies presented in Chapter 1 are inserted in the simulated data. The objective of the simulation study is to evaluate the false positive and false negative rates of the algorithm, because the anomalies are known in advance. The error between reported consumption and estimated replacement values can be calculated to determine how well the imputation model approximates the data. Various data sets from utilities are used in the second case study to test the algorithm. Both the original data sets (not cleaned) and the cleaned data sets are used to train a forecasting model. The forecasting models are evaluated on a testing set. The improvement in forecasting accuracy obtained by cleaning the data is the measure of effectiveness of the algorithm. To evaluate statistically the performance of the algorithm, a cross-validation study is conducted. This provides a statistical test of the difference in forecasting models trained on clean data versus those trained on original data. The simulation and utilities' data case studies are two different

ways of evaluating the performance of the algorithm. The chapter concludes with an analysis and discussion of the performance of the data cleaning algorithm.

5.1 Data Sets Description and Pre-processing

Data sets used in this chapter are confidential energy data obtained from Local Distribution Companies (LDC), representing the reported consumption of natural gas or electric energy on a given time period (daily or hourly). The data set used for the simulation study also is derived from the reported consumption of an utility.

The energy time series data sets are detrended before being used for anomaly detection. Energy time series data usually do not have the same trends for all years in the data set. Depending on the efficiency of energy systems, increased number of customers, or other reasons, the trends can be ascending or descending. This is a challenge for anomaly detection because the years with low trends might be mistaken for aggregated data with mismatched units. Also, the parameter estimation for a linear regression model might not provide good results if the slope of the trend is large. Therefore, the energy time series is detrended before being used for anomaly detection. However, the original data set is used for data replacement. There are various techniques that can be used to detrend energy time series data. The technique developed by Brown et al. is used in this dissertation to

detrend energy time series data before detecting anomalies [19]. The next section of this chapter presents the simulation study.

5.2 Simulation Study

The simulated data set presented in Figures 5.1 and 5.2 is derived from the daily reported natural gas consumption of operating area 7. The data set is from 01 November 2001 through 11 August 2014 for a total of 4,667 data points.

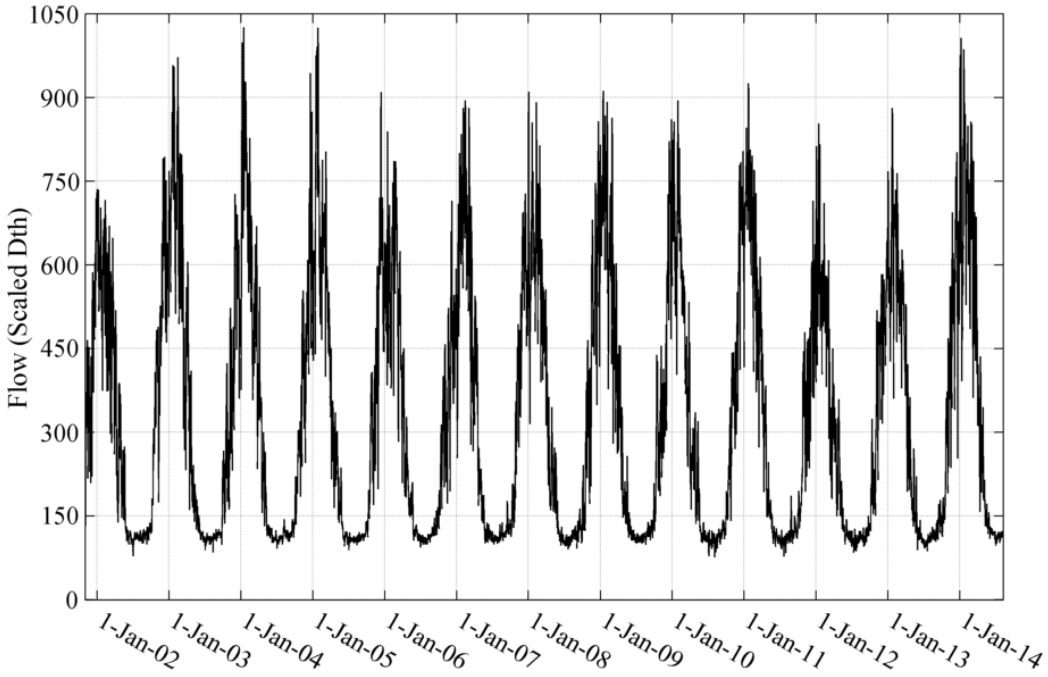

Figure 5.1: Time series plot of the simulated natural gas time series data set

No anomalies were found in the original reported natural gas consumption for operating area 7. Additionally, the data set is detrended to have approximately the same trends for all years in the data set. The detrended data set constitutes the

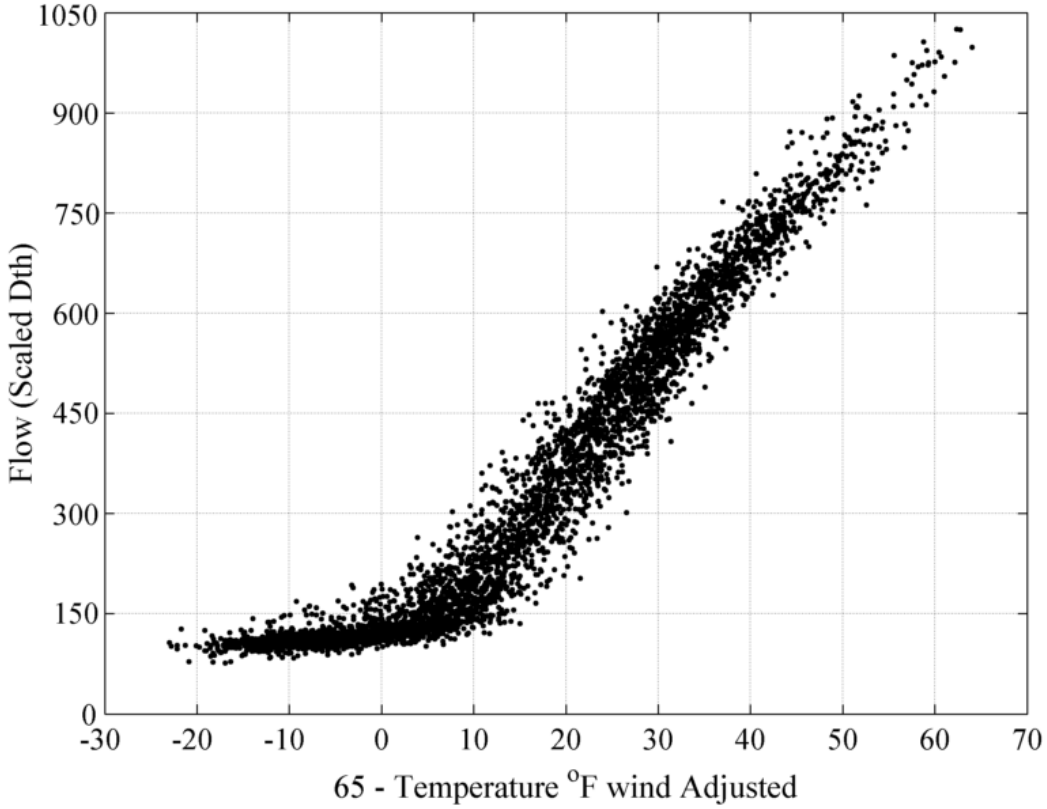

Figure 5.2: Scatter plot of the simulated natural gas time series data set

simulated data set. Different types of anomalies are inserted artificially in the data set, and the results of the linear regression data cleaning algorithm are presented for each case. The absolute percentage error (APE), relative to reported flow, is calculated between imputed and reported values in each case.

$$\text{APE} = \left| \frac{\text{Reported flow} - \text{Imputed flow}}{\text{Reported flow}} \right| \times 100. \tag{5.1}$$

An analysis of all the results is presented at the end of this section.

5.2.1 Missing Values

Eight missing flow values are inserted into the data set from 05 through 12 January 2014. Missing values are found in the pre-processing step of the algorithm (see Section 4.2). This case is studied to determine whether the algorithm returns false positives in the presence of missing values. The data cleaning algorithm identifies only the missing values as anomalies, and the results are presented in Figures 5.3 and 5.4. Figure 5.3 depicts the missing values in the time series plot and the replacement values calculated for all eight missing points.

Table 5.1: Imputation results for the simulation case of missing values

Date	Reported flow	Imputed flow	APE
05-01-2014	602.40	543.15	9.84
06-01-2014	733.05	708.82	3.30
07-01-2014	1006.35	939.03	6.69
08-01-2014	870.82	833.62	4.27
09-01-2014	754.35	716.17	5.06
10-01-2014	605.47	572.92	5.37
11-01-2014	392.02	358.51	8.55
12-01-2014	541.72	495.49	8.53

Table 5.1 presents reported and imputed flow values and the absolute percentage errors. Table 5.1 shows that the imputation model approximates the reported flow within a 10% error.

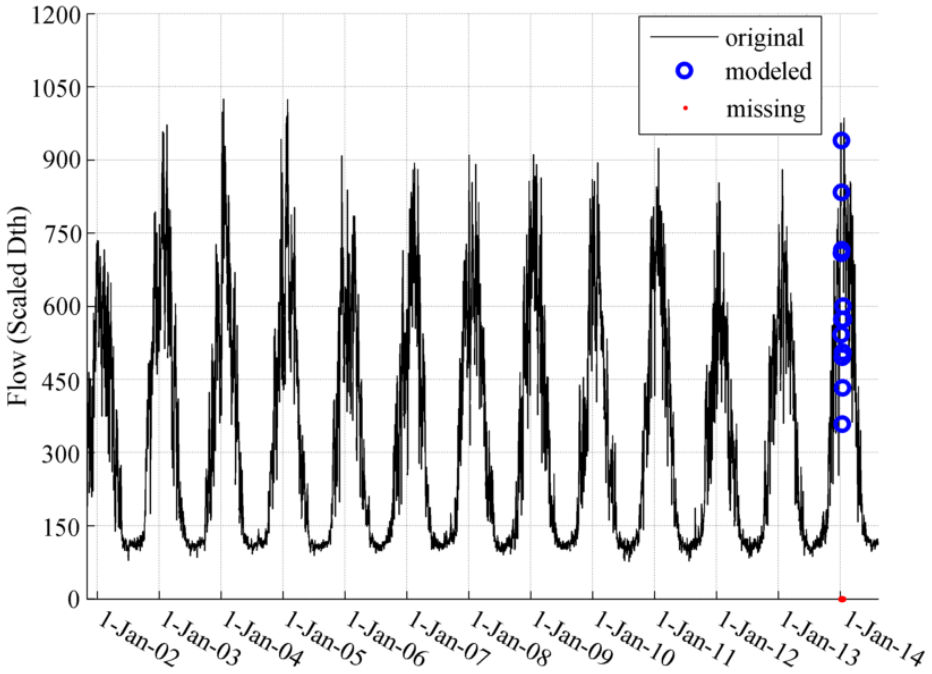

Figure 5.3: Time series plot of the data cleaning results for the simulation case of missing values

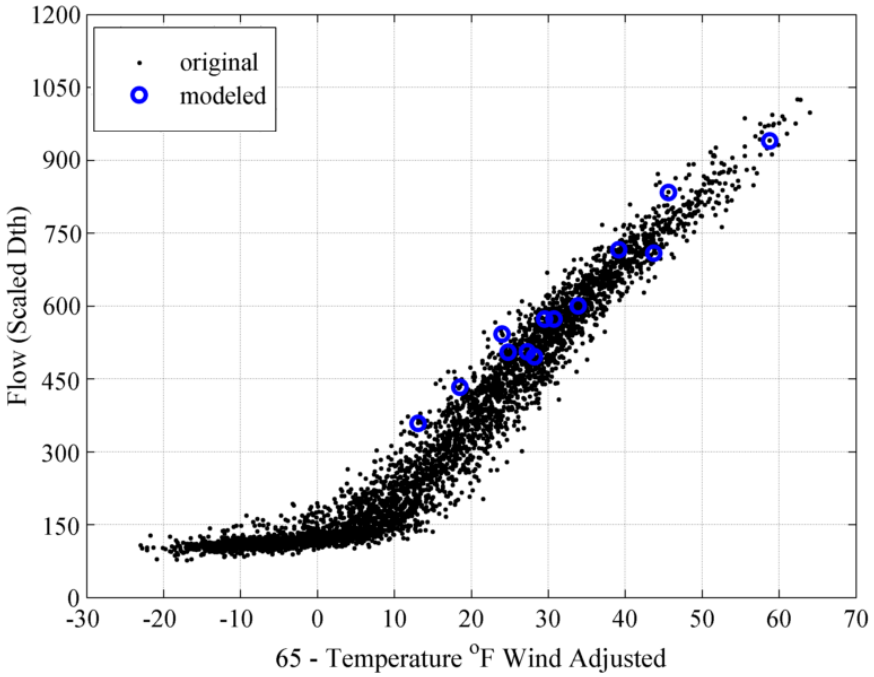

Figure 5.4: Scatter plot of the data cleaning results for the simulation case of missing values

5.2.2 Extremely High Flow Values or Main Breaks

Two extremely high flow values are inserted into the data set to simulate main breaks. The first value is inserted on 15 January 2004, which is in the winter season. The second value is inserted on 03 August 2005, which is in the summer season. The data cleaning algorithm results, presented in Figures 5.5 and 5.6, depict the two anomalies found and their replacement values.

Table 5.2 presents the imputation results for this case and shows that the imputed values are nearly the same as the reported flow values. There is a maximum absolute percentage error of 1% between reported and imputed values in the case of imputing extremely high flow values.

