UPPSALA
UNIVERSITET

UPTEC STS 20024
Examensarbete 30 hp
Juni 2020

Applying Machine Learning Algorithms for Anomaly Detection in Electricity Data

Improving the Energy Efficiency of Residential Buildings

Herman Guss
Linus Rustas

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Applying Machine Learning Algorithms for Anomaly Detection in Electricity Data

Herman Guss, Linus Rustas

The purpose of this thesis is to investigate how data from a residential property owner can be utilized to enable better energy management for their building stock. Specifically, this is done through the development of two machine learning models with the objective of detecting anomalies in the existing data of electricity consumption. The dataset consists of two years of residential electricity consumption for 193 substations belonging to the residential property owner Uppsalahem.

The first of the developed models uses the K-means method to cluster substations with similar consumption patterns to create electricity profiles, while the second model uses Gaussian process regression to predict electricity consumption of a 24 hour timeframe. The performance of these models is evaluated and the optimal models resulting from this process are implemented to detect anomalies in the electricity consumption data. Two different algorithms for anomaly detection are presented, based on the differing properties of the two earlier models.

During the evaluation of the models, it is established that the consumption patterns of the substations display a high variability, making it difficult to accurately model the full dataset. Both models are shown to be able to detect anomalies in the electricity consumption data, but the K-means based anomaly detection model is preferred due to it being faster and more reliable. It is concluded that substation electricity consumption is not ideal for anomaly detection, and that if a model should be implemented, it should likely exclude some of the substations with less regular consumption profiles.

Handledare: Åsa Engström och Tomas Nordqvist
Ämnesgranskare: Fatemeh Johari
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 20024

Populärvetenskaplig sammanfattning

Detta examensarbete är skrivet med handledning från Uppsalahem, det kommunala fastighetsbolaget i Uppsala. Uppsalahem är idag det största fastighetsbolaget i Uppsala och har över 17 000 lägenheter till förfogande. Uppsalahem har högt uppsatta mål mot att bli alltmer hållbara och detta återspeglar sig i att de försöker energieffektivisera sina byggnader, exempelvis genom att minska onödig konsumtion av energi och samtidigt upprätthålla en hög komfort. De har idag en avvikelsedetektering som grundar sig i att undersöka hur konsumtionen av olika energislag förändras. Månadsvärden är idag det huvudsakliga tillvägagångssättet för att samla in data för denna avvikelsedetektering. Idag jämförs månadskonsumtion för en given månad med samma månad föregående år, och en avvikelse detekteras när en tillräckligt stor förändring har skett jämfört med samma månad föregående år. Exakt hur stor skillnaden måste vara för att detektera en avvikelse beror på arean av det undersökta området eller hur mycket avvikelsen bedöms kosta. Uppsalahem har ett intresse i att analysera om denna feldetektering kan uppdateras och skötas snabbare och mer träffsäkert. Detta vill de undersöka genom att analysera data som istället är insamlad på timbasis. Denna rapport kommer examinera data över elkonsumtion hos substationer för 2018 och 2019. Syftet med detta projekt är även att undersöka vilka andra datakällor som är av intresse för Uppsalahem utifrån ett energieffektiviseringsperspektiv och om det är möjligt att utnyttja Uppsalahems tillgängliga elektricitetsdata för att skapa en modell för snabbare avvikelsedetektering.

Tillvägagångssättet för att undersöka huruvida det är möjligt att nyttja Uppsalahems tidigare data för snabbare avvikelsedetektering delades upp i två modeller. Den första modellen utnyttjar klustring. Klustring är en procedur som undersöker en mängd data för att sedan samla ihop data som liknar varandra. Exempelvis så är frukter klustrade i matbutiken då päronen inte befinner sig i samma korg som äpplena och bananerna. Klustring bygger på att försöka se mönster i den data som finns för att samla substationer som liknar varandra i samma "korg". När man sedan har klustrat stationer med liknande mönster kan man då ta fram en centroid (ett genomsnitt) som förväntas representera detta kluster på ett bra sätt. Det här genomsnittet betraktas som en representation för hur konsumtionen borde se ut för de stationer som samlats i det klustret. Vid en jämförelse mellan de individuella substationerna och detta genomsnitt så indikerar en avvikelse mellan dessa att ett fel har skett.

Den andra modellen utnyttjar prognostisering. Tanken är att det ska finnas underliggande trender och mönster i elkonsumtionen hos substationerna. En regressionsmodell anpassas till datan för att lära sig detta mönster, regressionsmodellen skapar således en funktion som motsvarar datan till så bra som möjligt. Denna modell använder sig av begreppen träningsdata och testdata. Träningsdatan är den data som modellen lär sig mönstret på medan testdata är data som jämför hur väl modellen lyckas prognostisera de framtida värdena. Om det finns en stor likhet mellan den prognosticerade konsumtionen och den faktiska konsumtionen går det att argumentera för att modellen kan prognostisera framtida konsumtion. Vid en tillräckligt stor skillnad mellan den faktiska konsumtionen och den prognostiserade så indikerar detta att en avvikelse har skett.

En av de slutsatser som dras i detta examensarbete är att den studerande datan är väldigt varierande i termer av regelbundenhet, kvalitet och upplösning. Detta gör att det är svårt att förstå varför konsumtionen ser ut som den gör och hur den lämpligast modelleras.

De två modellerna för feldetektion som har prövats i detta arbete varierar kraftigt i prestanda vilket i sin tur beror på den variation som finns i datan. Den klustringsbaserade modellen bedöms prestera bättre för målet att lokalisera avvikelser än den regressionsbaserade modellen. De substationer som har en klar regelbundenhet kan predikteras väl av regressionsmodellen, dessa är dock bara en mindre del av alla substationer. Av den anledningen presterar den klustringsbaserade modellen generellt bättre för datasetet. Jämfört med Uppsalahems nuvarande modell för att detektera avvikelser så finns det förbättringsmöjligheter vid en implementation av en klustringsmodell. Detta exemplifieras främst i de fall där klustringsmodellen lyckas hitta avvikelser på tider som är avsevärt kortare än en vecka.

Acknowledgements

This degree project within sociotechnical systems engineering (STS) is conducted with the contribution of Uppsalahem. The supervisors from Uppsalahem have been Åsa Engström and Tomas Nordqvist and have throughout the project been of help to us with their advice and support. Fatemeh Johari has been the subject reader of this project and has contributed with expertise in the studied area which has been immensely helpful in understanding the project. We would like to thank all of the above-mentioned persons for taking the time and guiding us through this project with their expertise in the subject.

Herman Guss & Linus Rustas

Uppsala, June 2020

Table of contents

1 Introduction

Residential buildings are among the largest energy consumers in Sweden. According to the Swedish Energy Agency the residential and service sector in 2017 had a total energy usage of 146 TWh, accounting for 39 % of the total energy consumption (Energimyndigheten, 2019). The residential subsector is the single largest consumer within this sector, with a total energy consumption of 87 TWh.

For residential buildings, the energy demand can be divided into different segments such as heating and electricity, which are measured in different capacities. The level of detail in these measurements may however vary between buildings based on different features such as size or the year of construction. For buildings reliant on district heating, heating is usually the largest share of the energy consumption followed by electricity. For apartment buildings in Sweden, district heating is by far the most common source of heating, according to IVA (2012b) being present in approximately 93 % of the multi-family residential buildings. Electricity accounts for roughly 20 % of the energy demand of Swedish residential buildings, although for a variety of reasons this share is gradually increasing, and electricity is predicted to account for as much as 40 % of the residential energy consumption by 2050 (IVA 2012a).

Lowering this energy consumption would have economical as well as environmental benefits. In addition to this, political regulations such as *Boverket's mandatory provisions and general recommendations* (BFS 2011:6) (Boverket, 2019) place increasing legal demands on energy efficiency for new, reconstructed or expanded buildings. Thus, there are several strong incentives for property owners to optimize buildings to reduce the demand of energy. Through more efficient energy use, there is potential to reduce the energy consumption while retaining utility and comfort for the end users. Such possibilities include energy profiling in order to single out buildings or areas which have an uncharacteristic consumption behavior over time or to group buildings or areas where the energy consumption follows distinct seasonal or diurnal patterns, as well as fast detection of anomalies in the energy consumption.

Due to technological development and the increasing digitalization there is today a rather large set of data associated with most residential buildings, which is also continually increasing in size. This data relates both to the characteristics of the buildings as well as their energy consumption, which might aid in increasing the energy efficiency in buildings if properly analyzed. Energy efficiency can be defined as when an appliance utilizes less energy but the performance remains the same or if the performance increases while using the same amount of energy (OVOEnergy, n.d.). This definition can also be applied to buildings where energy efficiency can be seen as increasing the comfort using the same amount of energy or decreasing energy consumption while retaining a good comfort (IVA, 2012b).

Large datasets are typically well suited for analysis using methods which fall under the umbrella term machine learning. *The hundred-page machine learning book* (Burkov, 2019) defines machine learning concisely in the following way:

> *"Machine learning is a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon. These examples can come from nature, be handcrafted by humans or generated by another algorithm.*
>
> *Machine learning can also be defined as the process of solving a practical problem by 1) gathering a dataset, and 2) algorithmically building a statistical model based on that dataset. That statistical model is assumed to be used somehow to solve the practical problem."* (Burkov, 2019).

In recent years several different machine learning techniques have been proposed and implemented for estimation of heating and cooling loads, energy consumption and performance in the building sector (Seyedzadeh et al., 2018). There are several applications within this sector including the development of new low energy building stock, energy retrofitting for old stock and optimization of energy management systems and heat, ventilation and air conditioning systems. Energy management systems have been utilized for energy data collection and consumption control which are fundamental to energy waste reduction. Therefore, a great amount of data related to different sensors is often readily available, and there is a demand for analytical tools that make use of that data to enable assessment of energy performance.

One possible area of application for machine learning techniques is learning patterns and creating models to predict the behavior of different objects. This might also be applied within the field of energy efficiency as models may bring new knowledge of the energy system behavior whether it be on the supply or demand side. Such models can predict energy production and demand which can be used to match needs on the energy market. Alternatively, models which learn an expected pattern from historical data may compare predictions to future data in order to detect unexpected patterns, which may be indicative of anomalies. Detecting anomalies may thus increase the energy efficiency as the performance remains the same while the energy consumption decreases. In this thesis, the possibility of such anomaly detection is investigated as the intent is to utilize two different models, utilizing energy profiling and forecasting respectively, in order to detect anomalies in the energy consumption of a residential building stock owned by the largest housing company in Uppsala, Uppsalahem.

1.1 Purpose

Using data mining and machine learning techniques, this thesis aims to improve the energy management system in residential buildings owned by the housing company Uppsalahem. The purpose is then to investigate a dataset in order to analyze and determine its applicability for being utilized in an anomaly detection system. Within this analysis, two models for anomaly detection are developed. A comparison is then made between the two different models to determine if either or both of the approaches are viable. The first model utilizes clustering and then generates a model based on the aggregate behavior of the cluster. The second model considers each substation individually. It is built upon regression and applies a Gaussian process regression to learn patterns from historical behavior and fits a model to predict future consumption. The goal is then to compare these models to the individual electricity consumption data series and determine their usefulness for detecting anomalies in the electricity consumption.

Thus, the aim of this thesis is to investigate whether the available data from measurements of electricity consumption at a residential property owner may be utilized to create machine learning models to enable a fast and reliable anomaly detection system. Such anomaly detection models should ideally detect probable faults as often as possible without signaling for anomalies during intervals which do not show abnormal behavior. Additionally, it should ideally allow for the detection of anomalies in energy within as short a timeframe as possible, preferably within a daily time range. To further this aim a set of data over electricity consumption is acquired from Uppsalahem.

The research questions of this master thesis are thereby:

- Can the available data at Uppsalahem be utilized for machine learning algorithms in order to detect anomalies affecting electricity consumption?
- What additional data may be of interest to collect in order to allow for further development of energy efficiency algorithms?

1.2 Methodology overview

This thesis strives to develop an accurate anomaly detection model for the real estate company Uppsalahem through an analysis of the available data. The dataset is first subjected to preprocessing where data which is not deemed to meet the quality requirements of the study is either interpolated or removed. Utilizing the preprocessed dataset, two models are then developed with the goal of enabling faster and more reliable anomaly detection. The first is a clustering based model, performing K-means clustering on normalized data. The second is a probabilistic regression based model, Gaussian process regression which considers relationships between the electricity consumption and factors such as time and temperature. Two separate performance evaluations are conducted for the developed models before proceeding to implement them in order to determine their capability to detect abnormal electricity consumption patterns in the studied dataset. Anomaly detection models are developed utilizing each of these techniques, differing slightly in implementation due to the different specifics of each model. These anomaly detection models subsequently undergo a basic optimization using optically identified anomalies as validation data, and their respective performance is evaluated through comparisons of their detection speed and ability to accurately detect anomalies.

1.3 Limitations

One of the limitations of this thesis is the availability of data at the desired spatiotemporal resolution. The collected electricity consumption data is on a substation level, which prevents the linkage of electricity profiles to individual physical buildings, which could otherwise have allowed for considering the physical properties of buildings in the analysis. Also, if the collected data had comprised a longer time interval, it could also have allowed for more extensive validation of the anomaly-finding algorithms, as the short timeframe of the testing intervals limits the possibility of evaluating the model behavior on a longer time scale. Additionally, due to not having access to data of heat usage in this thesis, only electricity is evaluated during the anomaly detection. However, as mentioned previously heat is the largest share of building energy demand and analysis of heat usage measurements can therefore bring considerable advantages, particularly in district heated buildings. Finally, there is no proper record on previous

anomalies in electricity measurements that can be used for performance evaluation of the developed anomaly detection model. The performance of the anomaly detection models is therefore evaluated on a rather small set of data where anomalies have been labeled by optical analysis.

1.4 Report overview

The report is outlined as follows, firstly, in Section 2, Background, necessary information about the housing company, Uppsalahem, and required background on machine learning and its application in increasing energy efficiency in buildings are presented. Section 3, Methodology and data, contains descriptions of the method utilized in this study, relevant theoretical concepts and their implementations. Thereafter the results that are produced in this thesis are presented with a continuous analysis in Section 4, Results and Analysis. Following the results is Section 5, Discussion, which compares the results from this thesis with relevant results from similar research articles to give perspective on the results. Lastly, the Conclusion of this thesis is presented in Section 6, which summarizes the report and the research findings.

2 Background

The background firstly presents information about Uppsalahem, in Subsection 2.1, which aims to give a short description of who they are and how their data management currently functions. Thereafter, Subsection 2.2 provides an overview of machine learning in the energy sector. This section also aims to explain how machine learning can be utilized to improve residential facility energy management. Following this, Subsection 2.3 provides more detailed information about specific applications of machine learning to building energy data. These different applications include energy profiling, forecasting and anomaly detection. Lastly, Subsection 2.4 is presented, to give a clearer definition of what is considered an anomaly within the scope of this study.

2.1 Uppsalahem

Uppsalahem is the single largest real estate owner in the city of Uppsala. The vast majority of Uppsalahem's property stock consists of 17000 apartments in which more than 30 000 people reside (Uppsalahem, n.d.b). Being part of the Swedish public housing, Uppsalahem has a mission from the municipality of Uppsala to provide access to good quality housing. Within this mission there is also a social responsibility, meaning that they must take social impacts, environmental effects and sustainability into account in their work. Furthermore, Uppsalahem actively works towards being even more sustainable. (Uppsalahem, n.d.a) This is executed through improvement of energy efficiency in building renovations, but also the development of efficient energy analysis and management systems. One of their current goals toward increased sustainability is to be increasingly efficient in detecting anomalies which in turn will lead to better and faster maintenance and a reduced use of energy.

Uppsalahem also collects and stores a rather large amount of data about these apartment buildings and their energy demand, and for that reason they provide a suitable foundational dataset for the types of analysis described in Subsection 2.2. In terms of energy related data, they measure the consumption of electricity, cold water, hot water and district heating on a monthly basis, often gathered through manual readings of measurement devices. A consequence of this is that they collect the data at somewhat irregular intervals, sometime around the turn of the month and not always at the same date. The consumption is collected at a substation level meaning that not just one but a number of buildings are connected to a single measurement point.

Uppsalahem's current system for detecting anomalies in energy consumption compares the monthly consumption to that of the same month the year before, for instance March 2019 compared to March 2018. Uppsalahem's buildings are grouped into residential districts, and the deviation reports are generated based on these districts. The anomalies found by these reports are investigated according to a certain order of priorities, these priorities are described in Table 1.

Table 1, the prioritization of Uppsalahem's anomaly detection

Prioritization	Situation
1st	Increase in energy consumption over 50% or an increase in cost by 100 000 SEK or more, annually
2nd	Increase in cost by 50 000 to 100 000 SEK annually
3rd	Area > 15 000 m^2 and a change in the consumption that is greater than 5 % or area 5 000-15 000 m^2 and a change in the consumption that is greater than 7.5 % or area 1 500-5 000 m^2 and a change in the consumption that is greater than 10% or area < 1 500 m^2 and a change in the consumption that is greater than 25%

Uppsalahem also manages anomalies on a substation level. However, for the individual substation the threshold for anomalies is set to 25% always. The anomalies are withdrawn from Uppsalahem's program Insikt. These reports only contain information about detected anomalies on a month-by-month basis. Thus, there is likely room for improvements towards detecting anomalies on a daily or even hourly basis.

2.2 Use of machine learning for energy efficiency

Machine learning is an algorithm that uses a real phenomenon and tries to create a model that replicates the phenomenon to the highest degree possible. These phenomena can be distributed throughout the world and can be ordinary situations as well as highly complex situations. Machine learning algorithms are broadly categorized into two groups known as, supervised and unsupervised. (Burkov, 2019)

The supervised machine learning algorithms aim to produce a model from a labeled dataset. This means that the model should take some input vector that describes a set of features and give information deduced from the input as the output. The aim of the model is therefore to learn a functional relationship between input variables and output variables. A typical example of supervised learning is the regression problem, $y = f(x)$ where the goal is to learn the dependent variable y as a function of the independent variable x. Unsupervised machine learning algorithms also use a dataset, however, in this case the data is unlabeled, and no distinction between independent and dependent variables is made. The goal is instead to learn an expected distribution of the data as a whole. The model might also create a vector with known parameters which is a transformation from the original unknown dataset. (Burkov, 2019) This project will explore the applicability of techniques from both these groups of machine learning algorithms to a problem defined for a set of energy data.

Over the past decades, energy demand in the building sector has been steadily increasing and according to Amasyali and El-Gohary (2018) it might be due to the increase in population combined with urbanization and increased social demands. As discussed by Allouhi et al. (2015) buildings contribute to the world's energy consumption and consequently greenhouse gas emissions, considerably. Thus, in line

with global climate mitigation and energy efficiency goals, a more energy efficient approach to building energy data is necessary. Building energy efficiency measures and analysis are among other things necessary to help reduce greenhouse gas emissions and the ability to predict future energy consumption is an important enabler of energy efficiency improvements. This ability is also highly useful to actors on the energy market such as utility companies, facility managers and end users, who may increase their efficiency by adapting their behavior to expectations. Knowledge about energy consumption patterns is vital for scheduling maintenance and ordinary operations to enable retainment or improvement of the energy performance of buildings. (Pham et al., 2020) Additionally, better understanding of energy consumption data might lead to increased financial savings and enhancement of the energy security of customers (McNeil et al., 2019). One example of such energy data which is highly interesting to study is that of time series data for buildings. Horrigan et al. (2018) uses time series data in order to improve building operational behavior, i.e. the energy and environmental performance of the building, by conducting a fault detection analysis. They further state that the detection of statistically significant faults in building performance data is an asset to building managers and that it can lead to significantly reduced energy losses.

2.3 Applications of machine learning to building energy data

Applications of machine learning are common in the field of residential energy data analysis. For the scope of this project, the main focus is on applications of machine learning for energy profiling and energy forecasting which can later be used in development of the anomaly detection algorithms. Examples of supervised and unsupervised machine learning models as previously described in Subsection 2.2 and examples of applications of these models for building energy data are provided.

2.3.1 Electricity profiling

An energy load profile contains information about the energy demand of a consumer or set of consumers, and how this demand is distributed. Electricity load profiles provide an approach to describe the typical behavior of electricity consumption (Zhang et al., 2018). This is utilized to quantify the total consumption contribution of different sub-components and features of the buildings or to distinguish usage characteristics. Profiling electricity use has the potential capability to educate end-users through feedback on how to change their consumption behavior. For utility companies these load profiles may be utilized to reach a certain load-shape objective. The most commonly implemented methods for electricity profiling are according to Wei et al. (2018) clustering methods, such as K-means or hierarchical clustering. Electricity profiles can be used to approximate the demand during critical periods and the load placed on the electricity supplier during those times. Load profiles in the area of electricity consumption have applications both on a general level as well as in more specific cases. On the general level they are useful to utility companies that wish to estimate the pattern of the total load placed upon the grid by a set of consumers (Singh & Yassine, 2018). On a more specific level, they may be utilized for approximating the contribution of individual parameters to the total load, for instance the load profile of household appliances (Issi & Kaplan, 2018) or electric vehicles (Lu et al., 2017). Knowledge of total load profiles as well as those of individual items can be exploited to inform and adapt consumer behavior to create opportunities for management of energy

consumption (Issi & Kaplan, 2018) and balance of supply and demand on the electric market (Damayanti et al., 2017; Zhang et al., 2018), which is of relevance to major distribution companies as well as micro grid owners (Damayanti et al., 2017).

The energy data exploited to create a profile is normally collected with certain constant time intervals, such as every 10, 15 or 30 minutes (Damayanti et al., 2017). Electricity load profiles are typically created on a daily or weekly time window to display consumption as a function of the day-to-day behavior of individuals, for instance typical electricity load profiles may display peak electricity usage in the evening for residential buildings (Marszal-Pomianowska et al., 2016) or markedly different load curves between weekdays and weekends for an office building (Bedingfield et al., 2018). Widén et al. (2009) generate electricity consumption profiles based on historical data. Their data collection intervals vary between one minute at the most frequent and hourly measurements at the least frequent. Furthermore, the authors conclude that the model they implement can generate close-to-reality electricity consumption predictions.

Clustering algorithms are commonplace within energy profiling, and they may work with either the consumption data within the time domain or other attributes constructed to represent the load curve (Zhang, 2018). The goal of clustering is to group data points based on similarity as measured by some metric, commonly Euclidean distance. This may be used to attain representative electricity profiles for groups of electricity consumers. According to Bedi and Toshniwal (2019) cluster analysis is utilized to collect groups of data that have a high similarity to each other and are highly unlike the other clusters, which might assist in finding natural groups with similar patterns in the data. They argue that clustering analysis helps identify trends in the consumption which can then be applied in load characterization to achieve a deeper understanding of consumption patterns. Nepal et al. (2019) implement K-means clustering to create day profiles for a set of university buildings, and arrive at a clear relationship of increased electricity usage during daytime hours and weekdays compared to weekends. Similarly, Damayanti et al. (2017) apply K-means, Fuzzy C-means and K-Harmonic Means to obtain two clusters for electricity consumption in West Java, one representing weekday profiles and the other weekends. K-means clustering has been used in a variety of circumstances, amongst them is the identification of daily electricity consumption of buildings (Miller et al., 2015; Miller and Schluter, 2015). Chicco (2012) performs a thorough investigation on several different clustering techniques and finds that the K-means algorithm performs best in determining the typical load pattern. A further description of the K-means algorithm is provided in Subsection 3.2.2.

2.3.2 Forecasting

Forecasting of energy consumption is an essential part of energy management, system operation and market analysis. Increased accuracy of predictions has the potential to increase savings and create new benefits, as described in Subsection 2.2. There is an emerging demand of customer flexibility in the energy system to increase efficiency, and proper prediction models are a core constituent of such initiatives. (Zhang, 2018) Estimations of energy usage in the long-, medium- and short-term are of importance for planning and investments on the energy market. This becomes particularly visible on the electricity market, where estimations of electricity demand hours or minutes ahead can exert an important influence over the dispatch of national electricity. More precise

predictions can therefore lead to improved energy management and considerable cost reductions for both energy suppliers and end-users. (Wei et al., 2018)

Load forecasting algorithms mostly fall into the subfield of supervised learning according to Zhao et al (2020), as the electrical load is considered a dependent variable. Forecasting models may further be divided into single-valued forecasters, in which the output is a single value, and probabilistic models, which provide a probability distribution of the dependent variable, meaning that the model output contains both an expected value and a standard deviation, which describes the region where the output is likely to appear (Brusaferri et al., 2019). One of the main strengths of the regression based predictors is that the models theoretically are able to learn complex relationships if the data is sufficient (Zhao et al., 2020). There is a vast amount of regression models that can be implemented for forecasting of electricity consumption (Bedi and Toshniwal, 2019). Amongst the successfully implemented regression models are Support vector machine (SVM), Artificial Neural Networks (ANN) and Gaussian process regression (GPR) (Zhao et al., 2020). It is commonplace for regression models to minimize the sum of squared errors or in the case of probabilistic models the marginal likelihood between the output values of the function and the data. This means that a regression model of electricity consumption fits the prediction to match the actual consumption to the highest degree possible. In electricity load forecasting, regression models utilize historical data to enable prediction of future electricity load. The regression models applied for electricity load forecasting differ depending on the application. Parameters that differ are amongst other forecasting horizons (hourly, daily, weekly, monthly and yearly) and the dependent variables (time, weather, historical consumption, etc.). (Yildiz et al., 2017) Regression models are thus able to learn functions from historical data that are able to forecast the electricity in cases where there is a pattern in the electricity consumption.

In a review concerning probabilistic forecasting on electricity consumption van der Meer et al. (2018c) discuss several different statistical methods for forecasting. Amongst the discussed are statistical techniques like quantile regression, Gaussian processes and K-nearest neighbor models. The conclusions of this review are that Gaussian processes might be a powerful procedure to predict systems which are dynamical and nonlinear. Van der Meer et al. (2018c) implement a Gaussian process in combination with historical data points to predict future household electricity consumption. Gaussian processes are given a more in-depth look in Subsection 3.4.

2.3.3 Anomaly detection

According to Seem (2007) the amount of data that the facility managers sometimes must take into consideration is immense. The datasets connected to residential buildings are often too large to consider the totality of the data for human analysis. There are however technologies available to support the facility manager such as alarm and warning systems. The setting of thresholds for these systems is a complex task, if they are set too tight it will generate false faults, if they are set too loose the system does not find all the faults. (Seem, 2007) One of the main reasons to analyze big data regarding the electricity consumption is, according to Zhang et al. (2018), to increase the capability of finding, fixing and isolating faults in a distribution system. The possible reduction of duration of energy consuming faults is also one of the main reasons for analyzing energy consumption data. As stated by Bang et al. (2019) if fault detection is

performed properly, it might be able to determine the characteristics of the fault and thereby assist in correcting it properly.

Fault detection can be divided into three different categories as stated by Kjøller Alexandersen et al. (2019). These different categories are quantitative model-based methods, qualitative model-based methods and process history based methods. Quantitative model-based methods are often obtained from modeling physical behavior of a studied phenomenon according to Kim and Katipamula (2017). Furthermore, these methods, as stated by Bynum et al. (2012), may be based in a detailed or simplified physical model depending on how well the mathematical models represent the reality. Qualitative based models are according to Bang et al. (2019) models based on a priori knowledge. This means that some prior knowledge about the system is needed to determine a model. One of the most commonly utilized qualitative models is the rule based model which often implements multiple if-then statements.

The focus of this thesis is, however, on the process history based models. This model family is purely data-driven and is therefore according to Katipamula and Brambley (2011) one of the more popular models due to the reduced complexity of the model. Bang et al. (2019) describes the fact that process history based models do not take into consideration any physical model of a system or a process, the model instead solely relies on the historical data that is available for analysis. These facts give the model an advantage when the process or system is poorly described by mathematical or physical models. The main disadvantages of the process history based models are according to Bang et al. (2019) the need for an abundance of data, if the available dataset is too small the analysis results would be less reliable. Another disadvantage that the authors bring up is the fact that the data might contain errors, in these cases there is a need for extensive preprocessing.