Table 5.2: Imputation results for the simulation case of extremely high flow values

Date	Reported flow	Anomalies	Imputed flow	APE
15-01-2004	1025.17	2525.18	1025.71	0.05
03-08-2005	103.56	1035.45	104.54	0.94

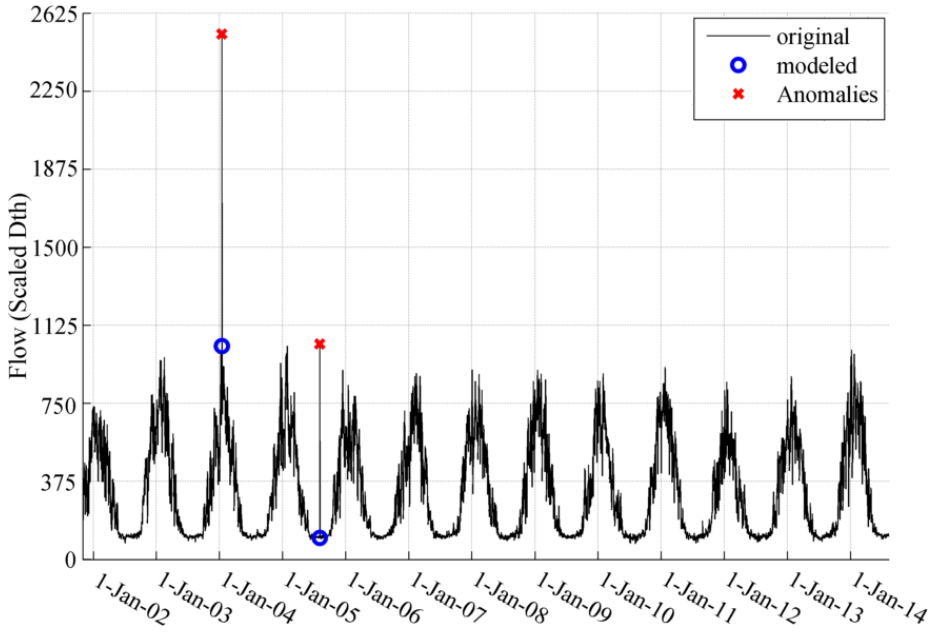

Figure 5.5: Time series plot of the data cleaning results for the simulation case of extremely high flow values

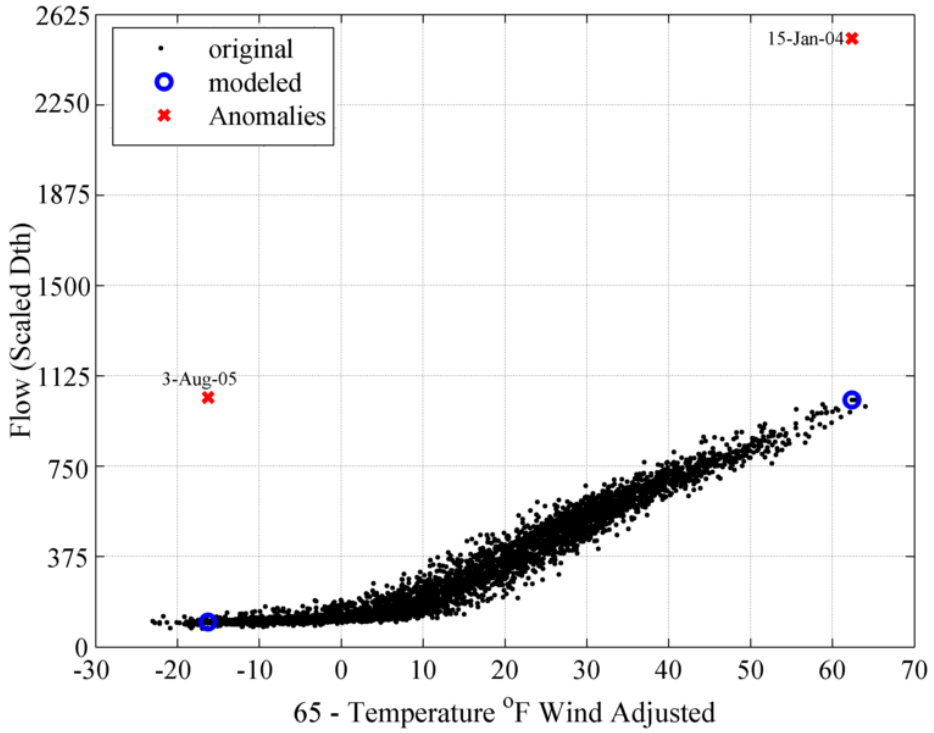

Figure 5.6: Scatter plot of the data cleaning results for the simulation case of extremely high flow values

5.2.3 Negative Flow Values

Seven negative flow values are inserted into the data set, from 24 through 27 February 2009 and 17 through 19 June 2009. The data cleaning algorithm results, presented in Figures 5.7 and 5.8, depict the seven anomalies found and their replacement values.

Table 5.3 presents the imputation results and shows that the maximum absolute percentage error between imputed and reported flow values is about 2% in the case of negative flow values. While any energy domain expert recognizes negative flow values as anomalous, the point is that the data cleaning algorithm can match the domain expert and therefore saves valuable time. In addition, the algorithm provides good imputation values.

Table 5.3: Imputation results for the simulation case of negative flow values

Date	Reported flow	Anomalies	Imputed flow	APE
24-02-2009	688.65	-344.33	692.55	0.57
25-02-2009	591.91	-295.96	600.38	1.43
26-02-2009	547.42	-273.68	547.57	0.03
27-02-2009	421.95	-210.97	426.68	1.12
17-06-2009	120.53	-120.53	121.51	0.81
18-06-2009	124.80	-124.80	123.45	1.08
19-06-2009	111.34	-111.34	113.49	1.93

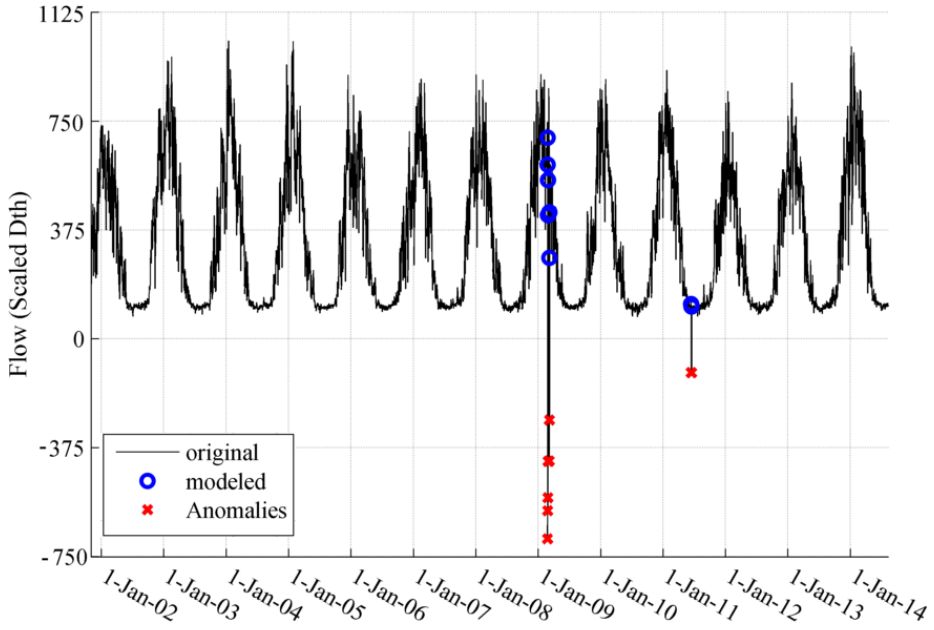

Figure 5.7: Time series plot of the data cleaning results for the simulation case of negative flow values

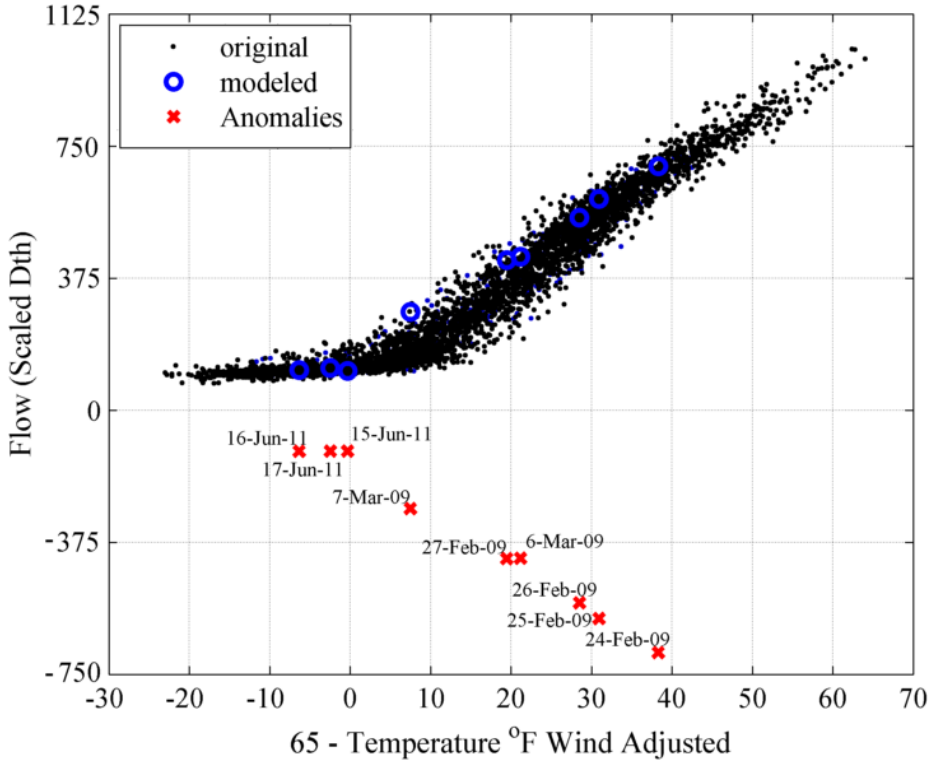

Figure 5.8: Scatter plot of the data cleaning results for the simulation case of negative flow values

5.2.4 Naïve Disaggregation or Stuck Meter

Five values are artificially inserted into the data set to represent a stuck meter. The flow value for 20 April 2004 is used as the fixed value for all days from 21 through 25 April 2004. In this case, the data cleaning algorithm identifies four anomalies out of the five inserted in the data set. The maximum absolute percentage error between reported values for 20 and 21 April 2004 is 5.6%. Therefore, it is expected that the reported value for 21 April 2004 is not considered anomalous.

The data cleaning algorithm results, presented in Figures 5.9 and 5.10, depict the four anomalies found and their replacement values. Table 5.4 also shows the imputation results for the case of stuck meter. Looking at Table 5.4, the maximum absolute percentage error between reported and imputed values is about 4.5%.

Table 5.4: Imputation results for the simulation case of a stuck meter

Date	Reported flow	Anomalies	Imputed flow	APE
20-04-2004	160.43	160.43	160.43	–
21-04-2004	169.95	160.43	160.43	5.60
22-04-2004	312.31	160.43	326.11	4.42
23-04-2004	326.85	160.43	324.92	0.59
24-04-2004	305.10	160.43	291.45	4.47
25-04-2004	307.06	160.43	309.82	0.90

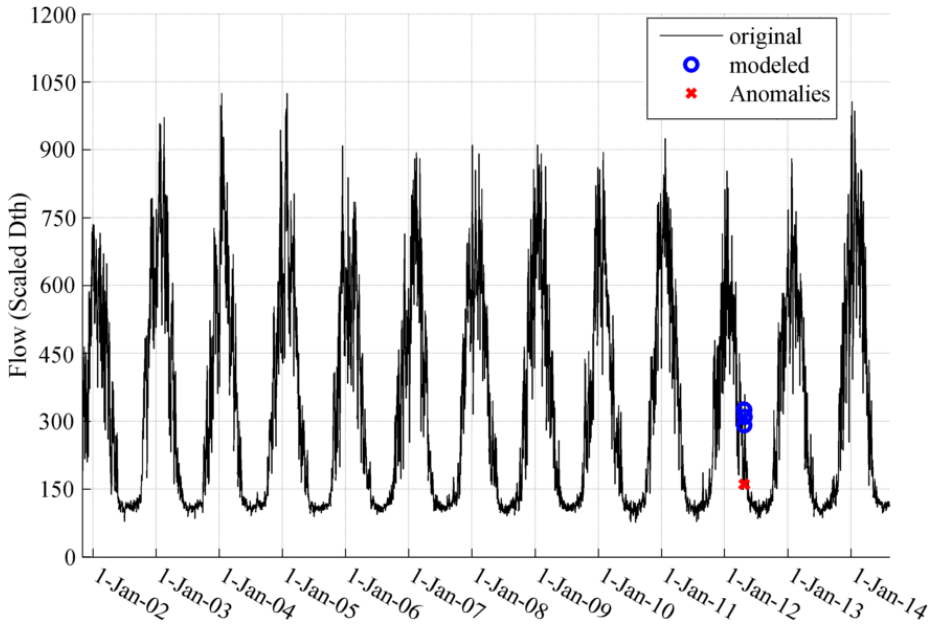

Figure 5.9: Time series plot of the data cleaning results for the simulation case of a stuck meter

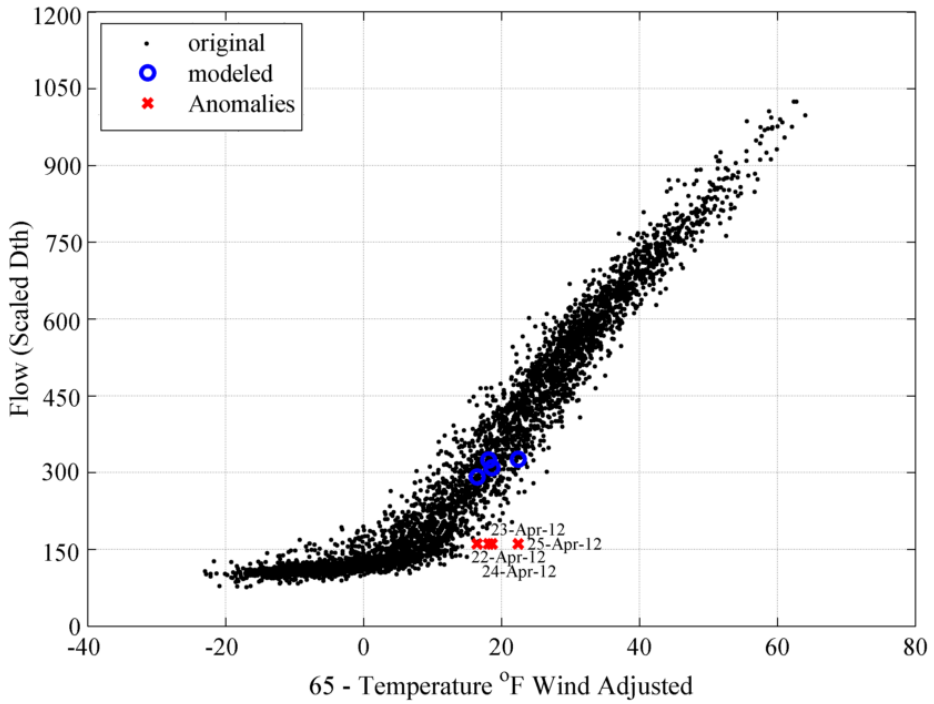

Figure 5.10: Scatter plot of the data cleaning results for the simulation case of a stuck meter

5.2.5 Power Generation Load

Eight power generation simulated flow values are inserted in the summer season, from 01 through 08 August 2003, to avoid confusion with the case of extremely high flow values. The data cleaning algorithm results, presented in Figures 5.11 and 5.12, depict the eight anomalies found and their replacement values.