Machine learning has lately been utilized to a great extent to detect errors and faults in consumption of electricity since deviations can easily be found when big amounts of data are investigated. Clustering techniques can be implemented for finding deviations, some examples of models applied for this purpose include K-means, Gaussian Mixture Models and DBSCAN. (Zheng et al., 2017) With regards to fault detection and diagnosis there are a vast number of different models and different applications. Zhao et al. (2020) mention examples of both supervised and unsupervised machine learning models being used for fault detection and diagnosis. However, these two broad categories contain several different models and due to the wide array of possibilities only a few will be mentioned in this report. Around 20% of fault detection methods that implement artificial intelligence are regression based and 24% utilize unsupervised learning methods, which includes clustering (Zhao et al., 2019). Zhao et al. (2020) mention some weaknesses of using regression models for fault detection. If the underlying data is insufficient then the resulting model might not accurately capture system behavior and the predictions might be a poor representation of the reality. This shortcoming might create situations where a model could detect errors simply because it has insufficient data. Furthermore, they discuss that the data should ideally be labeled to enable an optimal fault detection model. The authors also conclude that instances where the data is labeled correctly are scarce due to the reason that expertise often needs to manually label the data.

The Gaussian process regression mentioned in 2.3.2 is one of several models that can be applied to detect and diagnose faults (Zhao et al., 2020). A Gaussian process regression model is implemented by Van Every et al. (2017) to estimate the different flows of air into buildings with the goal of determining the supply of air needed for the ventilation to detect abnormal ventilation activity. Farshad (2019) describes a method for using the K-means method for fault detection. Farshad utilizes the K-means model's centroids in an essential part of producing a model applied for fault detection. The author compares the cluster centroid with a cluster individual to determine if there exists such a distance between them that it exceeds a predetermined threshold, when it does it is labeled as a fault. These two models, Gaussian processes and K-means, are chosen as the two candidate models examined in this thesis.

2.4 Examples of anomalies

To enable anomaly detection by utilizing machine learning it is initially important to determine what an anomaly is and if it is detectable. While there is an anomaly detection system present at Uppsalahem, it is deemed to be too different in functionality to the algorithms developed within this thesis to be used as evaluation data for the developed models. The evaluation data therefore instead consists of a small set of electricity profiles for which anomalies have been labeled through a simple optical analysis. This paragraph aims to provide some examples of the types of behaviors which are of interest to capture as anomalies. There exist obvious anomalies such as a sudden increase or decrease of the consumption for the substation. Other more subtle behaviors that can be classified as anomalies are a slow and steady increase or decrease of consumption. These anomalies are harder to find since the data might indicate that it is normal consumption even when the drift might indicate a successive decline in building performance.

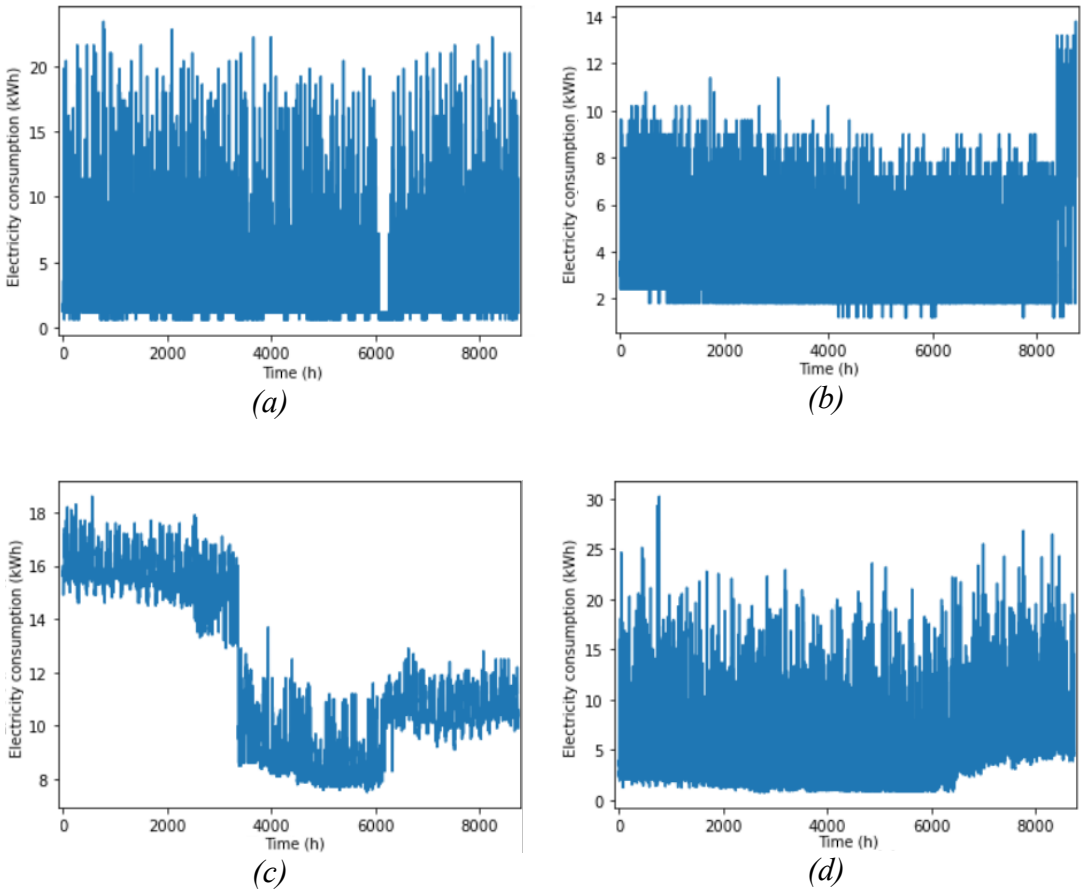


Figure 1(a, b, c, d), illustrations of different kinds of anomalies.

The figure above visualizes different examples of optically identified anomalies that exist in this data set. Figure 1(a) is an illustration of a substation that loses its highest consumption while maintaining its lowest consumption. Figure 1(b) illustrates an anomaly located in the latter part of the year. This is an anomaly that should be easier to find since there is an increase of both the lowest consumption and the highest consumption. Figure 1(c) similarly depicts a clear example of a radical change in base consumption. The last type of anomaly is illustrated in Figure 1(d). The anomaly in this case begins somewhat after hour 6000 when a slow and steady increase of the consumption occurs, commonly referred to as a *drift*. If this anomaly continues the potential energy losses will accumulate over a long time. However, these types of anomalies are difficult to detect since there is no clear difference in the day to day consumption. These anomalies need to be observed on a longer time scale than hours or days as to not give off false detections.

3 Methodology and data

This section presents the data, methodology as well as the theory applied in this project. It begins with Subsection 3.1 which describes the data which is subjected to analysis and the preprocessing steps necessary to enable its use in the models later created. Subsection 3.2 then presents the first model developed, the K-means clustering model, as well as the relevant design choices and metrics of evaluation of that model. The following subsection subsequently describes how the K-means model is applied as an anomaly detection model. Subsection 3.4 presents the second model, the Gaussian process regression model and its evaluation procedure, and the subsection following that proceeds to describe the implementation of anomaly detection based on the Gaussian process regression model.

For the purposes of the study, electricity data for the two most recent full years, 2018 and 2019, was acquired from the electricity service provider E.ON, and in the following methodology a division is made where the first year is considered for training and testing the K-means and Gaussian process models, while anomaly detection is conducted and evaluated on the second year of data.

3.1 The data

To allow for fast detection of anomalies, data should be gathered with a dense time interval. A dataset composed of hourly values of electricity consumption for roughly 600 substations is acquired from E.ON. The range of the data is at its lowest 0 to 107.7 kWh at the highest. The data for the 600 substations is downloaded as a set of roughly 30 excel files, which are then converted into a single Pandas dataframe in Python.

The electricity data is preprocessed in two steps. The first step aims to deal with missing or low time resolution data, and does so either through interpolation of missing or low-resolution intervals in the data, or through the removal of data where these insufficient intervals are too long to be deemed interpolatable. The second step handles low resolution of the values for electricity consumption, and does so through a simple moving average smoothing. These techniques are elaborated further below. Lastly a section is written about the acquisition of data of outside temperature. This data is not deemed to be in need of preprocessing.

3.1.1 Dealing with missing data

For some substations hourly values are only a disaggregation of daily or lower resolution measurements, so that each hourly value is set to a proportional share of the lower resolution measurement. This is not deemed to conform to this study's earlier established need for hourly resolution data. Additionally, many of the data series are missing values for some parts of the 2 years. Interpolation or removal of data series with missing values are two of the most common ways of handling situations with missing data according to Zhang, et al. (2018). Lepot et al. (2017) state that incomplete time-series hinder an optimal analysis of the data and development of models. It is determined that intervals shorter than 336 hours (2 weeks) may be interpolated for the purpose of this study, while all data series with missing or low time resolution data for consecutive intervals longer than 336 hours are removed from the analyzed dataset.

Furthermore, the interpolation should ideally be a function or an algorithm that represents the rest of the observations. Lepot et al. (2017) presents a vast amount of theoretical interpolation methods to use since time-series differ depending on the data source, economical, electrical, financial, etc. According to Beveridge (1992) four criteria need to be met for the option of an interpolation to be available. These four criteria are summarized in the citation below.

> *"(i) not a lot of data is required to fill missing values; (ii) estimation of parameters of the model and missing values are permitted at the same time; (iii) computation of large series must be efficient and fast, and (iv) the technique should be applicable to stationary and non-stationary time series."*

Other than the above-mentioned criteria the model should be robust and accurate (Lepot et al., 2017). There are two steps that need to be taken prior to the implementation of the interpolation method. Firstly, separation of the signal (relevant trend of interest) from the noise to only capture the relevant trends in the data. Secondly, understanding of the present and past data to improve the future forecasting and ability to fill in the absent data points to complete the interpolation. (Musial et al., 2011)

Missing or low-quality data is commonplace in the dataset utilized in this study. In the case when a substation is missing high resolution values there is an individual assessment of the possibility to interpolate the missing data. The data also contains several instances where the values are constant for a period of time. These instances are also subjected to interpolation. Interpolation is performed when there are intervals of constant data longer than 24 hours and shorter than 336 hours (2 weeks) or missing data intervals between 1 and 336 hours. Intervals longer than 336 hours are not deemed to be interpolatable and those data series are instead removed from the final dataset. The anomalies are however retained for the anomaly detection model since they are a vital part for the evaluation of the model.

The function chosen for interpolating data is a sum of a sine function and a linear function given in Equation 1:

$$f(x) = A \times \sin\left(\frac{2\pi}{24}x + \phi\right) + Cx + B, \qquad (1)$$

where A is the amplitude, ϕ is the phase shift, C determines the angle on the curve and B is a constant. This function is chosen over simpler interpolation methods as it better reflects the periodic patterns and variability within the data series. The frequency of the sine function is kept fix to assure the interpolation polynomial has a period of 24 hours (reflecting diurnal patterns in electricity consumption) while the parameters A, C, and B are optimized using a least square fit to the data 24 hours before and 24 hours after the interpolated range.

After a screening of the data the majority of the electricity data series are dropped due to either missing data or lacking hourly resolution data for intervals longer than two weeks. The set of data left after the data preprocessing stage is 193 time series with complete hourly resolution data for the years of 2018 and 2019.

3.1.2 Dealing with low resolution data

Additionally, there are different resolutions in the data of electricity consumption ranging from 0.01 kWh up to 0.6 kWh. This fact presents some problems in analyzing the low resolution 0.6 kWh data. There is a risk that the low-resolution data does not present enough variation to display the more fine-grained temporal changes in electricity consumption. This effect might make predictions harder or in the case of cluster analysis lead to a cluster containing mainly low-resolution data that is similar only because of the data resolution, which does not mirror patterns in actual consumption behavior.

This issue is remedied through smoothing the data. Smoothing is a technique which may be applied to time series data in order to reduce the minute variation between measurements at different time steps. As the smoothing process however also affects the distribution of the data (i.e., the electricity data normally contains momentary variation from one hour to the next, while smoothed data displays an aggregated pattern) the smoothing is applied universally to all data in the acquired dataset. Smoothing additionally has the benefit of reducing noise in the data. As a result of this noise reduction, the basis for pattern analysis is improved. There are however drawbacks as it also leads to some loss of specificity, as very short and sharp patterns may be smoothed out. Zytkow and Rauch (1999) use the moving average as a method for preprocessing. They highlight the fact that moving average smoothing is used to extract periodicity by removing noise from data that is collected in a fixed interval. There are several different algorithms for smoothing but this study uses sliding moving average (SMA). The SMA uses historical data to improve the smoothness of the studied subject. In this study the calculation of moving average is executed using the same method as Hyndman and Athanasopoulos (2018) described in Equation 2:

$$\widehat{T}_t = \frac{1}{m}\sum_{j=-n}^{n} y_{t+j}, \tag{2}$$

where,

$$m = 2n + 1, \tag{3}$$

In Equation (3), *m* represents the number of surrounding values that are used to determine the moving average. The parameter *n* is the width of the window of the moving average, that is, the number of hours before and after the investigated hour taken into consideration when calculating the average. The last notation that needs to be considered is the variable *t*, which is the location of the value currently being smoothened. Following the steps described in this subsection and Subsection 3.1.1 the complete preprocessing procedure is depicted in the flowchart seen in Figure 2.

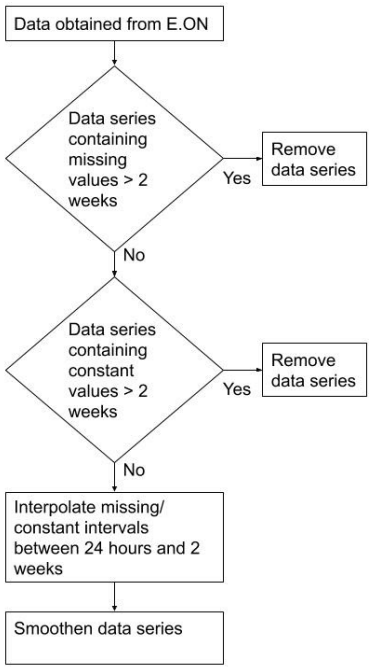


Figure 2, a flowchart depicting the complete preprocessing procedure

3.1.3 Weather data

Electricity consumption is impacted by multiple weather variables according to Yang et al. (2018). Auffhammer et al. (2017) mention that much previous research has been done on the relationship between electricity demand and outside temperature, and in a study of general grid demand establish that electricity consumption is responsive to temperature on a daily time frame. Within this study a set of models is therefore also created utilizing outside temperature as a dependent variable, for which a set of hourly weather data for the 2-year period of study is acquired from the Swedish meteorological and hydrological institute (SMHI) and added to the analysis.

Additionally, it is established in this project that electric heat pumps are connected to a share of the studied substations. Heat pumps use electricity for purposes such as powering radiators and heating tap water, and have a major impact on electricity load profiles compared to other heating systems such as district heating. Therefore, a correlation between lower temperature and higher electricity consumption might be expected. Thus, a model that takes the outdoor temperature into consideration should theoretically perform better than a model that does. The outdoor temperature data implemented in the model is data downloaded from SMHI (SMHI, n.d.). The utilized data consists of hourly measurements from one weather station in Uppsala. The data stretches over the same period and retains the same resolution as the electricity consumption data.

3.2 K-means model

The first anomaly detection method developed in this study is based on the principle of clustering. The clustering algorithm chosen is K-means clustering, in which the model is represented by a set of centroids, the mean value of each cluster, and every data series is grouped to one of these centroids based on proximity and centroids are then

recalculated in an iterative process. The K-means model requires the selection of the desired number of clusters, *K*, before execution. The resulting clusters represent a grouping of electricity profiles based on similarities in their behavior. To ascertain that the results of the K-means clustering represent patterns, rather than differing scales of the clustered data, the time series data is first subjected to Z-score normalization, where all data is scaled to have the same mean and variance, which is described in Subsection 3.2.1. The following subsection, gives a more in-depth description of the K-means model. The following two subsections then describe two common ways of determining the optimal number of clusters for K-means, the Elbow method and the Silhouette index. The final subsection offers a short description of how these validation scores are implemented to come to a conclusion about the final model.

3.2.1 Z-score normalization

One of the main problems that occur when developing a cluster based model for detecting anomalies on individual substation data is that the scales of the electricity consumption profiles are different from one substation to the other. Distance-based classification algorithms such as K-means are very likely to be affected by normalization, as they are built on the idea of calculating the distance between different entries. (De Jaeger et al., 2020; Viegas et al., 2016) For the purposes of this study, the clustering results should ideally not be impacted by scale differences in the baseline consumption, but only patterns and relative consumption changes which appear abnormal.

There are several alternatives for normalization of data (Cheng et al., 2019), this study, however, uses Z-score normalization as it takes into consideration and deals with data containing outliers. Also known by the name standardization, Z-score normalization transforms the input data by subtracting the mean of each feature from the original values and normalizing it to have unit variance (Gasser,2020). The process can according to Zhang (2019) be described by Equations 4-6:

$$\hat{X}[:,i] = \frac{X[:,i]-\mu_i}{\sigma_i}, (\mu_i, \sigma_i), \tag{4}$$

$$\mu_i = \frac{1}{N}\sum_{K=1}^{N} X[k,i], \tag{5}$$

$$\sigma_i = \sqrt{\frac{1}{N-1}\sum_{K=1}^{N}(X[k,i]-\mu_i)^2}, \tag{6}$$

where *X[:,i]* represents the feature at position *i*, and μ_i, σ_i are the mean and variance of that feature for the dataset. The normalization of the data utilizes the data for 2018 to transform all the 2018 time series to have mean zero and unit variance to enable clustering. The 2019 data is then normalized based on the mean and standard deviation of the 2018 data. The reasoning behind this is that clusters should be created using the first year of data, and data from the second year should not affect the clustering process.

3.2.2 K-means clustering

After the time series data is scaled to have similar means and variances it is ready for clustering. The K-means model is chosen for clustering the data. K-means is an iterative

method based on mean values of the data which divides the data into a set number, *K*, of clusters (Tan et al., 2018 p. 535). The K-means model is summarized in four main steps as:

1) To initiate the algorithm, *K* initial centroids are introduced within the dataspace at random coordinates.
2) Each datapoint's distance (normally measured as Euclidean distance) to the centroids is calculated and assigned to the centroid with the shortest distance metric, creating the clusters.
3) New centroids are calculated based on the mean value of the coordinates of all the data points in respective clusters.
4) Steps 2 and 3 are repeated until the assignment in step 2 stops changing between iterations, meaning that the clusters have converged.

Due to the way clusters are randomly initialized, K-means is considered a stochastic method, meaning that it does not yield exactly the same results every time the algorithm is executed (Tan et al., 2018 pp. 539-41). Due to this fact the initialization of the centroids also becomes important to the end result.

The K-means model has one hyperparameter, *K*. Hyper-parameters are parameters chosen by the model creators. Since this choice affects the result of the model the choice of hyper-parameters is an important decision in the creation of machine learning models. (Burkov, 2019)

The hyper-parameter *K* determines the number of clusters created at initialization. *K* must be a positive integer and its value may not exceed that of the total number of points in the dataset. Sometimes the selection of *K* might be inferred from the context of the problem being studied, but at other times the optimal number of clusters might not be clear. In those latter cases, multiple values for *K* may be tried iteratively and the different resulting models evaluated. (Tan et al., 2018 pp. 539-41)

This study calculates several common measurements of clustering performance, the silhouette index, the within-cluster sum of squared errors (WSS) as well as within-cluster R^2, to determine the optimal value of *K*. The final decision is based on the WSS and R^2 and utilizes the elbow method to determine the optimal value of *K*. However, for the sake of comparison the silhouette index values for the respective values of *K* are also presented.

3.2.3 Elbow method

The elbow method plots a performance metric against an investigated value of model parameters, which is deemed to be a measurement of model complexity, resulting in an elbow point diagram. According to Masud et al. (2018) it is a well-known method for determining the number of clusters in a data set. Govender and Sivakumar (2020) states that the WSS is used to calculate the total sum of squared errors between each data point in the cluster and the cluster centroid. It is defined in Equation 7:

$$WSS = \sum_{J=1}^{K} \sum_{i \in C_j} \left\| x_i - c_j \right\|^2, \quad (7)$$

where C_j is the *j*th cluster object set and the amount of cluster is represented as K, x_i is the *i*th data point clustered to C_j and c_j is the centroid of the *j*th cluster (Govender and Sivakumar 2020).

As the WSS is calculated for different values of *K* there will be a point where the value WSS does not decrease substantially. Thus, there exists a point where further increase of the number of clusters does not make the model significantly better, illustrated in Figure 3 as "Elbow point".

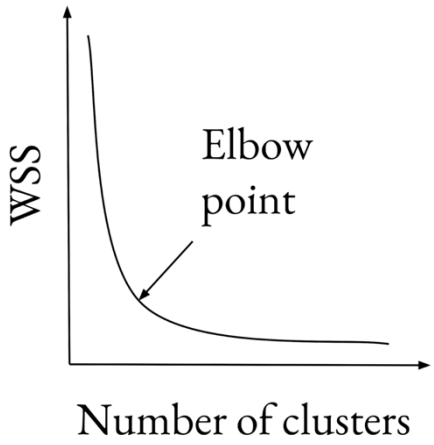


Figure 3, an illustration of the elbow point diagram for WSS, inspired by Pimentel and de Carvalho (2020) and Masud et al. (2018)

Another metric that can be utilized is the within-cluster R^2-value. Unlike the WSS it is a mean of the individual measurements to ease comparisons. It is defined in Equation 8:

$$Within\ cluster\ R^2 = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\#C}\sum_{i\in C_j}\left(1-\frac{\sum_{t=1}^{N}(c_j(t)-x_i(t))^2}{\sum_{t=1}^{N}(c_j(t)-\overline{x_i}(t))^2}\right), \qquad (8)$$

where *K* is the number of clusters, c_j and x_i are described as in Equation 7 and *N* is the length of x_i.

The optimal number of clusters is located at the x-axis at the elbow point in the diagram. Towers (2013) states that the elbow point is often determined by visually analyzing the diagram. Esteri et al. (2018) also highlights the fact that the elbow point can be hard to determine even for expert judgements due to the reason that it is executed optically. One issue with this method is when the investigated value increases for every cluster, then a single apparent elbow point might not be established (Masud et al., 2018; Pimentel and de Carvalho, 2020). In such an event, a sufficient elbow point is established to enable a specification of the number of clusters.

3.2.4 Silhouette index

The second performance metric calculated for the cluster model is the silhouette index. When a clustering model's ability to discover clusters in a dataset is evaluated the two criteria of interest are the compactness and separation of the clusters found (Tardioli et al., 2018). A good compactness is obtained when a cluster's data points are close to each other and the distance between them are small. The separation of clusters regards

the distance between different clusters. A large distance between clusters means that the clusters are well defined and that there is a distinct difference between the clusters according to Tardioli et al. (2018). There are several validation indices measuring these properties that are useful for determining the success of a clustering technique. For the purpose of this study the silhouette index is chosen due to its relatively simple formulation and common usage.

The silhouette index measures the ratio between the separation and the compactness of a cluster. The ratio varies from -1 to 1. When the cluster has a good partition the index is close to 1. For a single data point *i* the silhouette index is described by Equation 9:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}, \tag{9}$$

where *a(i)* is the average dissimilarity comparing the data point *i* to all the other data points in the same cluster. Meanwhile *b(i)* is the lowest measure of dissimilarity between *i* and any data point which is not a member of the same cluster. To evaluate the silhouette index for the clustering as a whole all the silhouette indices must be considered. Equation 10, represents the overall silhouette index for all clusters and is defined as

$$S = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\#C}\sum_{i\in C_j} s(i), \tag{10}$$

where *K* is the number of clusters that is chosen and C_j is described as in Equation 7. This means that a single value of the silhouette index is determined for all the clusters. (Tardioli et al., 2018)

3.2.5 Implementation and validation

This project utilizes the implementation of the K-means model included in the Scikit-learn package for Python. A set of models are fitted to the data, varying the hyperparameter value of *K* for each run, for values of *K* in the range between 2 and 20. The WSS as well as within cluster R^2 are calculated for each run and subsequently displayed in an elbow diagram. The silhouette index is also calculated for each value of *K* and displayed in a similar diagram.

After the initial clustering is complete there is a division of all clusters containing five or less data points and redistribution of their members into the remaining clusters. If this is the case the model then removes the centroid of the cluster deemed to have too few members, and then repeats the clustering algorithm using the remaining cluster centroids from the initial run as the initial centroids. The result of this process is that the data points which were grouped to the removed cluster centroid are dispersed into the other clusters while the remaining clusters remain roughly the same. This procedure is conducted to ensure that each cluster has enough members to retain its overall shape even if drifts or anomalies happen in the individual data series.

3.3 K-means model anomaly detection

The anomaly detection model compares the values of the actual measurements from the individual substation to the mean value of the substation's assigned cluster in the

clustering model. The amount of anomaly detection models is therefore equal to the number of clusters. For the clusters, the anomaly detection model is the mean consumption of the specific cluster, described in Equation 11:

$$M_C(t) = \frac{1}{\#C}\sum_{i \in C_j} x_i(t), \tag{11}$$

where *C* is the studied cluster and #C is the size of the cluster. The variable $x_i(t)$ represents the electricity consumption for the *i*th substation in cluster *C* at the hour *t*. This function $M_C(t)$ is used as the comparison for detecting anomalies in all substations which have been grouped to that cluster.

The question of how to define a fault has been discussed by multiple authors (Kjøller Alexandersen et al., 2019 and Bang et al., 2019). They argue that a fault could be described as when the consumption would deviate more than a predetermined bound. The bounds aim to take the uncertainties of the model into consideration. The upper bound and the lower bound illustrated in Figure 4 are located the same distance from the model at all times as it is a fixed value due to the normalization mentioned in Subsection 3.2.1.

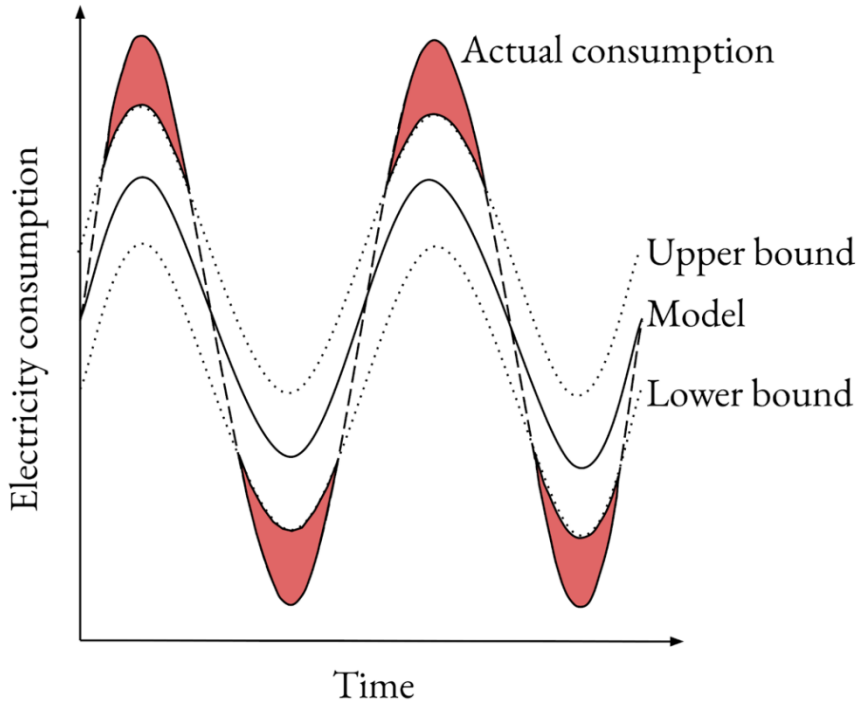


Figure 4, an illustration of the consumption where several anomalies should be detected (the red areas), inspired by Kjøller Alexandersen, et al. (2019) and Bang, et al. (2019)

The red area from Figure 4 can be interpreted as the size of the deviation, when the deviation size is large enough it is detected as an anomaly. As can be seen in the figure an anomaly is detected not only when the consumption is too high, but also too low. This is an important step to ensure that all possible anomalies are found. The deviation size is mainly chosen because it takes into consideration large anomalies that occur under a small period of time and small anomalies that occur under a long period of time. (Bang et al., 2019) The deviation size between two points in time T_1 and T_2 is defined in Equation 12:

$$Deviation\ size(T_1, T_2) = \sum_{t=T_1}^{T_2}(x(t) - M_C(t) + B)), \quad (12)$$

where $x(t)$ is the actual consumption at time *t*, $M_C(t)$ is the model described in Equation 12 and *B* is the above described bound. Furthermore, there is also the possibility that there is an anomaly below the consumption, it is defined in Equation 13:

$$Deviation\ size(T_1, T_2) = \sum_{t=T_1}^{T_2}((M_C(t) - B) - x(t)), \quad (13)$$

where every parameter is defined as for Equation 12.