Table 5.5 presents the imputation results in the case of power generation load and shows that the maximum absolute percentage error between imputed and reported flow values is about 3.5%.

Table 5.5: Imputation results for the simulation case of power generation load

Date	Reported flow	Anomalies	Imputed flow	APE
01-08-2003	112.80	225.52	109.65	2.79
02-08-2003	100.21	300.53	98.64	1.57
03-08-2003	107.17	321.60	106.42	0.70
04-08-2003	109.43	328.35	112.05	2.39
05-08-2003	108.08	324.22	108.98	0.83
06-08-2003	108.15	324.45	108.90	0.69
07-08-2003	111.97	336.00	113.47	1.34
08-08-2003	111.38	237.75	107.40	3.57

The next section summarizes the results obtained for all types of anomalies. An analysis of the results and the various imputation absolute percentage errors also is made.

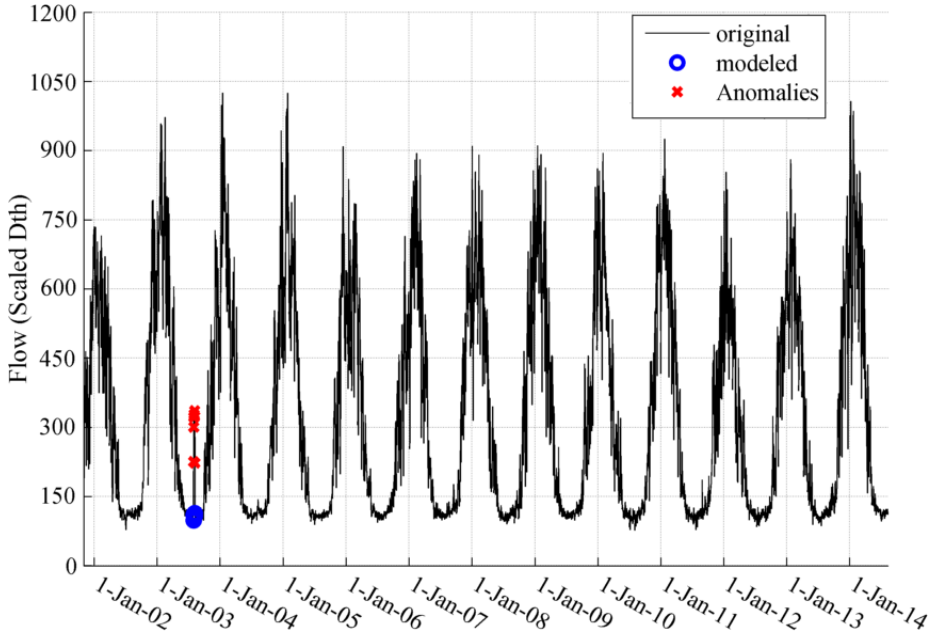

Figure 5.11: Time series plot of the data cleaning results for the simulation case of power generation load

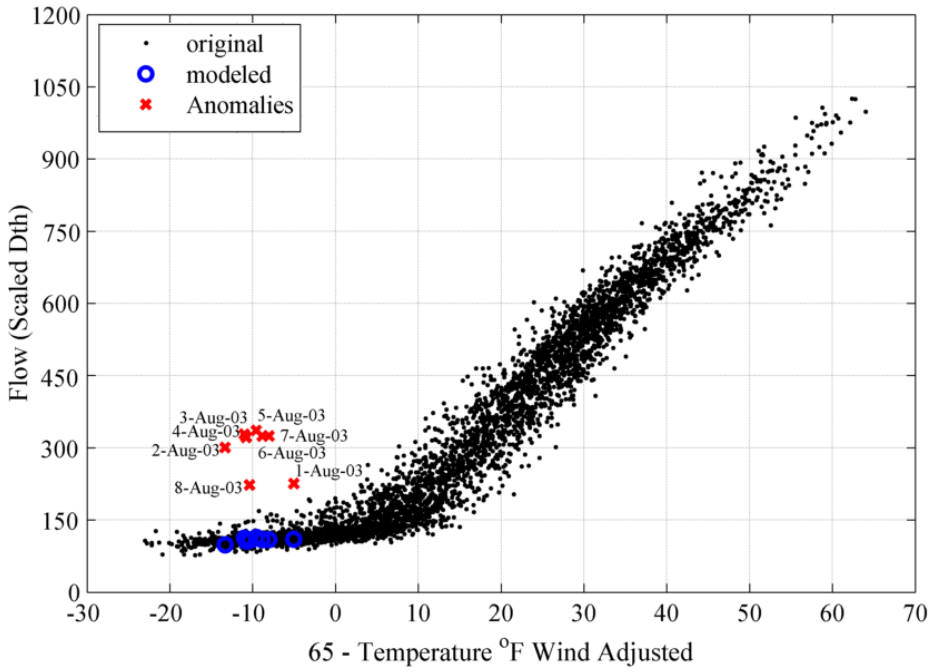

Figure 5.12: Scatter plot of the data cleaning results for the simulation case of power generation load

5.2.6 Simulation Study Analysis

Anomalies of various types were inserted artificially in various years and weather seasons of the simulated data set. The data cleaning algorithm detects all points that deviate considerably (the deviation is based on weather dependency and one day ago flow) from the remaining points in the data set.

The imputation model has an accuracy of about 5%, except in the case of missing data, where the accuracy is about 10%. The missing data case provides the largest error because its detection process is not iterative. In the case of missing data, all anomalies are detected at once in the beginning of the process and imputed because the linear data cleaning algorithm is autoregressive.

Overall, the anomaly detection provides expected results, and the imputation model provides good replacement values to the anomalous energy time series values. The next section of this chapter presents the evaluation of the data cleaning algorithm on data sets from local distribution companies. The improvement in forecasting accuracy obtained by cleaning the data is the measure of effectiveness of the algorithm because "actual" values do not exist.

5.3 Utilities Data Testing

The data cleaning algorithm is applied to reported electric and natural gas consumption data sets. The reported consumption is called the original data set,

while the algorithm output is the clean data set. The clean and original data sets are divided into training and testing sets. The training sets are used to train the same forecasting model derived from Vitullo et al. [98]. The models are used to calculate estimated values for the test sets and to compute out-of-sample errors. The root mean squared error (RMSE) and the mean absolute percentage error (MAPE) are both used as error measures in this dissertation. The RMSE is a unit measure that estimates the difference between observed and estimated values [45].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\text{Reported flow} - \text{Imputed flow})_i^2}{N}}. \tag{5.2}$$

The MAPE is a unitless measure that calculates the error as a percentage of the actual values [7].

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\text{Reported flow} - \text{Imputed flow}}{\text{Reported flow}}\right|_i \times 100. \tag{5.3}$$

For both RMSE and MAPE equations, N is the number of observations. Both error measures are two different ways of interpreting and analyzing the results. The errors calculated on the test set with models trained on both the original and clean data sets are compared to analyze the forecasting accuracy improvement obtained by cleaning the data. The original, anomalies, and clean data sets are presented here,

along with forecasting errors (RMSE and MAPE). An analysis of the error measures concludes this section.

The average RMSE and MAPE are calculated, along with the errors by month and by unusual day. "Unusual Day" is a term used to represent a day on which an unusual weather event occured [75]. These unusual events are weather-related events such as a sudden temperature increase or decrease, high or low humidity, or extreme cold. The unusual days encountered in this dissertation are coldest days, colder and warmer than normal days, windiest heating days, colder and warmer today than yesterday, the first cold and warm days, high and low humidity heating days, and sunny and cloudy heating days. An example of unusual days depicted on a natural gas data set is presented in Figure 5.13.

Three examples are presented in this section. The data cleaning algorithm is tested on two natural gas data sets and on one electric data set. In all the examples presented in this dissertation, the last year of data is the test set, while the previous years constitute the training set. The results graphs show the time series plot of energy consumption over time and the scatter plot which shows the energy consumption versus temperature, with the anomalies and replacement values depicted on both plots. The imputation results also are presented here. In the imputation tables, the anomalous values are presented along with their replacement values.

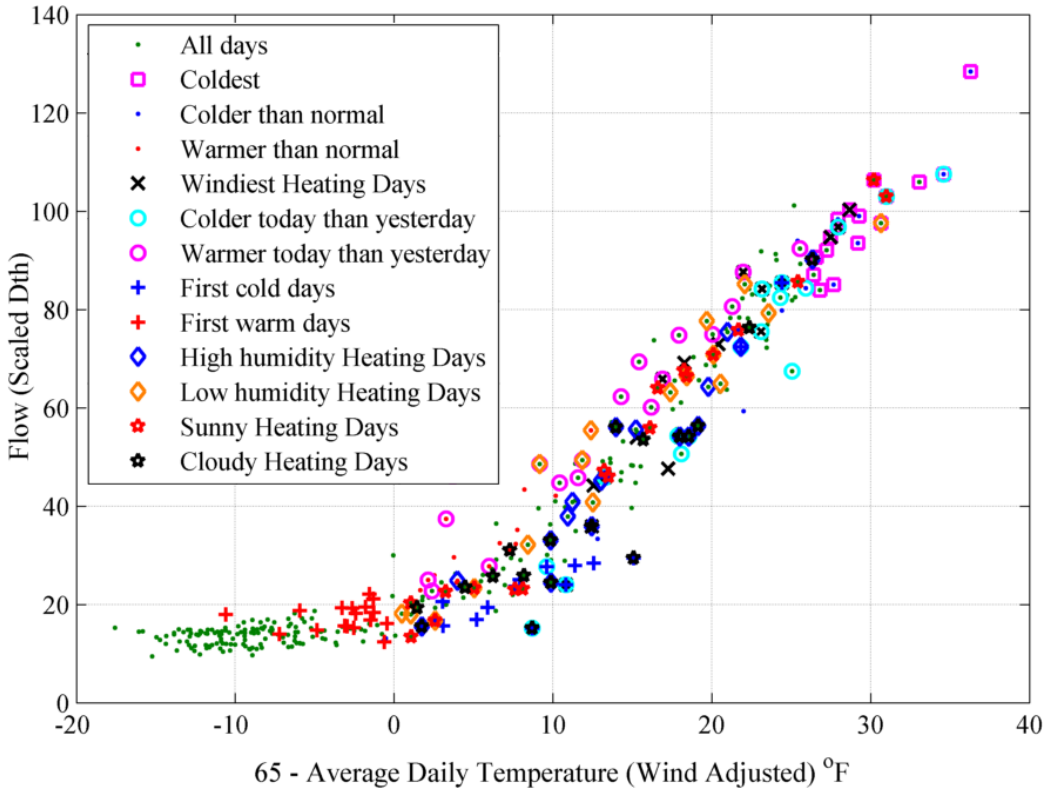

Figure 5.13: Example of unusual days for a natural gas data set

5.3.1 Example 1: Natural Gas Data Set of Operating Area 8

The first example is the reported natural gas consumption of operating area 8. The data set is from 01 May 2004 to 31 July 2012. The data cleaning algorithm results are presented in Figures 5.14 and 5.15. Additionally, the imputation results for the data set of operating area 8 are presented in Table 5.6.

The training set is data from 01 May 2004 through 31 July 2011. The RMSE and MAPE calculated on the test set from 01 August 2011 through 31 July 2012 by month and by unusual day are presented in Figures 5.16 and 5.17, respectively.

The RMSE and MAPE calculated with forecasting models trained on the

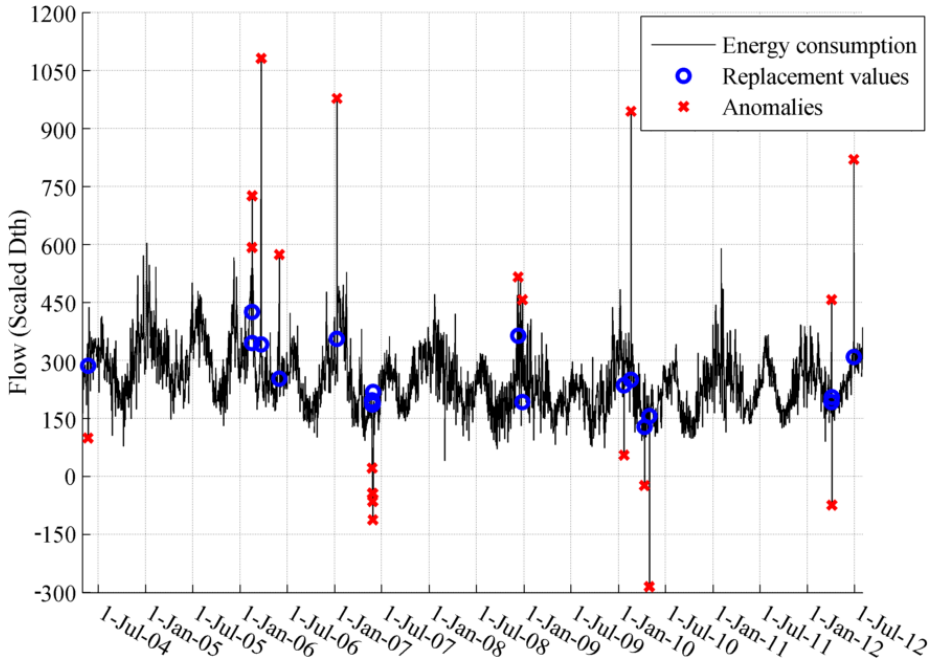

Figure 5.14: Time series plot of the data cleaning results for the natural gas data set of operating area 8