The comparison is executed for all values of *t* in the range from 1 to 8760 and the predicted consumption and from there iterates over all hours in the "unseen" data. The difference between the substation's actual consumption and $M_C(t)$ is then utilized to determine if an anomaly is detected, and a minimum deviation size, *L*, is implemented to determine when an anomaly has occurred. If the consecutive deviation is larger than a predetermined size it is reported as an anomaly. This leaves the anomaly detection model with two parameters to set in order to define an anomaly, the bound *B* and the size limit to determine an anomaly, *L*.

3.4 Gaussian process regression model

The second model of this thesis applies a regression model to make predictions which enable anomaly detection. In order to allow for sufficient predictions, the regression model utilizes training data to learn electricity consumption as a function of the time and eventually also temperature data. The specific regression model used is Gaussian process regression. The section begins with an introduction to Gaussian processes in Subsection 3.4.1. The two most important choices in the development of a Gaussian process regression model are the choice of kernel function and dependent variables, which are expanded on in the two following subsections. Subsection 3.4.4 describes dynamic Gaussian processes, which is a specific implementation utilized in this study to cope with computational complexity and learn seasonal trends in the data. Subsection 3.4.5 then describes the relevant metrics for measuring the performance of prediction models. The final subsection describes the practical implementation of the model and how the combinations of kernel function and dependent variables are selected from a cross-validation procedure.

3.4.1 Gaussian process

The Gaussian process is a non-parametric probabilistic tool which can be used to model non-linear functions. It may be utilized to solve classification as well as regression problems, the latter of which is the application for this study. A Gaussian process is a generalization of the Gaussian probability distribution, also known as the normal distribution. This distribution is parametrized by its mean and standard deviation (van der Meer, 2018a), in accordance with Equation 14:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{\sigma 2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad (14)$$

The Gaussian process is then formally defined as "a collection of random variables, any finite number of which have a joint Gaussian distribution" (Rasmussen & Williams,

2006). As the Gaussian process is used to model a function, the assumption then becomes that for all pairwise combinations of two function values, *f(x)* and *f(x')*, where *x* and *x'* are any two distinct values of *x*, they are jointly distributed according to Equation 15:

$$p\left(\begin{bmatrix} f(x) \\ f(x') \end{bmatrix}\right) \sim \mathcal{N}(\mu, K), \tag{15}$$

where μ is the mean and *K* the covariance matrix, which is defined as the matrix of all pairwise outcomes of the covariance function of any two values of *x*. The Gaussian process is then defined by its mean function *m(x)* and the covariance function *k(x, x')*, representing the real process *f(x)* resulting in the following connection in Equation 16-18:

$$m(x) = E[f(x)], \tag{16}$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))], \tag{17}$$

$$f(x) \sim GP(m(x), k(x, x')), \tag{18}$$

Thus, the random variable described by the Gaussian process is a representation of the value of *f(x)* at the point *x*. Gaussian processes are commonly considered as having a zero-mean function, albeit this is not strictly necessary. The covariance function *k(x, x')*, also called kernel function, specifies how the function value *f(x)* is related to other values *f(x')*.

For a vector of *n* points $X = \{x_1 \ldots x_n\}$ the distribution over functions described in Equation 14 is then characterized by $\mu^* = \mathbf{0}$ and the covariance matrix *K* which is seen in Equation 19:

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}, \tag{19}$$

The kernel function is an integral part of Gaussian processes as it determines the properties of the modeled function *f(x)*. There are a multitude of potential kernel functions to choose from when creating a Gaussian process (Rasmussen & Williams, 2006), some of which will be further described in Subsection 3.4.2. The kernel function is defined in such a way that the Gaussian process may also be generalized to unobserved function outputs, henceforth denoted as *f(x*)*. Assuming a zero mean, the joint distribution over functions may then be written as in Equation 20:

$$p\left(\begin{bmatrix} f(X) \\ f(X^*) \end{bmatrix}\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(X, X) & k(X, X^*) \\ k(X^*, X) & k(X^*, X^*) \end{bmatrix}\right), \tag{20}$$

The predictive distribution in unseen data points conditioned on observed data may then be obtained through Equations 21:

$$p(f(X^*)|X, f(X), X^*) = \mathcal{N}(\mu^*, \Sigma^*), \tag{21}$$

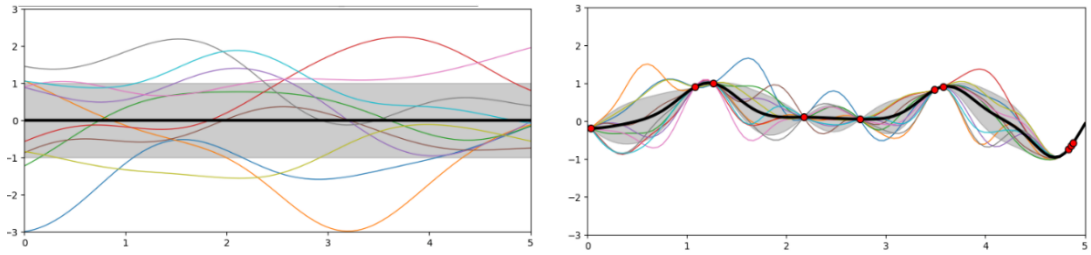
where the prediction's mean μ^* and covariance matrix Σ^* are described in equations 22-23:

$$\mu^* = K(X^*, X)[K(X, X) + \sigma^2 I]^{-1} f(X), \quad (22)$$

$$\Sigma^* = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma^2 I]^{-1} K(X, X^*) + \sigma^2 I, \quad (23)$$

Thus, for all predictions, the Gaussian process yields both an expected value (i.e. the relevant prediction) and a standard deviation, which may be used to establish a degree of confidence for the predictions. The latter is perceived to have a possible utility in the detection of anomalies, which is one of the reasons behind choosing a probabilistic model. (van der Meer, 2018a)

A benefit of utilizing a probabilistic model is that it allows for the specification of a prior distribution over functions, which expresses a belief about the system behavior imposed by the researcher before any data has been observed. Once a set of data has been observed this distribution is fitted to only include functions which pass through (or close enough to) the observed data points, resulting in the posterior distribution over functions. (Rasmussen & Williams, 2006) This posterior has reduced uncertainty in points close to the observed data points, as seen in Figure 5.



(a) *(b)*

Figure 5, illustration of the prior distribution (a) and posterior distribution fitted to a set of observations (b) of a Gaussian process using the Squared Exponential kernel. (image source: Scikit-learn, n.d.)

The training of the Gaussian process is inductive in the manner that it utilizes a limited set of data to adapt a function that can be used to predict all possible input values. The kernel function is parameterized by one or several hyper-parameters, θ, commonly written *k(x, x'; θ)*. These hyper-parameters are what allows for flexibility in how the data is modeled. The learning process for a Gaussian process model consists of learning these hyper-parameters from some training dataset. A benefit of working with Gaussian processes is that hyperparameters may be inferred directly from the training data, which reduces the need for cross-validating in order to infer values for the different hyper-parameters. This is performed by maximizing the log marginal likelihood with respect to θ (van der Meer, 2018) which is described in Equation 24:

$$\log p(f(X)|X, \theta) = -\frac{1}{2}\big(f(X) - m(X)\big)^T (K(x, x) + \sigma_n^2 I)^{-1}\big(f(X) - m(X)\big) - \frac{1}{2}\log|K(x, x) + \sigma_n^2 I| - \frac{n}{2}\log 2\pi, \quad (24)$$

This training process requires the inversion of this covariance matrix, which is a task of cubic complexity, making the training process computationally expensive as the number of observations in the training data grows large (Rasmussen & Williams, 2006). In the

Scikit-learn implementation of Gaussian processes, this optimization is carried out by the gradient based algorithm *L-BFGS-B* (Byrd, 1996) by default. There is however no guarantee that the marginal likelihood does not suffer from multiple local optima. It is normally not a problem for simpler kernel functions, but local optima may nevertheless exist and affect the resulting posterior. The Scikit-learn implementation of Gaussian processes also supports the implementation of other optimization algorithms through the passing of a function or the choice of not optimizing the hyper-parameters in order to make predictions based on the Gaussian process prior. (Scikit-learn, n.d.)

In creating a Gaussian process model the main choices left to the machine learning engineers are the choice of kernel functions and dependent variables. These choices are largely based on some prior knowledge of the dataset, established either through previous research or empirical study of the data. (Rasmussen & Williams, 2006) In this case conceptual knowledge about typical profiles of electricity consumption become important for making these choices.

3.4.2 Kernel functions

There is a multitude of kernel functions to choose from when creating a Gaussian process model, some relevant examples of kernel functions for this study include:

- Squared exponential kernel

Also called the radial basis function (RBF) kernel, is defined in Equation 25:

$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right), \tag{25}$$

This function has one parameter l defining the characteristic length-scale. The length-scale parameter can be understood as regulating how close in the feature space another point must be for it to have a significant impact on the distribution of values in x. The RBF function is infinitely differentiable, which means that a Gaussian process with this kernel function has mean square derivatives of all orders, making it very smooth. The RBF kernel is often cited as the most widely used kernel function in the field of kernel machines. (Rasmussen & Williams, 2006)

- Matern kernel

This kernel has the form that is described in Equation 26:

$$k(x, x') = \frac{1}{\Gamma(v)2^{v-1}}\left(\frac{\sqrt{2v}}{l}|x, x'|\right)^v K_v\left(\frac{\sqrt{2v}}{l}|x, x'|\right), \tag{26}$$

In addition to the length-scale parameter of the RBF kernel it has a parameter v which controls the smoothness of the resulting function. It is a generalization of the squared exponential function and converges to the RBF kernel as v approaches infinity. Stein (1999) argues that the Matern class yields more realistic models of real-world phenomena as it does not enforce the same smoothness as the RBF kernel. (Rasmussen & Williams, 2006)

- Exp-Sine-Squared kernel

This kernel has the form that is described in Equation 27:

$$k(x, x') = \exp\left(-\frac{2\sin^2(\pi|x,x'|/p)}{l^2}\right), \quad (27)$$

The Exp-Sine-Squared kernel allows for the modeling of periodic functions with a behavior which repeats its values in regular intervals. It has the length-scale parameter *l* and a periodicity parameter *p*, where the length-scale parameter fills the same function as in the above described kernels and the periodicity determines the distance between repetitions of the function. (Rasmussen & Williams, 2006)

- White Kernel

This kernel has the form which is defined in Equation 28:

$$k(x, x') = \sigma^2 I_n, \quad (28)$$

which explains the noise of the output signal as independently and identically normally-distributed. While it is of no use to model by itself it is an important part of the sum kernels utilized in this project as it allows the model to fit noise to the data when the signal is noisy. (Scikit-learn, n.d.)

Additionally, it is possible to create new kernel functions through addition and multiplication of existing functions (Rasmussen & Williams, 2006). The multiplication or addition of two positive semidefinite kernels always result in another positive semidefinite kernel. Duvenaud (2014) mentions some examples of combining kernel properties through multiplication. Among them is the "locally periodic kernel", acquired through the multiplication of the periodic Exp-Sine-Squared kernel with a Squared Exponential kernel. This kernel function results in functions over the data which have the property of periodicity but may also vary slowly over time. (Duvenaud, 2014)

3.4.3 Choice of dependent variables

When designing a regression model the choice of dependent variables is of paramount importance. Yang et al. (2018) propose a Gaussian process quantile regression model for making 1 hour ahead predictions of electricity consumption. Their dependent variables include calendar variables, weather conditions, electricity prices and historical load data. Thus, they arrive at the relationship described by Equation 29:

$$\hat{y} = f(t, d, v_l, v_t, p), \quad (29)$$

where $\hat{y}$ is the approximation of electricity consumption, where *"$t \in [0,24]$ is the hour of day, $d \in \{1,2, \dots ,365,366\}$ is the day of the year, v_l is a vector of the historical power load values, v_t is a vector of weather variables like temperature, p is the real-time price."* (Yang et al., 2018)

In order to infer relationships between electricity consumption and the hour of the day without utilizing a periodic covariance function, the time data is also mapped to a plane,

where each point in time may be described as a coordinate consisting of the day of the year and the hour of the day. As the original time resolution of the data is hourly, the conversion of the hour of the year into day of year and hour of day respectively could be achieved by the simple function described in Equation 30:

$$f(t) = (\left\lfloor \frac{t}{24} \right\rfloor, t - 24\times \left\lfloor \frac{t}{24} \right\rfloor), \tag{30}$$

where $\left\lfloor \frac{t}{24} \right\rfloor$ is the Euclidian division where t is the numerator and 24 the denominator.

This creates a coordinate system where the point (3,13) occupies a position closer to (4,13) than for instance (4,20), which also determines how much weight the Gaussian process regression model places on the different observations when making a prediction. Figure 6 shows an example of the considered distance between the observed data points and the point of interest for a prediction with a 24-hour horizon.

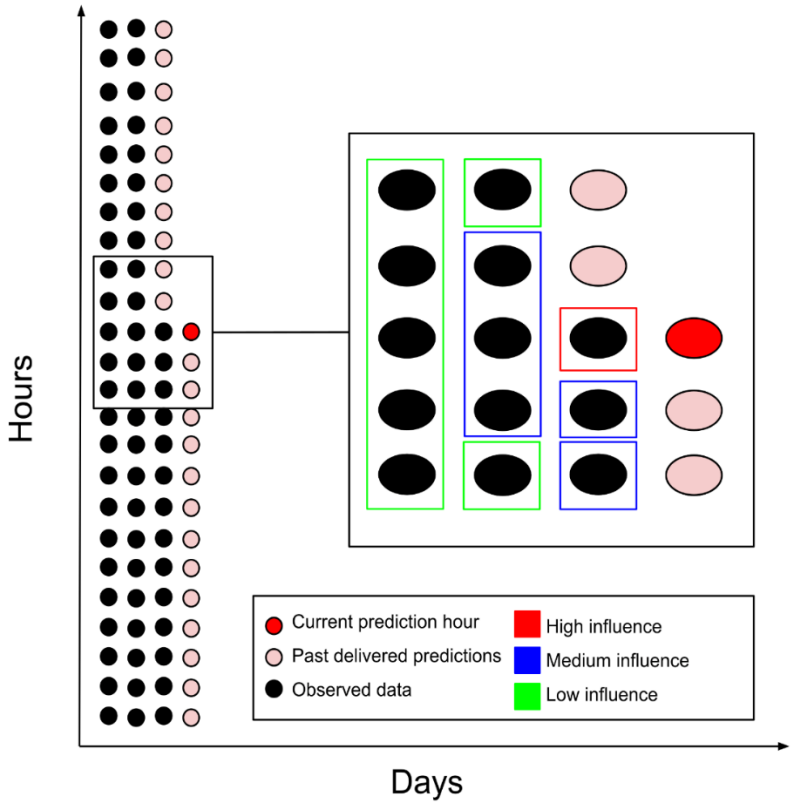


Figure 6, illustration of which data points the prediction regards.

The red point in Figure 6 is the point that the model aims to predict. The model takes the closest points in the plane into consideration when trying to predict the upcoming load. As can be seen in Figure 6 the model takes the point in the red square into most consideration, afterwards, the points in the blue squares are taken into consideration. The points that impact the predictions the least are the points in the green squares.

There are three alternative configurations of dependent variables tried in this study:

1) $\hat{y}$=f(t), where t is the specific hour of the two-year time period studied. This is used exclusively in combination with the periodic ExpSineSquared kernel.

2) ŷ=f(h, d) where $h \in [0,24]$ is the hour of the day and $d \in \{1,2, \ldots ,730\}$ is the day of the two year period studied.
3) ŷ=f(h, d, w) similar to 2 but with the added value of the outside temperature as *w*.

3.4.4 Measurements of model error

As the project is partly based on the utilization of regression models it is crucial to determine how accurate the developed models are. If a model is an insufficient fit it is not an adequate representation of the reality which in turn offers problems for the project. Therefore, four different, well known metrics are utilized, mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and R^2. The four metrics calculate the error and correlation between predictions and observed values and are further described in this subsection.

Mean absolute error
Mean absolute error (MAE) is the first measurement of model errors regarding forecasting. Van der Meer et al. (2018c) utilizes MAE to assess the performance of the forecasting model. Furthermore, they define it in Equation 31:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\widehat{y_t} - y_t|, \tag{31}$$

where *T* is the length of the time series, y_t is the measured value and the forecasted value is $\widehat{y_t}$. A lower value indicates a better performing model.

Mean absolute percentage error
Mean absolute percentage error (MAPE) is utilized by van der Meer et al. (2018c) as a performance metric to evaluate the predictions. Further it is defined in Equation 32:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\widehat{y_t}-y_t}{\max(y)-\min(y)}\right|, \tag{32}$$

where *max(y)* and *min(y)* represent the maximum and minimum value of the time series respectively. All other parameters are defined as for MAE above.

Root mean squared error
One of the most common approaches to determine the performance of a forecasting method is according to Bourdeau et al. (2019) the RMSE. The model is described in Equation 33:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2}, \tag{33}$$

where *N* is the number of values in the measurement series and $\widehat{y_i}$ is the model's predicted value at index *i* and y_i is the true value. A low RMSE value indicates a well fitted model. Depending on the implementation the value that is determined to be a good value varies, it is therefore difficult to set a fixed bar on what is a good value.

R-squared
The coefficient of determination or the R^2 measurement method is according to Cheng

et al. (2014) utilized to measure how well a regression model is fitted to a sample of observations. Furthermore, it is the squared value of several correlation coefficients between the model and the observations based on the studied sample. Bourdeau et al. (2019) defines the equation for the R^2-values as described in Equation 34:

$$R^2 = (1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}), \tag{34}$$

where $\hat{y}_i$ and y_i are defined in the same way as in the above paragraph. The variable $\overline{y}$ is the mean value of y_i for the whole year. According to Elsheikh et al. (2019) the R^2-value is utilized for determination of resemblance between two different time series, in this case a forecasted time series and an observed time series. Generally speaking, the better the value, the closer to 1 it is. Thus, a high number indicates a good forecasting model or method (Elsheikh et al. 2019).

3.4.5 Dynamic Gaussian process regression

Gaussian processes regressors for time series data may be trained iteratively in order to make one-step-ahead predictions, resulting in a dynamic Gaussian process model. The dynamic Gaussian process model is updated iteratively using a moving window. Girardi et al. (2003) conclude in their study that an iterative Gaussian process is often less computationally demanding, since it is being updated using a shorter length of training data instead of the whole available set. Van der Meer et al. (2018b) find that the dynamic Gaussian process approach produces sharper prediction intervals in a study of residential electricity data while also bringing a significant reduction to computational demand compared to a static Gaussian process regression model, but with the drawback that it also is less capable of predicting sharp peaks in electricity consumption.

This project opts for dynamic Gaussian processes primarily to cope with computational demand, as the training process for the static Gaussian process becomes very computationally expensive as the amount of training data considered grows. Additionally, the dynamic Gaussian process may benefit from an increased ability to learn seasonal patterns in electricity consumption, as the hyper-parameters are learned specifically from the local data close to the current time. When creating a dynamic Gaussian process for this project, a choice is made that the model hyperparameters should be updated based on 4 weeks of data. Thus, a call is made to the optimizer once every 672 hours (or 4 weeks) to update the model hyperparameters based on the latest 672 observations. Shorter intervals were originally considered but led to widely varying results from the hyperparameter optimization. The remaining time, the hyperparameters from the last model update are treated as a model prior and new data points are fitted to it without updating the hyperparameters. The amount of data considered for inferring the next prediction is always limited to the past week for computational reasons.

3.4.6 Implementation and cross validation

As the hyper-parameters in the dynamic Gaussian process are inferred from the training data as described in Subsection 3.4.1, the cross-validation procedure in this project instead serves to evaluate the different options for kernel functions, as well as the possible configurations of the dependent variables. Different combinations of kernel functions and dependent variables are therefore tried for a static Gaussian process

implementation. In this case the time horizon is 24 hours, which means that the model may consider observations up until 24 hours before the point being predicted, but not the 24 hours immediately preceding the point being predicted. More accurate predictions may be attained at an hourly prediction horizon, however that would also make it increasingly difficult to find anomalies due to the fact that predictions would largely be based on the observation of the previous hour, leading to a lack of differentiation between the predictions and the anomalies as the model also predicts the anomalies to a high extent. However, in the daily basis time horizon there exists a clear difference between the predictions and the expected consumption if an anomaly occurs. A weekly time horizon basis was also considered but deemed to be too inaccurate to yield sufficient predictions to enable a viable anomaly detection model.

This study utilizes a blocking based cross validation for time series (Racine, 2000) which is specifically developed to handle the issues of dependency between observations in time series data. Cross validation is conducted on the first half of each data series (consisting of the year 2018) so that anomaly detection models using the kernels and dependent variables chosen from the cross-validation step may then be deployed for the 2019 data. The chosen number of splits for the validation procedure is 6 where each split is divided into 50 % training and 50 % test data. These values are chosen as they mirror the functionality of the dynamic Gaussian process model described in Subsection 3.4.4.

The blocking cross validation procedure consists of splitting the validation data into equally sized blocks, for which each block is then split into two separate blocks of training and test data as illustrated in Figure 7. The Gaussian process regression model then learns hyperparameters from the training data block and subsequently makes 24-step ahead predictions for every hour in the test data set. The aforementioned performance metrics of MAE, RMSE and R^2 are then calculated for the predictions.

In this study, cross validation is conducted individually for each data series that is considered for the anomaly detection algorithm. As this procedure is computationally expensive, an exhaustive cross validation is not conducted for each data series in the acquired dataset.

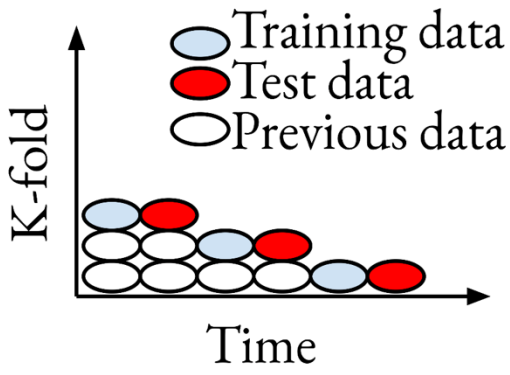


Figure 7. A representation of the K-fold cross validation procedure for time series. (Inspired by van der Meer et al., 2018b)

3.5 Gaussian process regression anomaly detection

The Gaussian process anomaly detection model differs from the K-means based anomaly detection model in a number of ways. Firstly, instead of setting a static bound for the full year, the gaussian process utilizes the standard deviation of the model to attain the bounds for comparing the model and actual data. This allows the anomaly detection model to dynamically compensate for varying degrees of certainty in predictions during the whole year. The anomaly detection model then multiplies this standard deviation by some constant which may be defined for the individual case, e.g. 1.96 to consider only observations outside the 95 % confidence interval as contributing to the anomaly detection. Also, the Gaussian process predictions often remain close to the observations even for what is considered to be abnormal data. The model's inability to accurately predict the variance of effect peaks would also often interfere with the process of finding long-term anomalies. Therefore, the anomaly detection model opts for trying to find situations where a big enough share of the observations is outside of the prediction bounds instead of looking for continuous intervals where the model output and data differ in a way that's indicative of over- or underconsumption. This however requires the model to restrictively consider some interval when calculating this share of abnormal observations. This results in a set of three configurable parameters for the second anomaly detection model; The constant for defining the confidence interval of the bounds, *s*, the length of the interval the model must consider in order to detect an anomaly, *w*, and finally the share of observations within said interval that must be outside the bounds for the data to be considered an anomaly, *p*. Thereby, the criterion for finding abnormally high consumption at a given time *t* can be defined as

$$\sum_{t-w}^{t} sign(y(t) - (E[f(t^*)] + s\sigma(f(t^*)))) > w \times p, \tag{35}$$

where *y(t)* is the observed consumption in *t* and $f^*(t)$ is the Gaussian process estimate delivered on a 24-hour prediction horizon. In similarity to the clustering based model there is another mirror function for unexpectedly low consumption:

$$\sum_{t-w}^{t} sign(y(t) - (E[f(t^*)] + s\sigma(f(t^*)))) < w \times p, \tag{36}$$

Additionally, since the Gaussian process estimates are based directly on the events during the previous 24 hours, while the timeframes for detecting anomalies are often significantly longer, a slight modification is made to the earlier described prediction model. If the model is allowed to make estimations based on abnormal data, these estimates are likely to mirror abnormal behavior, preventing the detection of anomalies longer than the prediction horizon. The anomaly detection for the Gaussian process model therefore requires the regression model to receive feedback from the anomaly detection model, i.e. when the anomaly detection model is in the process of detecting an anomaly, the regression model may not use those data points to make new estimations. Instead, the model then makes predictions based on the latest "good" data, resulting in a longer prediction horizon in those cases. As the prediction horizon grows longer, the insecurity in the predictions, represented by the standard deviation, increases, meaning the bounds for the anomaly detection gradually become broader towards the end of the interval. This allows the model to retain relatively strict bounds during normal consumption and near the beginning of the anomaly detection interval, making it preferable to always making estimates based on a longer prediction horizon.

4 Results and analysis

This section presents a detailed analysis of the results that have been obtained and analysis of the same. The results section begins with a thorough examination of the first implemented model, the K-means model and conclusions about the representativity of the clusters. Thereafter the Gaussian process regression model and an evaluation of its forecasting abilities on the data set is presented. Finally, the results on the implementation of the anomaly detection models are presented in two subsections, one discusses the anomaly detection model based on clustering and the other elaborates on the results from the Gaussian process based anomaly detection model. The aim is to evaluate the results from these models using different examples and illustrations from the studied data set. Some criticisms of the perceived flaws and insufficiencies of the respective models are also presented in these paragraphs. Finally, a comparison between the developed models is executed, in which their respective strengths and weaknesses are analyzed.

4.1 K-means model

In an iterative process the clustering model is implemented for values ranging from 2 to 20 representing the hyperparameter *K,* and the performance metrics Silhouette index, WSS and within-cluster R^2 is also calculated for each run. The silhouette index indicates that the optimal value of *K* is 2. For *K=2*, the calculated silhouette score is 0.12 as is presented in Figure 8, where the blue line depicts the calculated performance metric and the red lines mark the identified optimum for that performance metric. The blue curve depicts the performance metric and the red lines identifies the optimum. A silhouette score of 0.12 indicates a lack of compact or well separated clusters in the dataset. Therefore, the silhouette index is considered to be insufficient and therefore it is decided that the optimal value of *K* should instead be derived from the elbow point method.

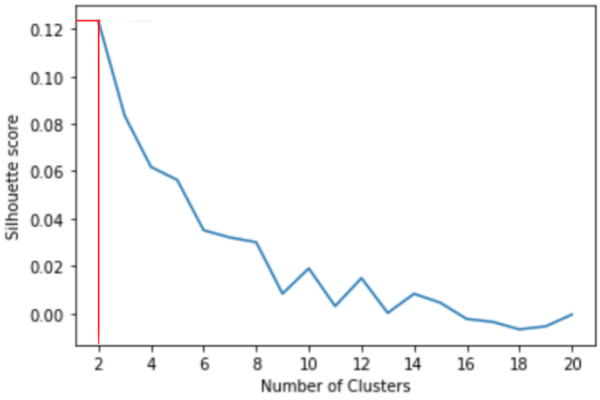
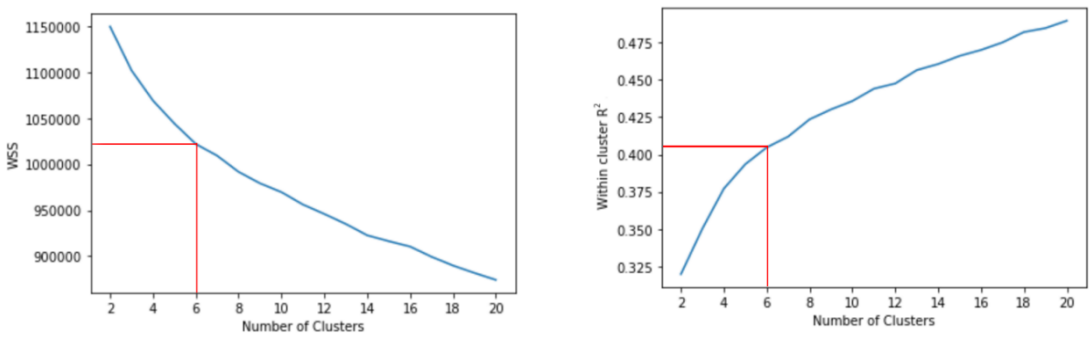


Figure 8, silhouette score for the clusters

The elbow diagrams for the two metrics of WSS and R^2 are displayed in Figure 9. Similarly to Figure 8 they depict the calculated performance metrics for all values between 2 and 20. Figure 9(a) illustrates how the WSS value changes as the number of clusters changes. Figure 9(b) depicts the same relationship for the within cluster R^2 value. These diagrams are analyzed optically and the identified elbow points have been marked in the diagrams.