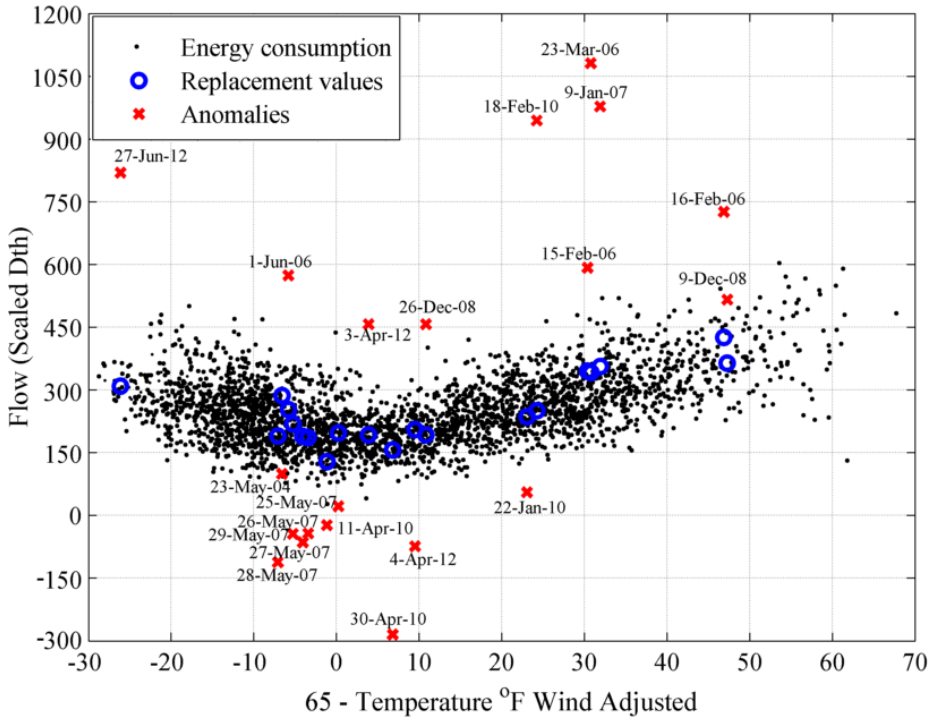

Figure 5.15: Scatter plot of the data cleaning results for the natural gas data set of operating area 8

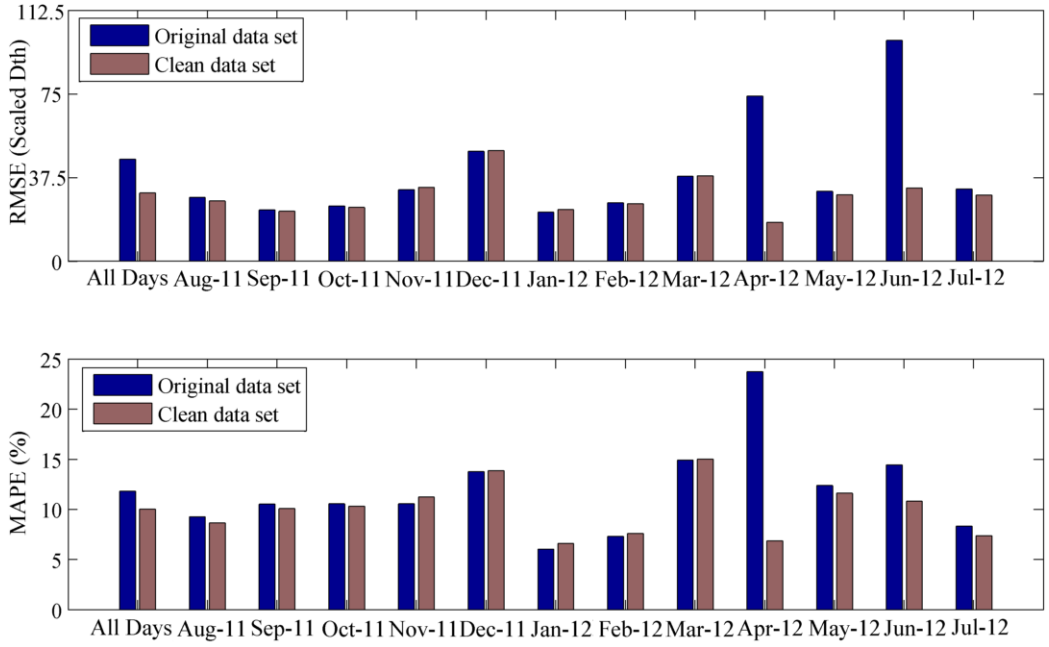

Figure 5.16: RMSE and MAPE by month for the original and clean data sets of operating area 8

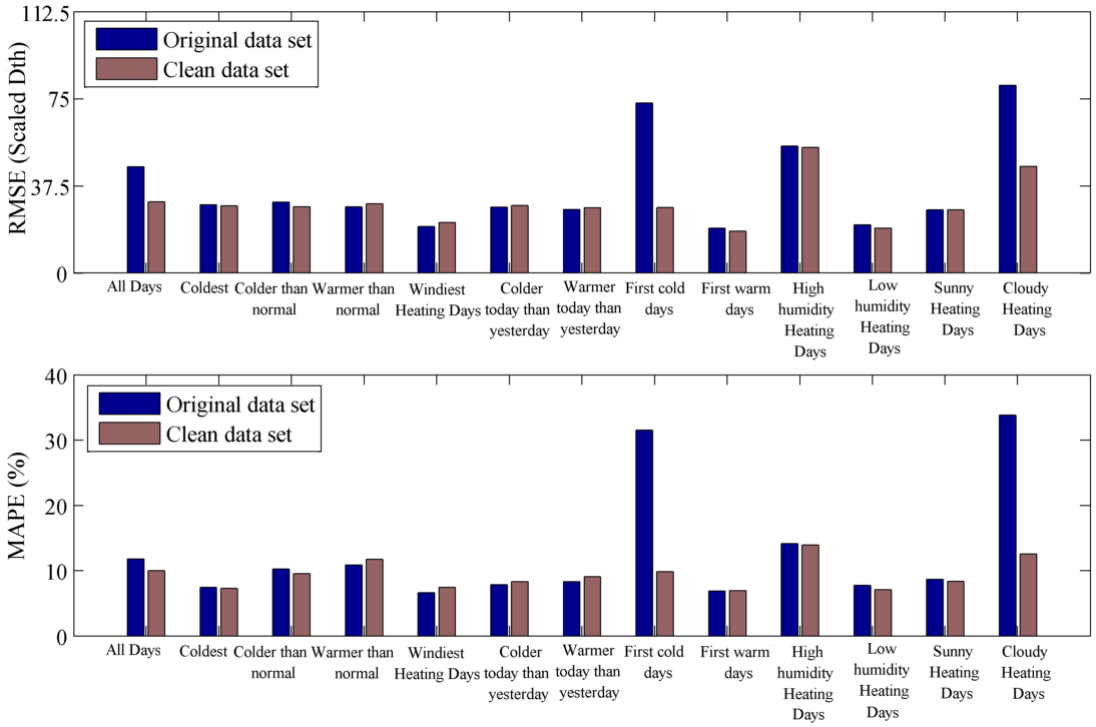

Figure 5.17: RMSE and MAPE by unusual day for the original and clean data sets of operating area 8

Table 5.6: Imputation results for the natural gas data set of operating area 8

Date	Reported flow	Imputed flow
23-05-2004	99.45	286.73
15-02-2006	592.12	344.92
16-02-2006	725.77	425.70
23-03-2006	1080.90	341.70
01-06-2006	573.83	253.43
09-01-2007	977.85	356.25
25-05-2007	21.52	196.95
26-05-2007	-43.28	186.08
27-05-2007	-64.50	188.62
28-05-2007	-112.20	189.98
29-05-2007	-44.40	218.61
09-12-2008	515.56	364.56
26-12-2008	457.21	193.05
22-01-2010	55.20	235.50
18-02-2010	944.40	250.80
11-04-2010	-23.32	128.49
30-04-2010	-284.85	156.45
03-04-2012	456.90	192.68
04-04-2012	-74.04	204.82
27-06-2012	819.60	309.37

clean data set are on average smaller than the error calculated on the original test set (about 33% in RMSE). The largest observed improvements are 76% in RMSE for April 2012, and 66% in RMSE for June 2012. All other months' performances are about the same. Looking at the RMSE and MAPE by unusual day calculated with models trained on both the original and clean data set, the largest improvements are observed for the first cold days (61.5% in RMSE and 20% in MAPE) and the cloudy heating days (43% in RMSE and 20% in MAPE).

5.3.2 Example 2: Natural Gas Data Set of Operating Area 9

The second example is the reported natural gas consumption of operating area 9. The data set is from 01 March 2003 through 30 November 2014. The data cleaning algorithm results are presented in Figure 5.18, Figure 5.19, and Table 5.7.

Table 5.7: Imputation results for the natural gas data set of operating area 9

Date	Reported flow	Imputed flow
03-08-2012	19.53	49.55
19-02-2014	540.38	440.30
22-02-2014	260.32	418.48
23-02-2014	286.07	514.00
30-07-2014	13.52	61.95
04-08-2014	17.17	54.66
18-09-2014	68.15	118.92

The training set is data from 01 March 2003 through 30 November 2013. The RMSE and MAPE calculated on the test set from 01 December 2013 through 30 November 2014 by month and by unusual day are presented in Figures 5.20 and 5.21, respectively.

The RMSE and MAPE by month and by unusual day are lower for the clean data set than for the original data set. The average observed improvement in RMSE is 37%. The forecasting improvement, obtained with models trained on clean data, are observed for February, July, August, and September 2014, with the largest percentage improvement of 77% in RMSE found in February 2014. No cloudy

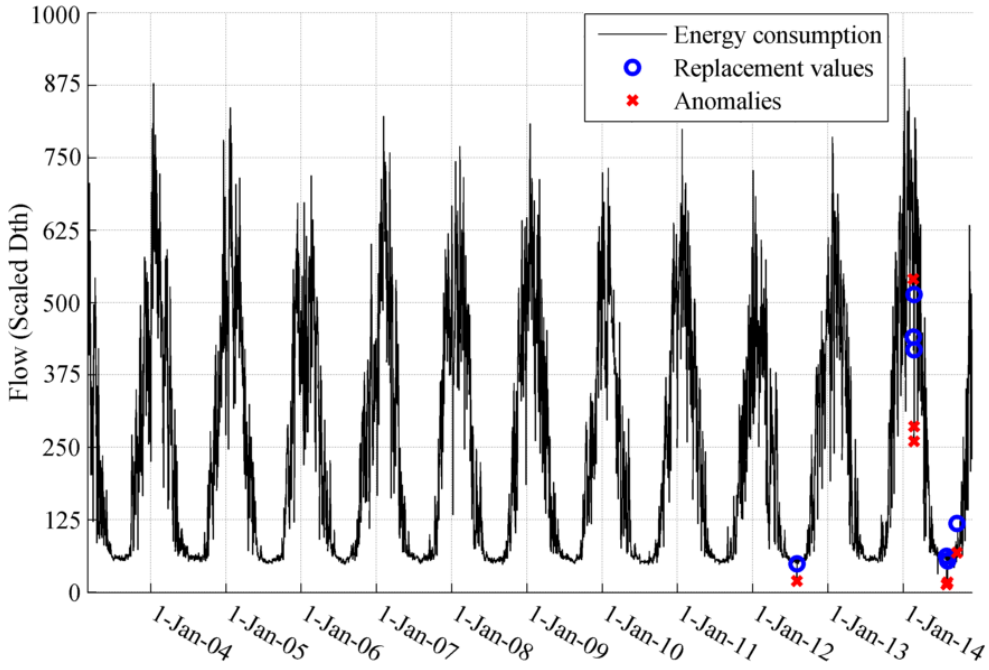

Figure 5.18: Time series plot of the data cleaning results for the natural gas data set of operating area 9

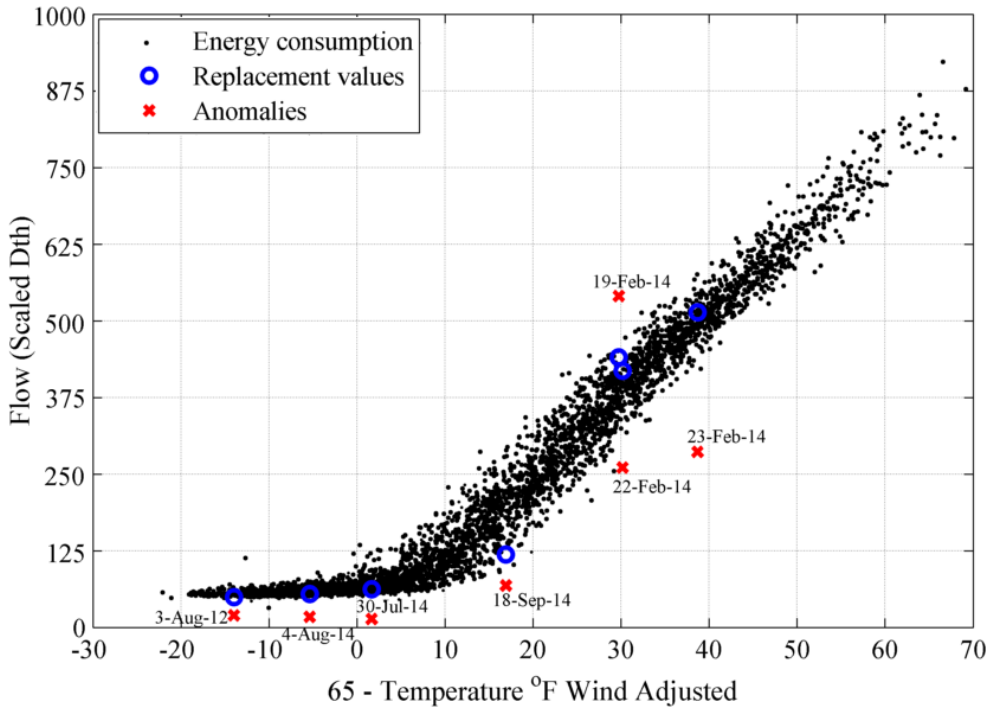

Figure 5.19: Scatter plot of the data cleaning results for the natural gas data set of operating area 9

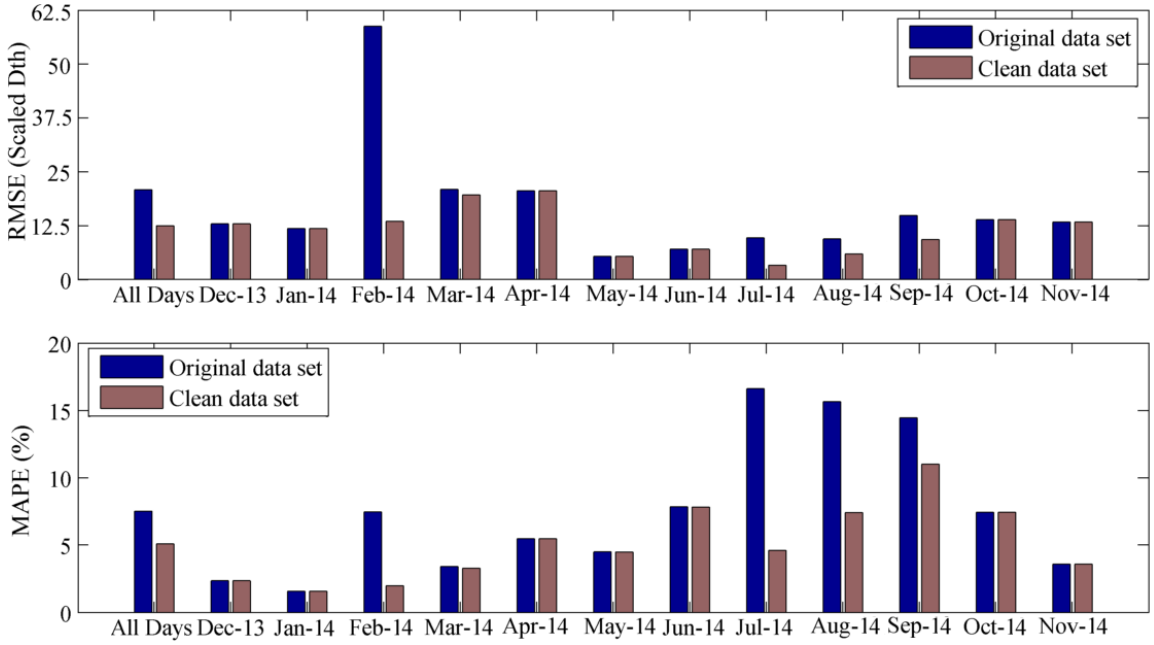

Figure 5.20: RMSE and MAPE by month for the original and clean data sets of operating area 9

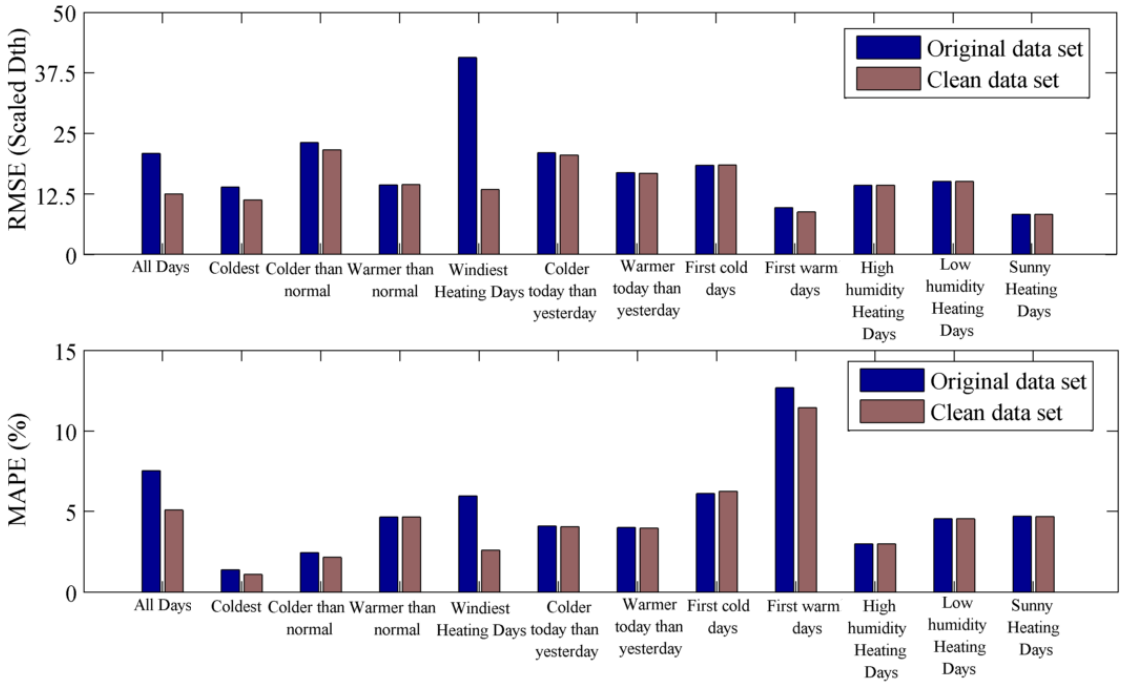

Figure 5.21: RMSE and MAPE by unusual day for the original and clean data sets of operating area 9

heating days were found in the weather of operating area 9. An improvement of 70% in RMSE is observed also for the windiest heating days.

5.3.3 Example 3: Electric Data Set of Operating Area 10

Example 3 is the reported electric consumption for operation area 10. The data set is from 01 February 2004 to 31 July 2013. The data cleaning algorithm results are presented in Figure 5.22, Figure 5.23, and Table 5.8.

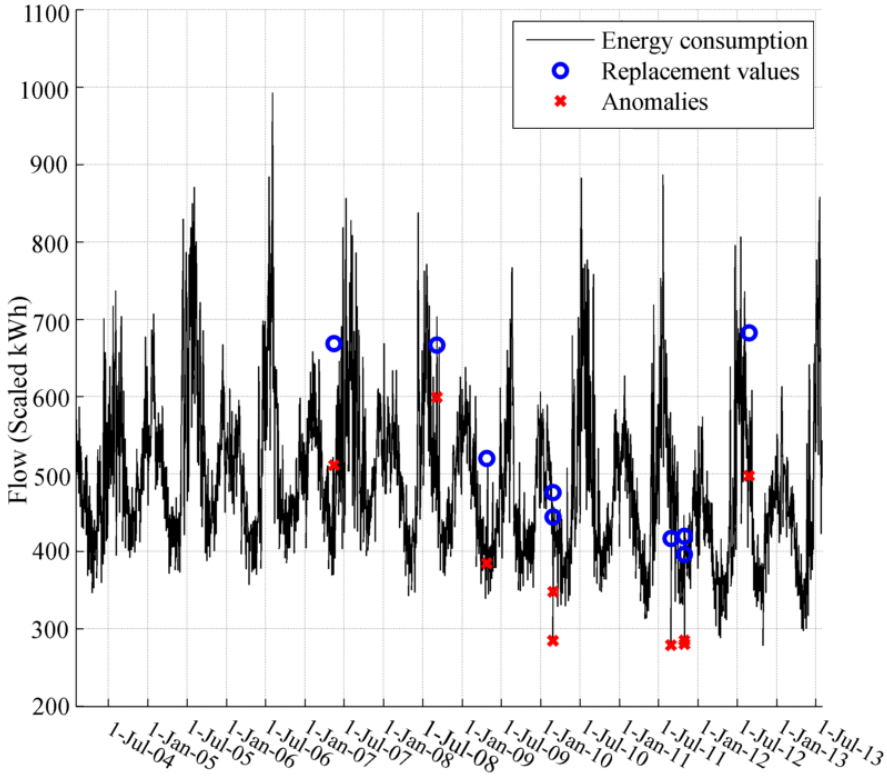

Figure 5.22: Time series plot of the data cleaning results for the electric data set of operating area 10

The training set is data from 01 February 2004 through 31 July 2012. The test set is from 01 August 2013 through 31 July 2013. The same forecasting model

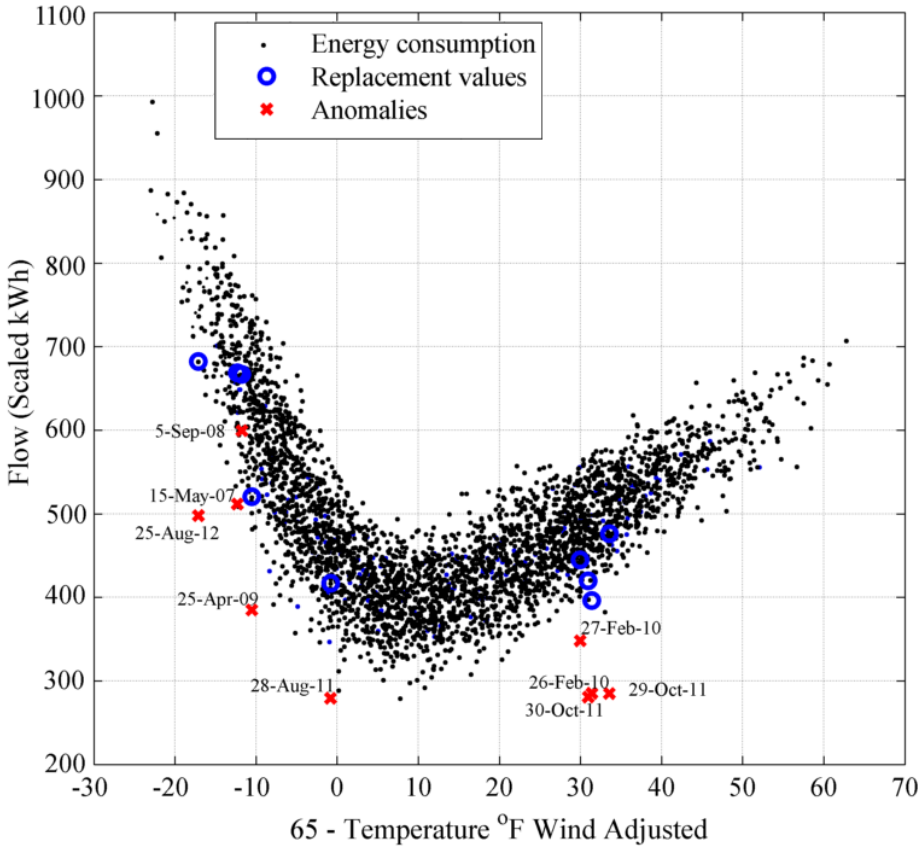

Figure 5.23: Scatter plot of the data cleaning results for the electric data set of operating area 10

used to calculate replacement values and error measures for natural gas data sets is applied to this example. The RMSE and MAPE calculated on the test set by month and by unusual day are presented in Figures 5.24 and 5.25, respectively.

The RMSE and MAPE are only improved for the months of August and September 2012. The forecasting errors for all other months are about the same. However, the anomalies were found in February, May, August, September, and October. Since one of the primary uses of electric load is cooling, the improvement of 28% found in August 2012 is practically significant. The RMSE and MAPE by

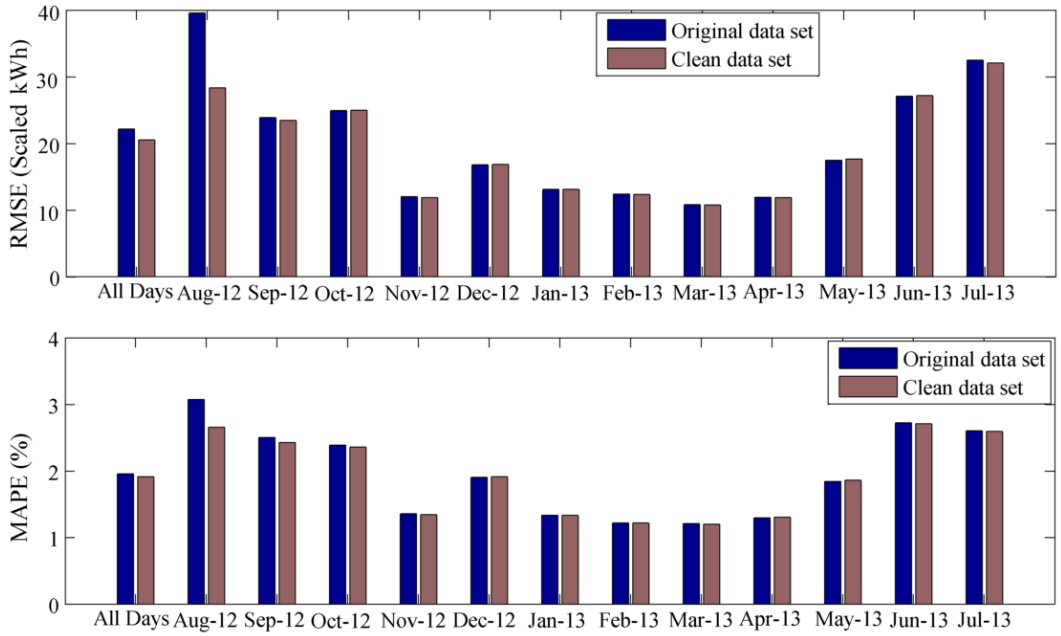

Figure 5.24: RMSE and MAPE by month for the original and clean data sets of operating area 10

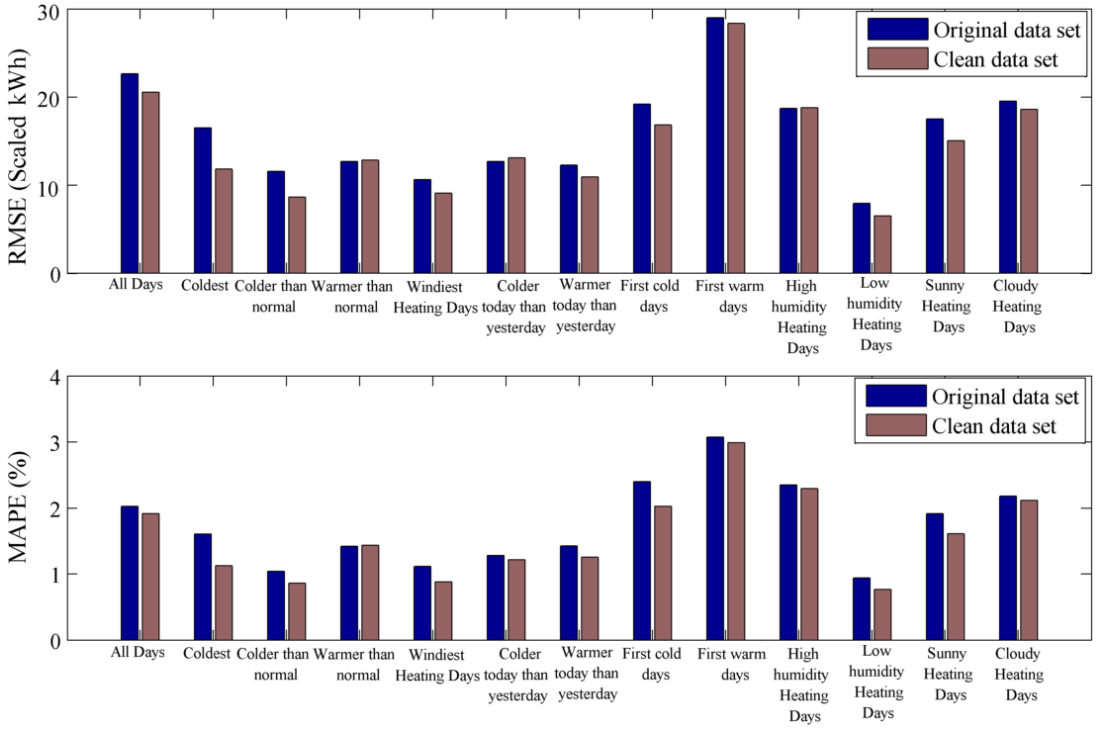

Figure 5.25: RMSE and MAPE by unusual day for the original and clean data sets of operating area 10

Table 5.8: Imputation results for the electric data set of operating area 10

Date	Reported flow	Imputed flow
15-05-2007	511.34	668.31
05-09-2008	598.82	666.33
25-04-2009	384.36	520.32
26-02-2010	284.38	476.35
27-02-2010	347.87	444.50
28-08-2011	278.88	416.48
29-10-2011	284.83	396.36
30-10-2011	280.40	419.35
25-08-2012	497.34	681.80

unusual day are lower for the clean data set than for the original data set for both warmer and colder days. The average observed RMSE improvement is 7%. There is no improvement found for high humidity heating days and warmer than normal days. Note that the imputation model used to calculate replacement values in this case is the same model used for natural gas data sets, and it yields good results.