(a) (b)

Figure 9, elbow method based on the WSS (a) and within cluster R² (b) with respect to the number of clusters

Neither the WSS nor R² value shows a sharp elbow where the error metrics completely stop declining as more clusters are added. When the elbow diagrams of the different performance metrics are evaluated it is hard to establish an elbow point, however when they are analyzed in conjunction, an elbow point is determined to exist at K=6 for both WSS and R².

Table 2, the clusters and their respective number of members

Cluster	Number of members
Cluster 1	12
Cluster 2	46
Cluster 3	28
Cluster 4	59
Cluster 5	8
Cluster 6	40

Table 2 depicts an example result of the clustering algorithm. The number of members in each cluster can be seen in the table as well. As the clustering algorithm is stochastic the clusters and numbers of data points vary with each run, however, consistent cluster profiles are observed between multiple runs of the clustering algorithm. After the initial clustering an evaluation is performed to determine if the cluster centroid is deemed to be a sufficient representation of the cluster's individuals. For the specific run no cluster had less than 5 members, and thus no cluster is scattered due to its size being too small.

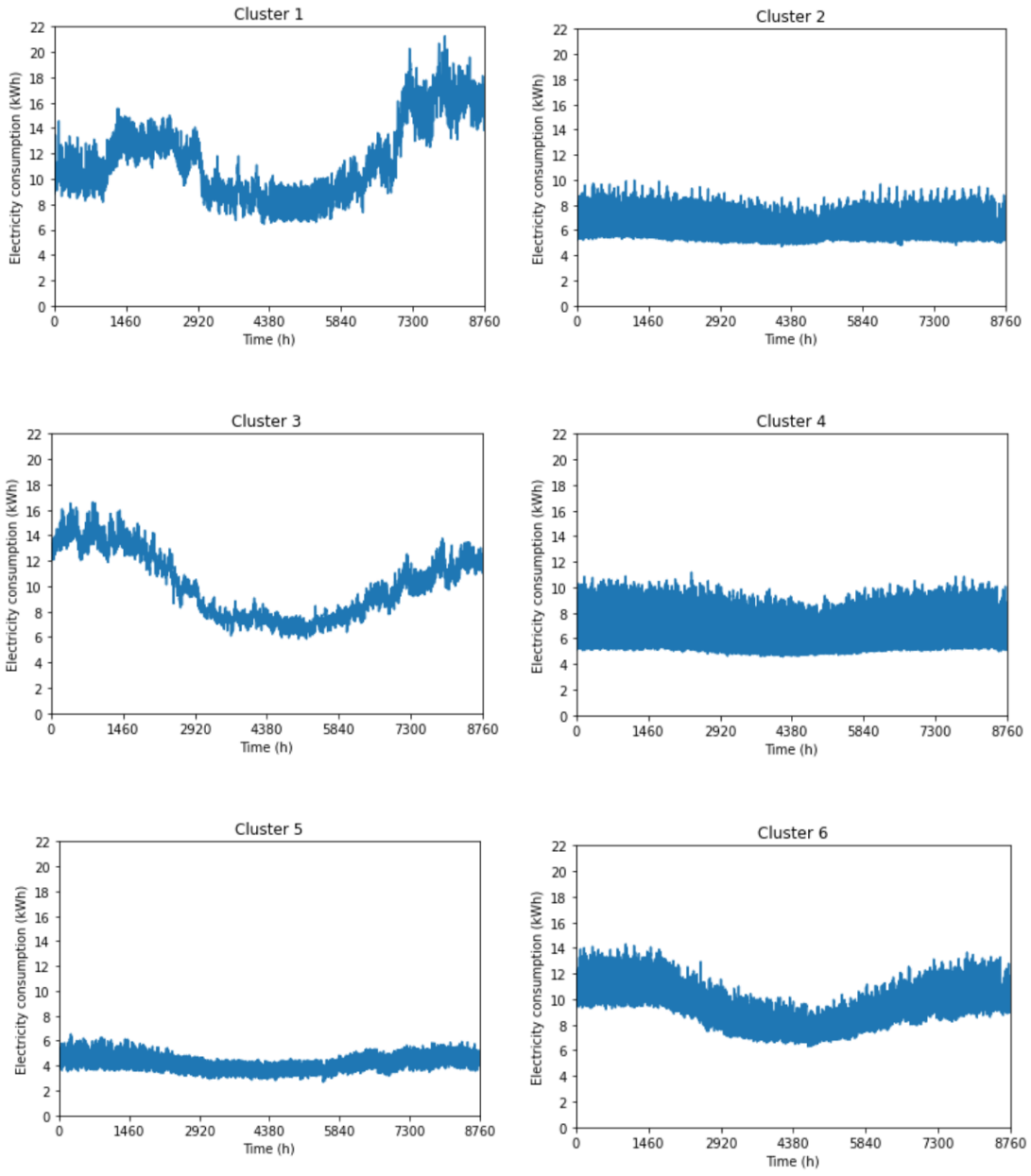


Figure 10, the centroids of the different clusters regarding the substations.

Figure 10 displays the centroids of the clusters given in Table 2. The cluster centroids all represent different patterns of electricity consumption in the dataset. It is important to remember that the cluster centroids are generated from normalized data. The clustering procedure thus ignores the scale of the data and only considers the patterns (seasonal and daily) when generating the clusters. This implies that individuals that have a mean consumption of e.g. 50 kWh might be clustered to cluster 5 if its consumption pattern is similar. When denormalizing the cluster centroids they are transformed to have the mean and standard deviation of the data which has been classified to the respective cluster. Some of the more irregular or seasonal consumption profiles display higher than average consumption during the late autumn, winter and early spring months, suggesting that they may contain some electric heat pumps. (It was

discovered during the study that many of the substations with high electricity consumption contained heat pumps.) Centroids for clusters 1, 3, 5 and 6 have more of a seasonal pattern, compared to the relatively numerous clusters 2 and 4 which lack this seasonal profile. Furthermore, it can be seen that the centroid for cluster 1 exhibits an atypical consumption pattern. It has a clear nonseasonal pattern in the early stages of the year. The behavior of initially low consumption for the first 1000 hours compared to hours 1000-2500 is quite hard to explain, yet, it could not be linked to a specific item disturbing the clustering and this cluster profile would also appear consistently between runs. One speculation is that Uppsalahem continuously works towards a more sustainable consumption by, for instance, installing heat pumps. These heat pumps can be a reason for the more uneven consumption pattern at cluster centroid 1 but also centroid 3.

4.2 Gaussian process regression model

It is established through early modelling attempts that the possibility of creating accurate models on a 24-hour basis varies a lot between the different substations. Due to the high variability in the dataset the optimal choice of kernel function and dependent variables also varies between the individual substations. For this reason, it is decided that the configuration of dependent variables and kernel function should be chosen individually for each of the substations and predictions given on an individual basis. However, to begin with, a general choice of kernel function is established and its results for the dataset in general are presented in Table 3. This is done to communicate a general idea of the performance of the Gaussian process regression model for the entire dataset while also showing the distribution of well-predicted and less well-predicted substations. Due to the large number of substations, these results are presented by grouping the prediction results into one of four buckets, for each of which the performance metrics of R^2 and MAPE are displayed as an interval.

A subset of the electricity data is chosen based on what has been identified as a representative distribution of the different patterns present in the dataset. The cross-validation procedure described in Subsection 3.4.6 is conducted for this subset of data series to establish an optimal kernel for the dataset in general. The kernels that are investigated are Matern 3, Matern 5, RBF, ExpSineSquared and combinations thereof, as well as different choices of dependent variables. Refer to Table 3 for a complete list of the combinations evaluated.

Table 3, Exemplification of cross validation for a substation. The dependent variables are represented as T for time, D for day, H for hour and finally W for the outside temperature.

Kernel function	Dependent variables	MAE	R^2	RMSE
Persistence	None	2.558	0.518	3.778
ExpSineSquared + RBF	*T*	2.058	0.721	2.867
ExpSineSquared × RBF	*T*	2.120	0.712	2.908
Matern 3	*D, H*	2.164	0.711	2.919
Matern 5	*D, H*	2.159	0.711	2.918
RBF	*D, H*	2.157	0.709	2.923
Matern 3 + RBF	*D, H*	2.098	0.717	2.886
Matern 3 × RBF	*D, H*	2.145	0.712	2.914
Matern 3	*D, H, W*	2.210	0.701	2.962
Matern 5	*D, H, W*	2.191	0.704	2.952
RBF	*D, H, W*	2.199	0.702	2.965
Matern 3 + RBF	*D, H, W*	2.148	0.706	2.939
Matern 3 × RBF	*D, H, W*	2.475	0.648	3.223

Table 3 displays an example of the cross-validation procedure for one of the selected substations. Additionally, the cross-validation results for the different metrics are compared to those of the persistence model. The persistence model is utilized to see if there are diurnal patterns in the electricity profiles that the regression model should pick up on, and as a benchmark of performance for the profiles where such patterns are present. The persistence model is a model that treats the observed values for the previous time horizon as a prediction for the upcoming time horizon, for a time horizon of 24 hours, this means that the observation of today's consumption is the prediction for tomorrow. It is important to stress the fact that the table only describes one specific instance of the cross-validation procedure, and that the numbers displayed are not representative for the dataset as a whole.

During this cross-validation process varying results are observed for the optimal choice of kernel function and dependent variables as well as performance gain compared to the persistence model. Some general observations are that the sum-kernels of Matern3 + RBF and ExpSineSquared + RBF sometimes outperform the remaining evaluated kernels by a wide margin, while rarely underperforming significantly, indicating that there might be benefits to the flexibility provided by the sum-kernels. Additionally, the periodic ExpSineSquared + RBF kernel would as might be expected perform significantly worse for the data where the results from the persistence model were poor, suggesting an absence of periodic patterns. Finally, the inclusion of outside temperature would often result in slightly lower accuracy of predictions. Examination of the hyper-parameter values after optimization showed that the optimizer would often fit a short length-scale to either weather or hour of the day, suggesting that the correlation between outside temperature and time of day may make it difficult to infer optima for the hyper-parameters. While this does not suggest that electricity consumption is unrelated to the outside temperature, it might not aid in making day-to-day predictions for most of the substations. This observation does however not hold true for the whole dataset, and some of the high consumption substations would benefit from using weather as a dependent variable, likely due to the presence of electric heating. However, no

comprehensive list of electric heating equipment connected to the different substations has been acquired, making it difficult to confirm this hypothesis. After the selected set is evaluated, the sum-kernel of Matern3+RBF without the inclusion of outside temperature as a dependent variable is chosen as a "best general model" for the dataset. Van der Meer et al. (2018b) similarly conclude that the sum-kernel Matern 3+RBF yields the best results for hour-ahead predictions of an electricity profile, further supporting this result.

However, as might be expected from the wide variation of shapes among the profiles, establishing an optimal choice of kernel function and dependent variables for the entirety of the data is a hard task and the optimal choice of kernel function and dependent variables varies between individual substations. Based on these differing results it is recommended that kernel functions and dependent variables should be cross-validated and chosen on an individual basis. However, an exhaustive search for the optimal combination of kernel functions and dependent variables on an individual basis is not conducted within this study due to the restraint of computational demand.

After a choice of a generally well-performing combination of kernel function and dependent variables has been established the cross-validation procedure is repeated for this specific kernel, and R^2 and MAPE values are calculated and the results from this process are presented in Table 4.

Table 4, results presenting the R^2-values and MAPE for the regression model for all the individuals in the data set.

Number of substations	R^2	MAPE
5	0.75-1.00	5.9%-7.9%
25	0.5-0.75	7.7%-12.5%
40	0.25-0.5	8.9%-15.3%
123	<0.25	9.5%-21.1%

The substations are grouped into the different buckets based on the metric of R^2, while the displayed interval for MAPE is simply the range of MAPE values of the substations grouped to that bucket. The relationship between the metrics of R^2 and MAPE is not clear cut, and substations with a high R^2 value will sometimes report a higher MAPE than those with a lower R^2. This is why the intervals for MAPE of the different buckets sometimes overlap. For the chosen general model, the MAPE of the predictions is in the 10 - 15 % range for the majority of the substations. It should be noted that some of the substations are expected to benefit more from a different combination of kernel function and dependent variables, but the results displayed in Table 4 are deemed to offer a relatively accurate representation of what may be achieved with the different models tried within this study.

Since the dataset is not cleared of what is considered to be abnormal behaviors, these often disturb the performance of the Gaussian process regression model. Whenever discrete behavior changes occur in a data series, the Gaussian process model requires multiple time horizons to learn the new behavior. While the behavior of the Gaussian process in these cases is desirable in the design of an anomaly detection model, it also hinders the performance whenever the data contains these abnormalities.

The wide variety in the degree of predictability in the day-to-day behavior of the substations is likely to present challenges for the application of anomaly detection. One of the initially perceived benefits of the Gaussian process regression model was the establishment of a standard deviation for the resulting models, which may allow for taking a margin of error for the predictions into account. However, when the predictability of the data is low this presents some problems, as showcased in Figure 12, which displays predictions and the 95 % confidence interval for what is deemed to be one example of relatively accurate predictions, and one example where predictions are less accurate.