5.3.4 Utilities Data Testing Analysis

In this section, the data cleaning algorithm is tested on both natural gas and electric data sets. For the natural gas data sets, operating area 8 is a region in the southern part of the United States, while operating area 9 is a region in the northern part of the United States. Therefore, they both experience different climates. The principal use of natural gas and electric energy by residential, commercial, and industrial customers is heating and cooling. Therefore, the improvement of forecasting errors in winter months for the natural gas data sets and

in summer months for the electric data sets are valuable. The energy data sets have many similarities because the electric data anomalies were imputed by replacement values calculated using forecasting models intended for natural gas, and provide a maximum observed RMSE improvement of 28%. The RMSE and MAPE presented in this section are only single point values calculated on the last year of the test set.

To verify the performance of the data cleaning algorithm throughout the data sets and also on smaller data sets, a cross-validation scheme is used. The cross-validation yields a set of results that can be used to find the mean of the error and also to test the statistical significance of the improvement in forecasting accuracy. The next section of this chapter explains the cross-validation scheme and presents the RMSE results obtained for operating area 8 and 9. The cross-validation scheme is not applied to the electric data set of operation area 10 because of the lack of good electric forecasting models.

5.4 Cross-validation

The goal of data cleaning is to improve energy demand forecasting accuracy. Therefore, the improvement in forecasting accuracy by cleaning the data is the measure of effectiveness of the algorithm. Cross-validation is used to verify the accuracy and consistency of the results. A cross-validation or random rotation is a validation technique for assessing how the results of a statistical analysis generalize

to an independent data set [54]. The objective is to show that if the results obtained on various subsets are consistent and statistically significant, a conclusion can be made about the performance of the data cleaning algorithm in general. Also, the cross-validation demonstrates that the accuracy of the results does not depend on the choice, type, or length of the data set.

5.4.1 Cross-validation Scheme

An example of cross-validation diagram for the natural gas data set of operating area 8, composed of eight years of data from 01 May 2004 to 31 July 2012, is depicted in Figure 5.26. The data is divided into subsets starting with a minimum

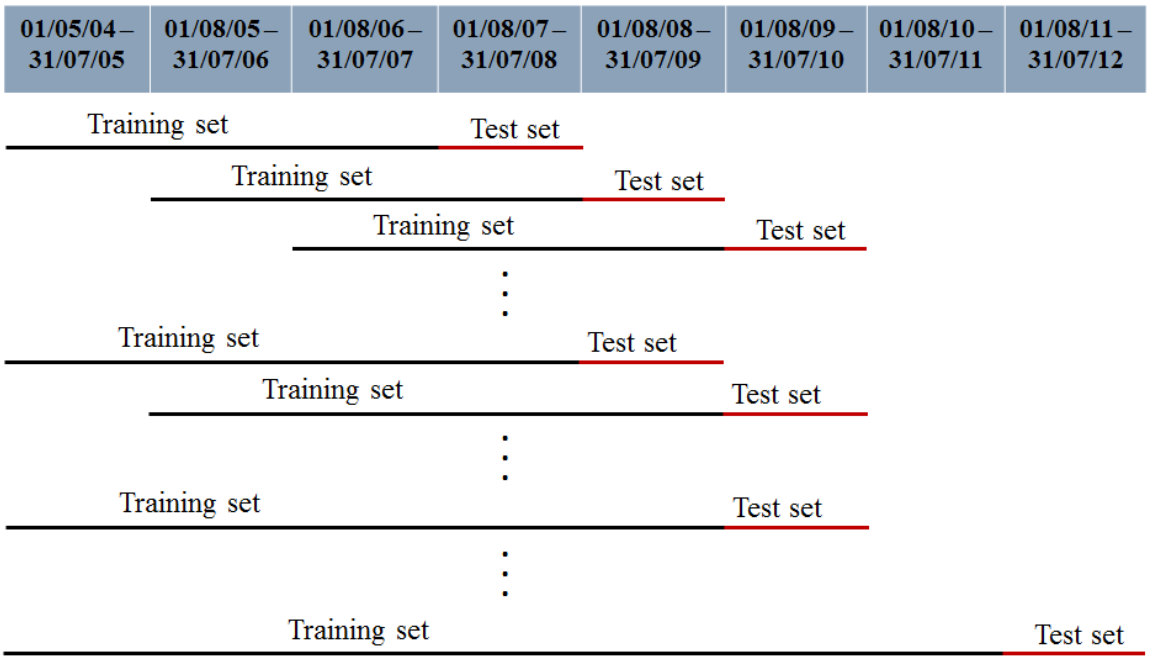

Figure 5.26: Example of cross-validation scheme for the data set of operating area 8

of three years for the training set. The minimum number of years in the training set is three because the training set should be large enough to be able to train the forecasting models without memorization. Combinations are built using the subsets. Each combination is divided into a training set used to train a forecasting model, and a test set used to calculate out-of-sample errors.

Table 5.9: Subdivision of the data set of operating area 8

Naming convention	Subset date range
S1	01-05-2004 to 31-07-2005
S2	01-08-2005 to 31-07-2006
S3	01-08-2006 to 31-07-2007
S4	01-08-2007 to 31-07-2008
S5	01-08-2008 to 31-07-2009
S6	01-08-2009 to 31-07-2010
S7	01-08-2010 to 31-07-2011
S8	01-08-2011 to 31-07-2012

Table 5.9 and Table 5.10 present the data subdivision into subsets and the combination of those subsets into crosses. The test set length is one year. The years in the data set must be consecutive because of the nature of the forecasting models used.

The cross-validation scheme for operating area 9 is presented also in Tables 5.11 and 5.12. The cross-validation results for operating areas 8 and 9 are presented in the next section.

Table 5.10: Cross-validation table for the data set of operating area 8

Number of subsets used for the training set	Cross number	Training set	Testing set
Training set composed of 3 subsets	1	S1 S2 S3	S4
	2	S2 S3 S4	S5
	3	S3 S4 S5	S6
	4	S4 S5 S6	S7
	5	S5 S6 S7	S8
4 subsets	6	S1 S2 S3 S4	S5
	7	S2 S3 S4 S5	S6
	8	S3 S4 S5 S6	S7
	9	S4 S5 S6 S7	S8
5 subsets	10	S1 S2 S3 S4 S5	S6
	11	S2 S3 S4 S5 S6	S7
	12	S3 S4 S5 S6 S7	S8
6 subsets	13	S1 S2 S3 S4 S5 S6	S7
	14	S2 S3 S4 S5 S6 S7	S8
7 subsets	15	S1 S2 S3 S4 S5 S6 S7	S8

Table 5.11: Subdivision of the data set of operating area 9

Naming convention	Subset date range
S1	01-03-2003 to 30-11-2004
S2	01-12-2004 to 30-11-2005
S3	01-12-2005 to 30-11-2006
S4	01-12-2006 to 30-11-2007
S5	01-12-2007 to 30-11-2008
S6	01-12-2008 to 30-11-2009
S7	01-12-2009 to 30-11-2010
S8	01-12-2010 to 30-11-2011
S9	01-12-2011 to 30-11-2012
S10	01-12-2012 to 30-11-2013
S11	01-12-2013 to 30-11-2014

5.4.2 Cross-validation Results

The cross-validation results for the natural gas data set of operating area 8 and 9 are presented in Table 5.13 and Table 5.16, respectively. A dependent samples t test

Table 5.12: Cross-validation table for the data set of operating area 9

Number of subsets used for the training set	Cross number	Training set	Testing set
Training set composed of 3 subsets	1	S1 S2 S3	S4
	2	S2 S3 S4	S5
	3	S3 S4 S5	S6
	4	S4 S5 S6	S7
	5	S5 S6 S7	S8
	6	S6 S7 S8	S9
	7	S7 S8 S9	S10
	8	S8 S9 S10	S11
4 subsets	9	S1 S2 S3 S4	S5
	10	S2 S3 S4 S5	S6
	11	S3 S4 S5 S6	S7
	12	S4 S5 S6 S7	S8
	13	S5 S6 S7 S8	S9
	14	S6 S7 S8 S9	S10
	15	S7 S8 S9 S10	S11
5 subsets	16	S1 S2 S3 S4 S5	S6
	17	S2 S3 S4 S5 S6	S7
	18	S3 S4 S5 S6 S7	S8
	19	S4 S5 S6 S7 S8	S9
	20	S5 S6 S7 S8 S9	S10
	21	S6 S7 S8 S9 S10	S11
6 subsets	22	S1 S2 S3 S4 S5 S6	S7
	23	S2 S3 S4 S5 S6 S7	S8
	24	S3 S4 S5 S6 S7 S8	S9
	25	S4 S5 S6 S7 S8 S9	S10
	26	S5 S6 S7 S8 S9 S10	S11
7 subsets	27	S1 S2 S3 S4 S5 S6 S7	S8
	28	S2 S3 S4 S5 S6 S7 S8	S9
	29	S3 S4 S5 S6 S7 S8 S9	S10
	30	S4 S5 S6 S7 S8 S9 S10	S11
8 subsets	31	S1 S2 S3 S4 S5 S6 S7 S8	S9
	32	S2 S3 S4 S5 S6 S7 S8 S9	S10
	33	S3 S4 S5 S6 S7 S8 S9 S10	S11
9 subsets	34	S1 S2 S3 S4 S5 S6 S7 S8 S9	S10
	35	S2 S3 S4 S5 S6 S7 S8 S9 S10	S11
10 subsets	36	S1 S2 S3 S4 S5 S6 S7 S8 S9 S10	S11

is used to test the statistical significance between the results because the objective is to compare the means of paired samples [95]. The means of RMSE calculated on both original and clean data sets are compared in this case. The null hypothesis is H_0: the mean of the forecasting errors calculated on original data sets is less than the mean of the forecasting errors calculated on clean data sets. The alternative hypothesis is H_1: the mean of the forecasting errors calculated on original data sets is greater than the mean of the forecasting errors calculated on clean data sets.

Table 5.13: Cross-validation results for the natural gas data set of operating area 8

Cross number	Avg RMSE (Original data set)	Avg RMSE (Clean data set)
1	41.40	41.55
2	50.55	47.10
3	60.67	37.95
4	44.32	42.37
5	47.11	32.18
6	49.65	46.20
7	60.53	37.95
8	43.50	41.62
9	46.88	31.81
10	60.75	38.16
11	42.00	40.28
12	46.50	31.11
13	42.60	40.14
14	46.12	30.82
15	45.75	30.60
μ	48.55	37.99
σ^2	45.72	30.75

The results of the one-tailed dependent samples t test are presented in Table 5.14 and Table 5.15 for operating area 8 and 9, respectively. The results indicate that the sample of clean data sets yield smaller forecasting errors in average

Table 5.14: Results for dependent samples t test for operating area 8

degrees of freedom	14
t-statistic	4.68
t critical one-tail	2.62
p-value	1.77×10^{-4}
level of significance	0.01

than the sample of original data sets, and that the difference is statistically significant in both cases at the 1% level of significance.

Table 5.15: Results for dependent samples t test for operating area 9

degrees of freedom	35
t-statistic	3.18
t critical two-tail	2.43
p-value	1.51×10^{-3}
level of significance	0.01

The mean RMSE by month and by unusual day, calculated on all crosses, for the natural gas data set of operating area 8 are presented in Figures 5.27 and 5.28. They show that the results are also practically significant. The mean RMSE by month and by unusual day calculated on clean data sets are smaller than mean RMSE by month and by unusual day calculated on original data sets.

The largest observed improvements in RMSE are 45% in April, 41% in June, and 33% in February. The average observed improvement from data cleaning is 21%

Table 5.16: Cross-validation results for the natural gas data set of operating area 9

Cross number	Avg RMSE (Original data set)	Avg RMSE (Clean data set)
1	10.93	10.93
2	12.10	12.10
3	12.48	12.48
4	10.85	10.85
5	12.87	12.87
6	12.53	12.45
7	12.95	12.97
8	21.76	14.68
9	12.45	12.45
10	12.95	12.95
11	10.52	10.52
12	12.58	12.57
13	12.22	12.12
14	12.10	12.10
15	21.25	13.63
16	13.00	13.00
17	11.20	11.20
18	12.60	12.60
19	11.99	11.89
20	12.46	12.46
21	21.08	12.82
22	11.28	11.28
23	12.60	12.60
24	11.80	11.70
25	12.35	12.35
26	20.95	12.70
27	12.57	12.57
28	11.75	11.67
29	12.25	12.25
30	20.77	12.30
31	11.73	11.65
32	12.22	12.22
33	20.75	12.35
34	12.25	12.25
35	20.78	12.38
36	20.80	12.45
μ	14.10	12.28
σ^2	14.41	0.59

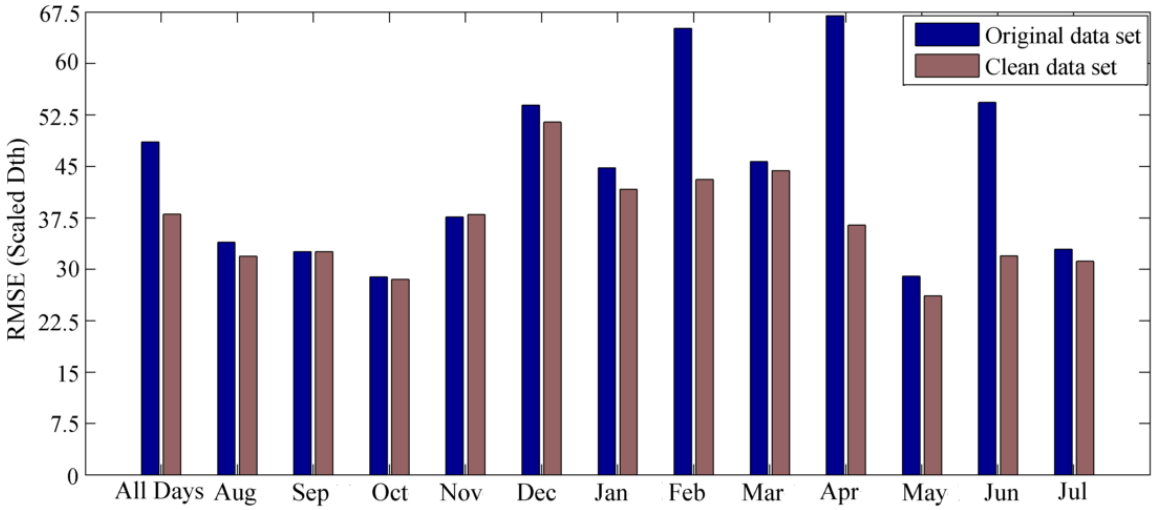

Figure 5.27: Mean RMSE by month for the cross-validation results of the data set of operating area 8

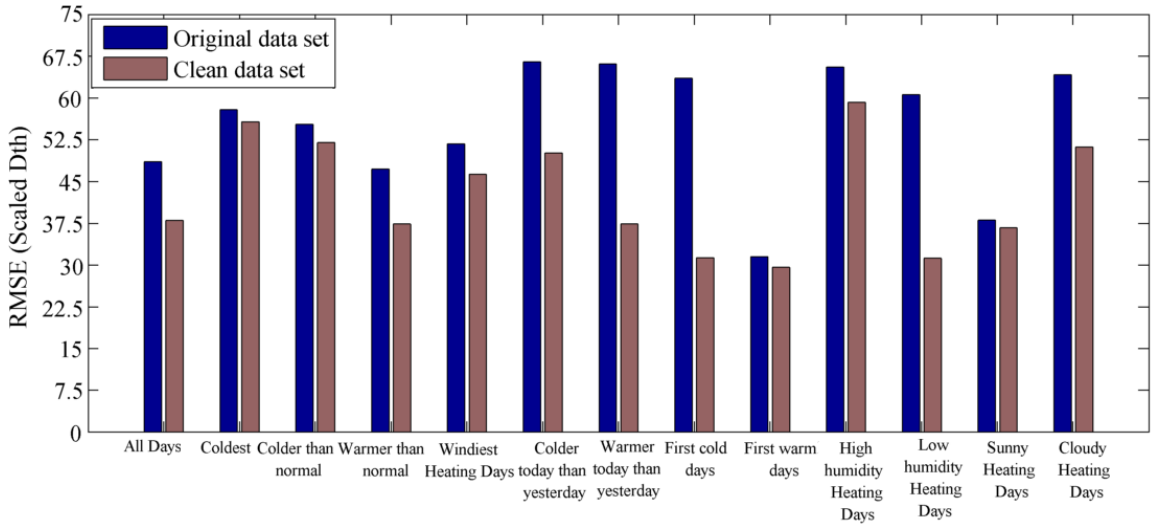

Figure 5.28: Mean RMSE by unusual day for the cross-validation results of the data set of operating area 8

for the natural gas data set of operating area 8. The largest observed improvement by unusual day is 51%, obtained for the first cold days.

For the natural gas data set of operating area 9, the comparison between

mean RMSE by month and mean RMSE by unusual day calculated on original and clean data sets are presented in Figures 5.29 and 5.30.

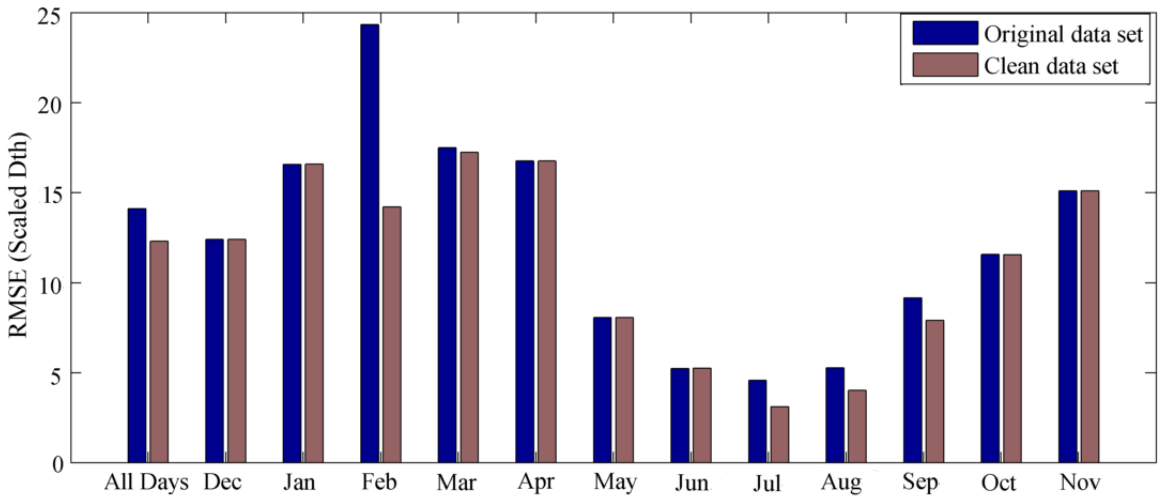

Figure 5.29: Mean RMSE by month for the cross-validation results of the data set of operating area 9

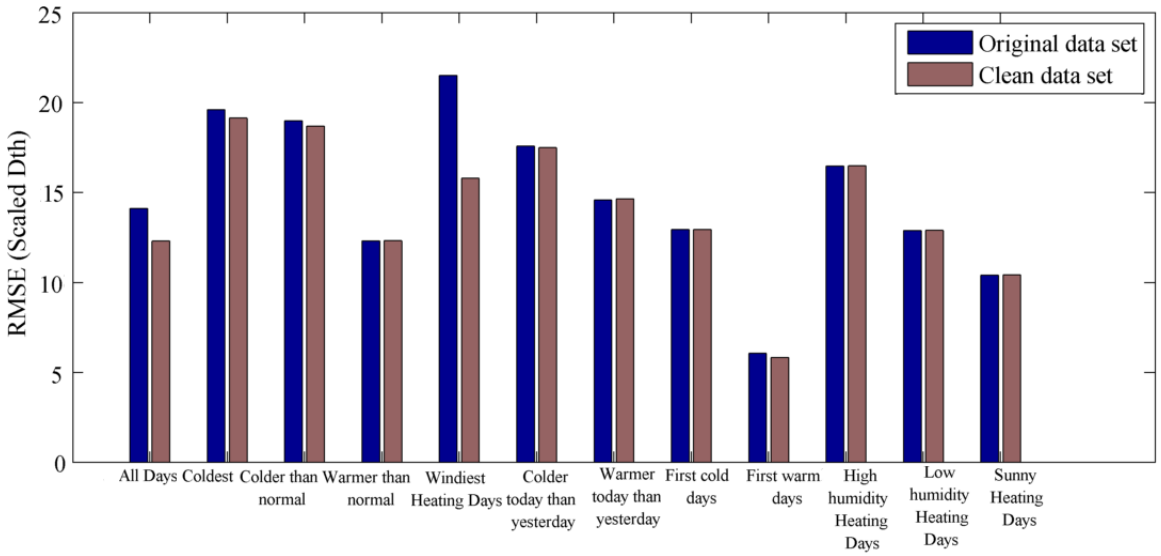

Figure 5.30: Mean RMSE by unusual day for the cross-validation results of the data set of operating area 9

For operating area 9, the results are also both practically and statistically significant at the 1% level of significance. The average observed improvement in RMSE is 12%, and the largest observed improvements in RMSE are 41% in February, 32% in July, and 24% in August. The largest observed improvement by unusual day is 26%, obtained for the windiest heating days. The variance of the results obtained on clean data sets is smaller (0.59) compared to the variance of the results obtained on original data sets (14.41), which indicates a better correlation between forecasting results in the case of clean data sets.

5.4.3 Cross-validation Analysis

For the natural gas data set of operating area 9, some months did not encounter any improvement because most of the anomalies were identified in the last year of data. The data cleaning provides average observed improvements of 21% and 12% for operating area 8 and 9, respectively. The largest observed improvements by unusual day varies depending on the weather and the region.

The cross-validation is not performed for the electric data set of operating area 10. The forecasting models used in this dissertation are designed for the prediction of natural gas demand. Therefore, the error associated with forecasting models should also be taken into consideration in this case. There is an error associated with forecasting models but they have been validated by Lim (2002),

Matin (2004), and Taware (1998) [60, 68, 92], and the assumption can be made that the natural gas forecasting models perform the same on all crosses.

The cross-validation determines that the improvement in forecasting error obtained by cleaning the data, which is the difference of errors between the performance of the models trained on original data set and the cleaned data set, is statistically significant at the 1% level of significance.

In this chapter, the data cleaning algorithm was evaluated both on simulated data (for all types of anomalies) and on utilities' data. They also were tested on both natural gas and electric energy data sets. A cross-validation is used to evaluate the performance of the data cleaning algorithm on smaller subsets and to make a general conclusion about the observed improvements. The next chapter of this dissertation presents a review of the contributions made and the results obtained. Recommendations for future work also are proposed to improve upon this research work.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

This chapter presents a summary of the contributions made by this research, the techniques developed, and the results obtained. Innovations proposed to improve the quality of this research work conclude this chapter. This dissertation presents energy time series data cleaning methods developed with the goal of improving forecasting accuracy. The literature survey presented in Chapter 2 shows that many techniques have been developed for anomaly detection, but very few have a practical usefulness. The techniques mostly are tested on simulated data sets and yield false positives in practice. The literature survey also underlines the fact that an accurate data cleaning tool should include domain knowledge of the problem. In this dissertation, energy demand forecasting knowledge is combined with probabilistic and statistical techniques to develop data cleaning algorithms for energy time series.

6.1 Summary of the Contributions

The contribution of this dissertation is the generalization of the data cleaning problem to energy time series. The hypothesis-driven anomaly detection algorithm is developed to identify anomalies in data sets with a 99% level of confidence. Statistical and probabilistic approaches are combined to detect anomalies in time series because probabilistic approaches do not yield good results on time series due

to their variability and exogenous factors. The energy time series domain knowledge also is incorporated into the algorithms to provide a practical significance. The data cleaning models are simple enough to extract exogenous factors influencing the data but also highlight the anomalies instead of modeling them. However, the imputation model is a complex forecasting model that models the trend, seasonality, and variability of the energy time series. Forecasting models are incorporated into the data cleaning algorithms and are used to improve the accuracy of data imputation. The data cleaning algorithms are applied to natural gas and electric time series data from utilities, as well as to simulated data sets. The analysis presented in Chapter 5 shows that cleaning the data provides a statistically and practically significant improvement in forecasting accuracy.

6.2 Summary of the Results

The data cleaning algorithm is tested on natural gas and electric reported consumption data sets from utilities. The measure of effectiveness of the results is the improvement in forecasting accuracy by cleaning the data. A cross-validation scheme with training subsets of various lengths is used to validate the data cleaning methods and to draw a general conclusion about the significance of the percentage of improvement. It is found that the observed improvements provided by data cleaning are both practically and statistically significant at the 1% level of

significance. The largest observed improvement in out-of-samples RMSE is 21%, which is also practically significant for energy demand forecasting.

6.3 Recommendations for Future Work

This research work can be improved by exploring other distributions and studying the impact of a better fit of the distributions on anomaly detection. From the example presented in Section 3.2, the normal probability distribution function does not fit the entire data set well but does approximate the tails of the distribution (which are the regions of interest). Other parametric and non-parametric distribution functions can be studied to determine their impact on data cleaning results. Also, the main areas of concern are the left and right tails of the distribution because they are the locations of the anomalies. Therefore, another idea is to use a percentage of the data instead of the entire data set to find the probability of a data point belonging to the distribution of the remaining data points.

Linear regression models are used in this dissertation to extract time series features and calculate the residuals of the data set. Machine learning techniques such as artificial neural networks (ANN) or support vector machines (SVM) could be studied to model time series in place of statistical methods. However, the models should not be over-trained, and the anomalous features should not be modeled in the process. Also, the linear regression models described in this dissertation use one

lag of the energy signal as autoregressive term. Because the data cleaning approach developed here is applicable to historical data sets, a forward autoregressive term of the energy signal can be studied in place of the lag of energy signal. Additionally, previous day temperature effects (lag of cooling degree days and lag of heating degree days wind-adjusted) can be included in the linear regression models, and their impact on time series data cleaning should be studied.

In this dissertation, the detrending and imputation models used are natural gas demand detrending and forecasting models, respectively. The imputation model provides an average observed improvement of 7% for the electric data set of operating area 10. That error can be reduced further by replacing the imputation and detrending models, in the case of electric time series data, with robust electric demand forecasting models and electric detrending models.

Machine learning classification techniques can be added also to the algorithms to improve the accuracy of the anomaly detection process and output a label for the category of an anomalous data point. The accuracy of the classifiers depends on the type and the numbers of anomalous features available. Also, the classifiers have to be re-trained per operating area, which is a tedious task. However, data transformation techniques such as surrogate data [18] could be used in this case to transform all anomalous features found on various data sets into a set of features used to train the classifiers and improve the accuracy of the data cleaning process.

Data cleaning is a problem found in other fields such as econometrics, finance, and medicine. The techniques developed in this dissertation can be expanded to other fields if the exogenous factors of the time series data are known and the imputation models for the particular context are well defined. The anomalies are found in the residuals of the time series. Therefore, the exogenous inputs are used in the linear regression models to extract the time series residual errors and to identify anomalies. The imputation models are used thereafter to calculate replacement values for the anomalies found.

BIBLIOGRAPHY

[1] R. Adnan, H. Setan, and M. N. Mohamad. Multiple outliers detection procedures in linear regression. *Matematika*, 1:29–45, 2003.

[2] M. Adya, S. J. Armstrong, F. Collopy, and M. Kennedy. Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, 17(2):143–157, 2001.

[3] H. N. Akouemo and R. J. Povinelli. Time series outlier detection and imputation. In *PES General Meeting — Conference Exposition 2014 IEEE*, pages 1–5, July 2014.

[4] R. R. Andridge and R. J. A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

[5] F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203–215, 2005.

[6] S. J. Armstrong, M. Adya, and F. Collopy. Rule-based forecasting: Using judgment in time-series extrapolation. In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell, MA, USA, 2001. Kluwer Academic.

[7] S. J. Armstrong and F. Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69–80, 1992.

[8] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, Hoboken, NJ, USA, 3rd edition, 1994.

[9] M. Beccali, M. Cellura, V. L. Brano, and A. Marvuglia. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area. *Renewable and Sustainable Energy Reviews*, 12:2040–2065, 2008.

[10] M. S. Beigi, S.-F. Chang, S. Ebadollahi, and D. C. Verma. Anomaly detection in information streams without prior domain knowledge. *IBM Journal of Research and Development*, 55(5 - Article 11):1–11, 2011.

[11] I. Ben-Gal. *Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for*

Practitioners and Researchers. Kluwer Academic Publishers, 2005.

[12] A. M. Bianco, M. García Ben, E. J. Martínez, and V. J. Yohai. Outlier detection in regression models with ARIMA errors using robust estimates. *Journal of Forecasting*, 20(8):565–579, 2001.

[13] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Pearson Prentice-Hall, San Francisco, CA, USA, 2nd edition, 2001.

[14] M. Bouguessa. A probabilistic combination approach to improve outlier detection. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, volume 1, pages 666–673, 2012.

[15] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, USA, 4th edition, 2008.

[16] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. OPTICS-OF: Identifying local outliers. In *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, volume 1704, pages 262–270, Prague, Czech Republic, 1999.

[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, volume 29, pages 93–104. ACM Press, 2000.

[18] R. H. Brown. Research results: The Heck-with-it hook and other observations. In *Southern Gas Association Conference: Gas Forecasters Forum*, Jacksonville, FL, USA, October 2007.

[19] R. H. Brown, Y. Li, B. Pang, S. R. Vitullo, and G. F. Corliss. Detrending daily natural gas demand data using domain knowledge. In *Proceedings of the 30th International Symposium on Forecasting*, 2010.

[20] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):1–47, 1998.

[21] O. Büyükalaca, H. Bulut, and T. Yilmaz. Analysis of variable-base heating and cooling degree-days for Turkey. *Journal of Applied Energy*, 69:269–283, 2001.

[22] G. Buzzi-Ferraris and F. Manenti. Outlier detection in large data sets. *Journal of Computers and Chemical Engineering*, 35:388–390, 2010.

[23] D. Carmona, M. A. Jaramillo, E. Gonzalez, and J. A. Alvarez. Electric energy demand forecasting with neural networks. In *IECON 02 [Industrial Electronics Society, IEEE 2002 28th Annual Conference of the]*, volume 3, pages 1860–1865, 2002.

[24] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):Article 15, 2009.

[25] I. Chang, G. C. Tiao, and C. Chen. Estimation of time series parameters in the presence of outliers. *Journal of Technometrics*, 30(2):193–204, 1988.

[26] S. Chawla and A. Gionis. k-Means: A unified approach to clustering and outlier detection. In *The 13th SIAM International Conference on Data Mining*, pages 189–197, Austin, TX, USA, 2013.

[27] C. Chen and L.-M. Liu. Joint estimation of model parameters and outlier effects. *Journal of American Statistical Association*, 88:284–297, 1993b.

[28] K. Choy. Outlier detection for stationary time series. *Journal of Statistical Planning and Inference*, 99:111–127, 2001.

[29] P. M. Dare. A study of the severity of the midwestern winters of 1977 and 1978 using heating degree days determined from both measured and wind chill temperatures. *Bulletin of American Meteorological Society*, 62(7):974–982, 1981.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(Series B):1–38, 1977.

[31] L. Denby and D. R. Martin. Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association*, 74(365):140–146, 1979.

[32] T. E. Dielman. *Applied Regression Analysis*. Brooks/Cole, 4th edition, 2005.

[33] R. A. T. Donders, J. Geert, T. Stijnend, and K. G. M. Moonsc. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59:1087–1091, 2006.

[34] Energy Information Administration. Energy explained - your guide to understanding energy, 2008. `http://www.eia.gov/energyexplained/`.

[35] C. Fauconnier and G. Haesbroeck. Outliers detection with the minimum covariance determinant estimator in practice. *Journal of Statistical Methodology*, 6(4):363–379, 2009.

[36] M. Goldstein and A. Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. In *Proceedings of the 35th German Conference on Artificial Intelligence (KI'12)*, pages 59–63, Saarbruecken, Germany, 2012. Stefan Wölfl.

[37] A. Grané and H. Veiga. Wavelet-based detection of outliers in financial time series. *Journal of Computational Statistics and Data Analysis*, 54:2580–2593, 2010.

[38] A. Gupta, A. Gupta, and A. Mishra. Research paper on cluster techniques of data variations. *International Journal of Advance Technology & Engineering Research*, 1(1):39–47, 2011.

[39] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson. *Multivariate Data Analysis*. Prentice Hall, 7th edition, 2010.

[40] D. M. Hawkins. *Identification of Outliers*. Chapman and Hall, England, United Kingdom, 1980.

[41] S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier detection using replicator neural networks. *Data Warehousing and Knowledge Discovery - Lecture Notes in Computer Science*, 2454:170–180, 2002.

[42] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Journal of Knowledge and Information Systems*, 26(2):309–336, 2011.

[43] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[44] T. Hong. Energy forecasting: Past, present and future. *Foresight: The International Journal of Applied Forecasting*, (32):43–48, 2014.

[45] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[46] International Bureau of Weights and Measures. *The International System of Units (SI)*, pages 118,144. 8th edition, 2006.

[47] N. K. Jajo. Graphical display in outlier diagnostics, adequacy and robustness. *Statistics and Operations Research Transactions*, 29(1):1 –10, 2005.

[48] J. M. Jerez, I. Molina, P. J. García-Laencinac, E. Albad, N. Ribellesd, M. Martíne, and L. Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Journal of Artificial Intelligence in Medicine*, 50(2):105–115, 2010.

[49] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. A local search approximation algorithm for k-means clustering. *Journal of Computational Geometry*, 28(2-3):89–112, 2004.

[50] A. Kaya. Statistical modelling for outlier factors. *Ozean Journal of Applied Sciences*, 3(1):185–194, 2010.

[51] O. Kaynar, I. Yilmaz, and F. Demirkoparan. Forecasting of natural gas consumption with neural networks and neuro fuzzy system. *Energy Education Science and Technology Part A: Energy Science and Research*, 26(2):221–238, 2011.

[52] E. N. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the International Conference on Very Large DataBases*, 1998.

[53] E. N. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: Algorithms and applications. *The International Journal on Very Large DataBases*, 8:237–253, 2000.

[54] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143, San Mateo, CA, USA, 1995.

[55] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica*, 31:249–268, 2007.

[56] H.-P. Kriegel, P. Kröger, and A. Zimek. Outlier detection techniques. *The 2010 SIAM International Conference on Data Mining*, 2010.

[57] D. Kunkle, V. Slavici, and G. Cooperman. Parallel disk-based computation for large, monolithic binary decision diagrams. In *Proceedings of the 4th International Workshop on Parallel and Symbolic Computation*, PASCO '10, pages 63–72, New York, NY, USA, 2010. ACM.

[58] S. Labovitz. Criteria for selecting a significance level: A note on the sacredness of 0.05. *The American Sociologist*, 3(3):220–222, 1968.

[59] A. H. Lee and W. K. Fung. Confirmation of multiple outliers in generalized linear and nonlinear regressions. *Journal of Computational Statistics and Data Analysis*, 25(1):55–65, 1997.

[60] H. L. E. Lim. Computational intelligence models for short term natural gas demand forecasting. Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, USA, August 2002.

[61] R. J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

[62] R. J. A. Little. Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[63] R. J. A. Little and D. B. Rubin. The analysis of social science data with missing values. *Sociological Methods & Research*, 18(2-3):292–326, 1989.

[64] H. Liu, S. Shah, and W. Jiang. On-line outlier detection and data cleaning. *Journal of Computer and Chemical Engineering*, 28(9):1635–1647, 2004.

[65] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Journal of Neural Networks*, 43:72–83, 2013.

[66] K. W. Magld. Features extraction based on linear regression technique. *Journal of Computer Science*, 8(5):701–704, 2012.

[67] M. Markou and S. Singh. Novelty detection: A review - part 1: Statistical approaches. *Journal of Signal Processing*, 83:2481–2497, 2003.

[68] I. Matin. Artificial neural network models to predict gas consumption. Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, USA, November 1995.

[69] E. Q. McCallum. *Bad Data Handbook: Mapping the World of Data Problems*. O'Reilly Media, Sebastopol, CA, USA, 2012.

[70] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, USA, 2012.

[71] C. R. Muirhead. Distinguishing outlier types in time series. *Journal of the Royal Statistical Society. Series B*, 48(1):39–47, 1986.

[72] N. E. Nahi. Optimal recursive estimation with uncertain observation. *IEEE Transactions on Information Theory*, IT-15(4):457–462, 1969.

[73] NaturalGas.org. Uses, 2014. `http://www.naturalgas.org/overview/uses`.

[74] P. Palaanen. Bayesian classification using Gaussian mixture model and EM estimation: Implementations and comparisons. Technical report, Lappeenranta University of Technology, Lappeenranta, Finland, 2004.

[75] B. Pang. The impact of additional weather inputs on gas load forecasting. Master's thesis, Marquette University, Department of Electrical and Computer Engineering, Milwaukee, WI, USA, Summer 2012.

[76] A. Papoulis and U. S. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Europe, Boston, MA, USA, 4th edition, 2002.

[77] C. Phong and R. Singh. Missing value estimation for time series microarray data using linear dynamical systems modeling. *22nd International Conference on Advanced Information Networking and Applications*, page 814819, 2008.

[78] T. D. Pigot. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

[79] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts, and J. Ye. Statistical models for reconstructed phase spaces for signal classification. *IEEE Transactions on Signal Processing*, 54(6):2178–2186, 2006.

[80] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, volume 29, pages 427–438. ACM Press, 2000.

[81] I. Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001*

workshop on empirical methods in artificial intelligence, volume 3, pages 41–46, 2001.

[82] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[83] D. B. Rubin. Multiple imputations in sample surveys - A phenomenological Bayesian approach. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 20–34, 1978.

[84] L. H. Rubin, K. Witkiewitz, J. St. Andre, and S. Reilly. Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. *The Journal of Undergraduate Neuroscience Education*, 5(2):71–77, 2007.

[85] T. Sauer, A. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65(3/4):579–616, 1991.

[86] J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147177, 2002.

[87] J. L. Schafer and M. K. Olsen. Multiple imputation for multivariate missing-data problems: A data analysts perspective. *Multivariate Behavioral Research*, 33:545571, 1998.

[88] W. Schen, V. Babushkin, Z. Aung, and W. L. Woon. An ensemble model for day-ahead electricity demand time series forecasting. In *Proceedings of the Fourth International Conference on Future Energy Systems*, e-Energy '13, pages 51–62, New York, NY, USA, 2013. ACM.

[89] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461464, 1978.

[90] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, chapter 10, pages 651–683. Pearson Addison Wesley, Boston, MA, USA, 2006.

[91] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEEE International Conference on Artificial Neural Networks*, volume 4, pages 442–447, 1995.

[92] A. Taware. Forecasting and identification methods applied to gas load estimation problems. Master's thesis, Marquette University, Department of

Electrical and Computer Engineering, Milwaukee, WI, USA, December 1998.

[93] D. M. J. Tax and R. P. W. Duin. Outlier detection using classifier instability. In *SSPR '98/SPR '98 Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition*, pages 593–601, 1998.

[94] R. S. Tsay. Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7(1):1–20, 1988.

[95] T. C. Urdan. *Statistics in Plain English*. Routledge Taylor and Francis Group, New York, NY, USA, 3rd edition, 2010.

[96] J. Van den Broeck, S. Argeseanu Cunningham, R. Eeckels, and K. Herbst. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medecine*, 10(2):0966–0970, 2005.

[97] S. Velilla. A note on the behaviour of residual plots in regression. *Statistics & Probability letters*, 37:269–278, 1998.

[98] S. R. Vitullo, R. H. Brown, G. F. Corliss, and B. M. Marx. Mathematical models for natural gas forecasting. *Canadian Applied Mathematics Quaterly*, 17(4):807–827, 2009.

[99] P. T. von Hippel. Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37:83–117, 2007.

[100] A. R. Weekley, R. K. Goodrich, and L. B. Cornman. An algorithm for classification and outlier detection of time-series data. *Journal of Atmospheric and Oceanic Technology*, 27(1):94–107, 2010.

[101] P. W. Wilson. Detecting outliers in deterministic nonparametric frontier models with multiple outputs. *Journal of Business & Economic Statistics*, 11(3):319–323, 1993.

[102] J. W. Wisnowski, D. C. Montgomery, and J. R. Simpson. A comparative analysis of multiple outlier detection procedures in the linear regression model. *Journal of Computational Statistics and Data Analysis*, 36(3):351–382, 2001.

[103] X. Xu, V. Kumar, R. J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Journal of Knowledge and Information Systems*, 14(3):1–37, 2008.

[104] K. Yamanishi, J. Takeuchi, and G. Williams. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 320–324. ACM Press, 2000.

[105] K.-V. Yuen and H.-Q. Mu. A novel probabilistic method for robust parametric identification and outlier detection. *Journal of Probabilistic Engineering Mechanics*, 30:48–59, 2012.

[106] A. Zaharim, R. Rajali, R. M. Atok, I. Mohamed, and K. Jafar. A simulation study of additive outlier in ARMA(1,1) model. *International Journal of Mathematical Models and Methods in Applied Science*, 3(2):162–169, 2009.

[107] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgensen, and J. Ucles. HIDE: A hierarchical network intrusion detection system using statistical preprocessing and neural network classification. In *Proceedings of IEEE Workshop on Information Assurance and Security*, pages 85–90, 2001.

[108] Y. Zhou, H. Yu, and X. Cai. A novel k-means algorithm for clustering and outlier detection. In *2009 Second International Conference on Future Information Technology and Management Engineering*, pages 476–480, 2009.

[109] C. Zou, S.-T. Tseng, and Z. Wang. Outlier detection in general profiles using penalized regression method. *IIE Transactions*, 46(2):106–117, 2014.