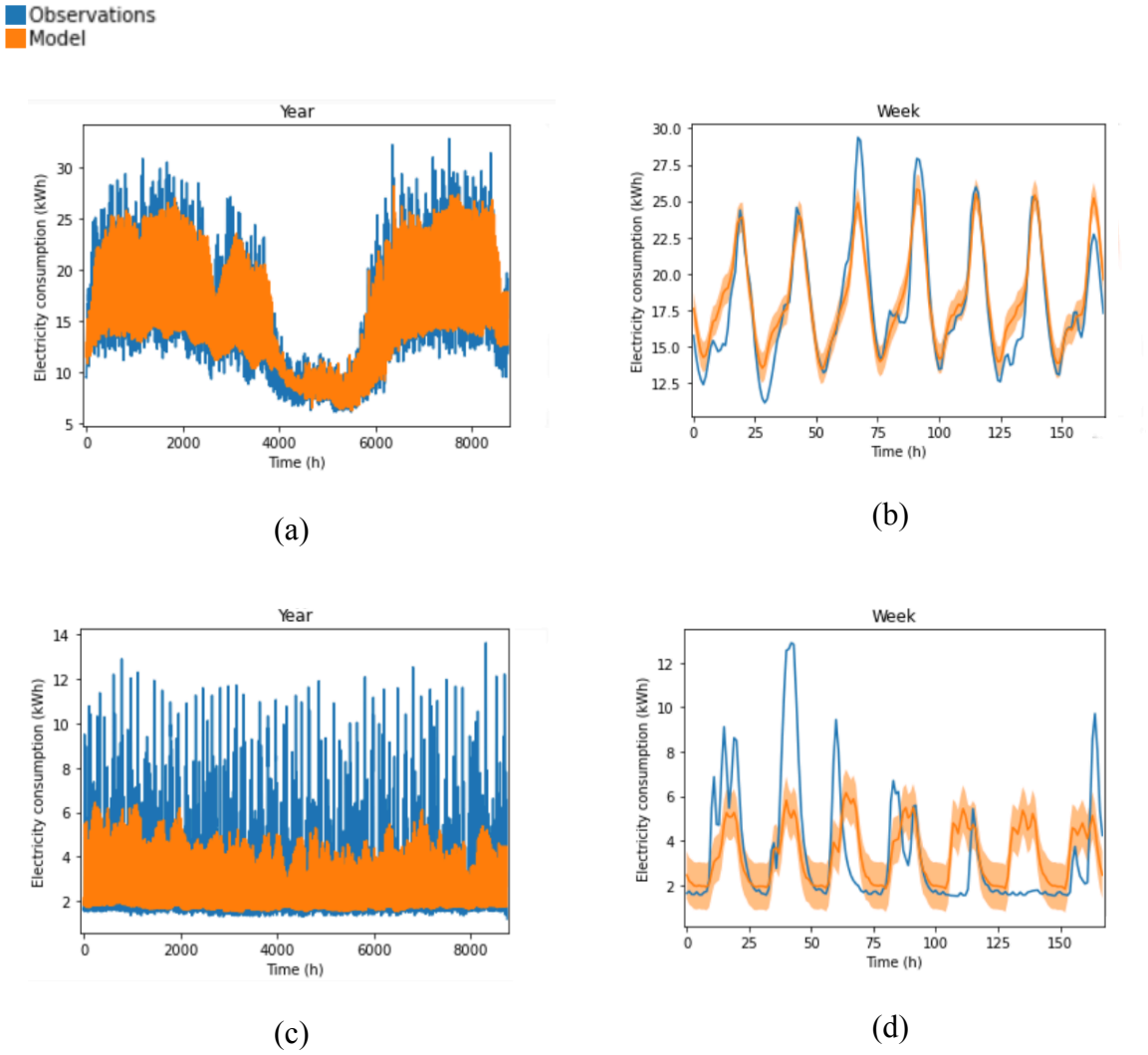


Figure 11, illustrating the differences of the forecasting between two different substations on two different time periods, yearly and weekly. The blue curve is the observed data and the orange curve is the prediction.

Figure 11(a) and Figure 11(b) illustrates a substation which consists of a clear diurnal pattern on a yearly and weekly basis respectively. Figure 11(c) and Figure 11(d) on the contrary shows a substation which the regression model struggles to fit a function to the observed data. The model's inability to predict the peaks is highlighted in Figure 11(d). When comparing Figure 11(b) and 11(d) it is apparent that the predictability of the diurnal patterns is highly dependent on the individual substation. The shaded area in the

graphs 11(b) and 11(d) displays the 95 % confidence interval of the predictions. As can be seen in Figure 11(d) the substation has a larger shaded bound due to the fact that is more difficult to predict.

Figure 11(d) also displays some of the main weaknesses of the Gaussian process when applied to electricity data. Gaussian processes often struggle to accurately model noisy data (Bijl et al., 2016), and the assumption of the White Kernel, that noise in the observations may be modeled independently as $N\sim(0, \sigma^2)$ with the same variance for all observations, does not hold very well for many of the examined electricity profiles. Contrarily, the unpredictable variance in the observations seems strongly linked to the relative values of x and y, i.e. the need to add noise to model observations is greater during certain hours of the day, when the peaks in electricity consumption occur. For future reference, some type of weighted noise function might improve model performance in these cases. Due to these issues with the forecasting model, in the implementation of the Gaussian processes based anomaly detection model a focus is placed on the cases where more accurate predictions could be attained.

4.3 Anomaly detection implementation

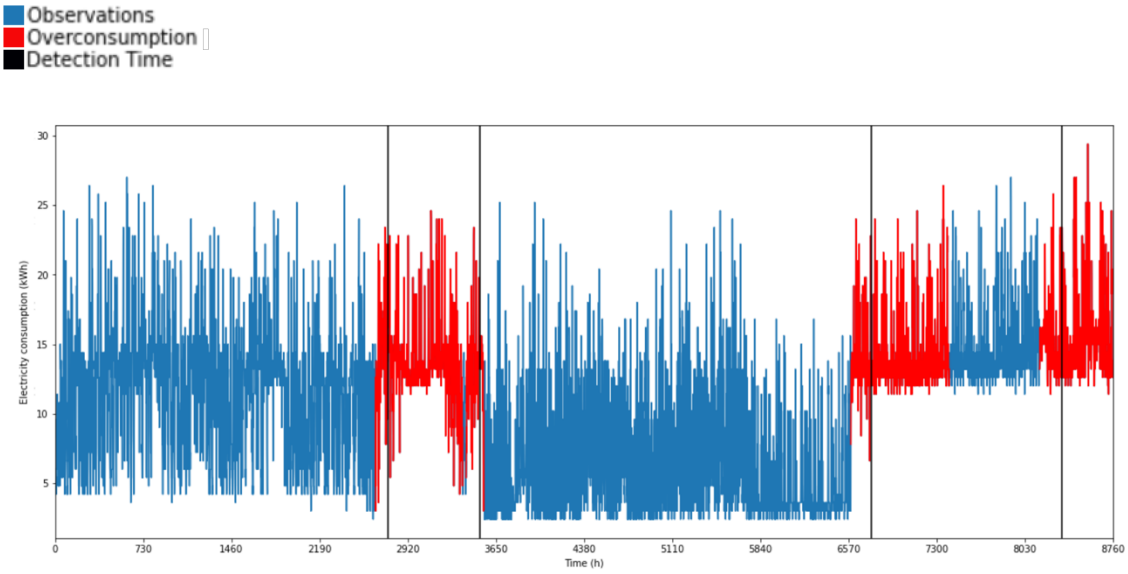
This section presents the examples of results of the two anomaly detection models developed. These examples are divided into two different subsections for each of the developed models. For each example some commentary on what the model detects as anomalies and analysis of the model's strengths and weaknesses is provided. The existing anomaly reports from Uppsalahem are also compared directly to the developed model in order to give an idea of the situations in which the developed model differs from the anomaly reports. Uppsalahem's current anomaly detection method on substation level described in Subsection 2.1 is the foundation of these reports.

Full-scale comparisons of the developed models and the existing anomaly reports at Uppsalahem have been attempted, but due to the significant differences in the method of the different models the results of the different models were too dissimilar for such a comparison to be deemed meaningful. The overlap in detected anomalies is observed to be limited to somewhere around 50-60 %, and most of these observed differences can be explained by the differences in how the different models define anomalies. Therefore, comparisons are limited to the individual examples presented in text, so that explanations of the observed differences may be provided. However, since said reports only contain information about abnormal energy consumption on a monthly basis, it is sometimes hard to pinpoint if the models are indeed reacting to the same data.

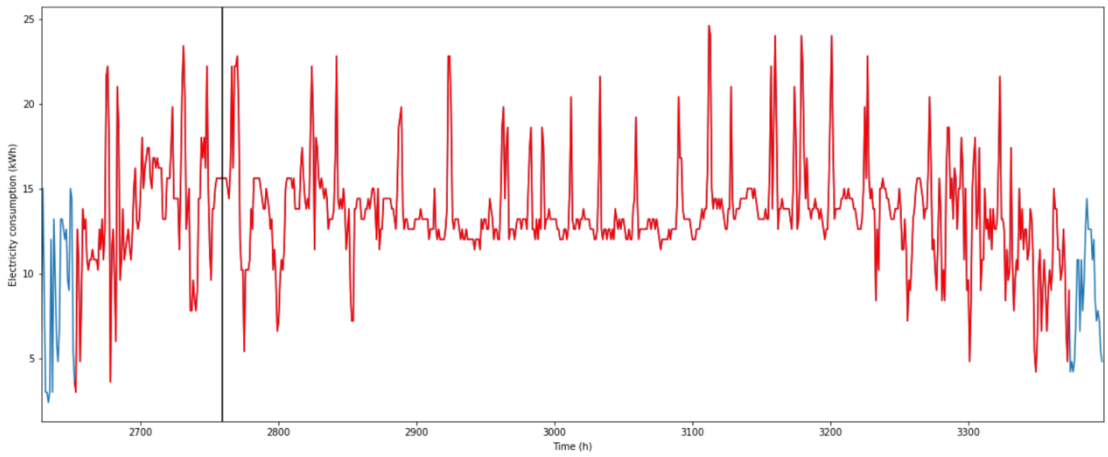
4.3.1 K-means anomaly detection

An analysis of a subgroup of substations is made to draw some conclusions about the results generated from the K-means model. Six cases are illustrated in this subsection to present the differences between the K-means model and Uppsalahem's model. In the graphs that follow, the anomalies that the model finds are colored differently. The blue color represents "normal" consumption, the red color portrays abnormal consumption as it is too high, the green color illustrates abnormal consumption that is too low. It should be noted that red and blue intervals depict what the respective models define as an anomaly, and that they therefore, do not correspond to any specific external definition of what an anomaly is. Finally, the black vertical line represents the time the anomaly is

detected by the model. The first three graphs display executions of the model where parameters have been configured to optimally detect what has been labeled as anomalies in the data series. The values on the x-axis are selected to mark intervals of two months.



(a)

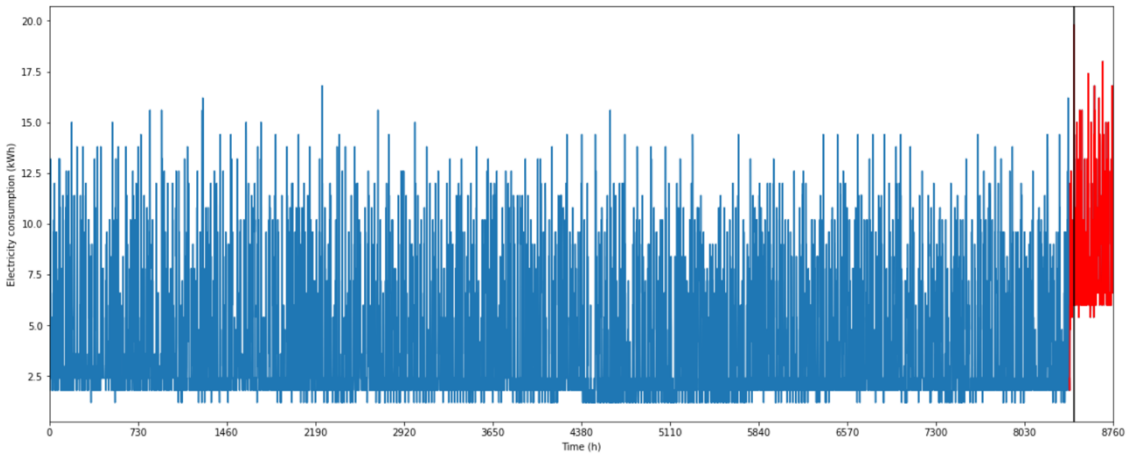


(b)

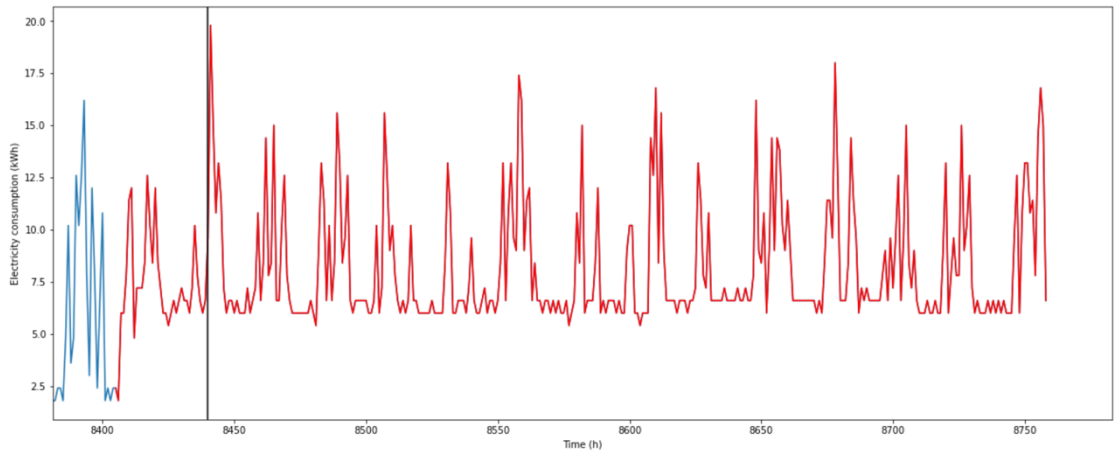
Figure 12, substation with the detected anomalies marked. Figure 12(b) is the interval surrounding the first anomaly detected.

Figure 12(a) displays the electricity load profile for the whole year of 2019 for which four anomalies have been detected. Figure 12(b) displays the first of the four anomalies on a shorter time scale. This arrangement of information is also used for the subsequent figures Figure 14 and Figure 15. Two anomalies have been identified during the analysis, the first taking place roughly between hours 2700 and 3400 and the second from 6500 lasting until the end of the interval. The model also detects anomalies in both of these regions. The model does however not detect an anomaly for a section of the latter of these intervals, between hour 7500 and 8100, while the sections before and after that interval are considered to be different from normal consumption.

Uppsalahem's existing anomaly detection model notes anomalies during all months except for June and September.



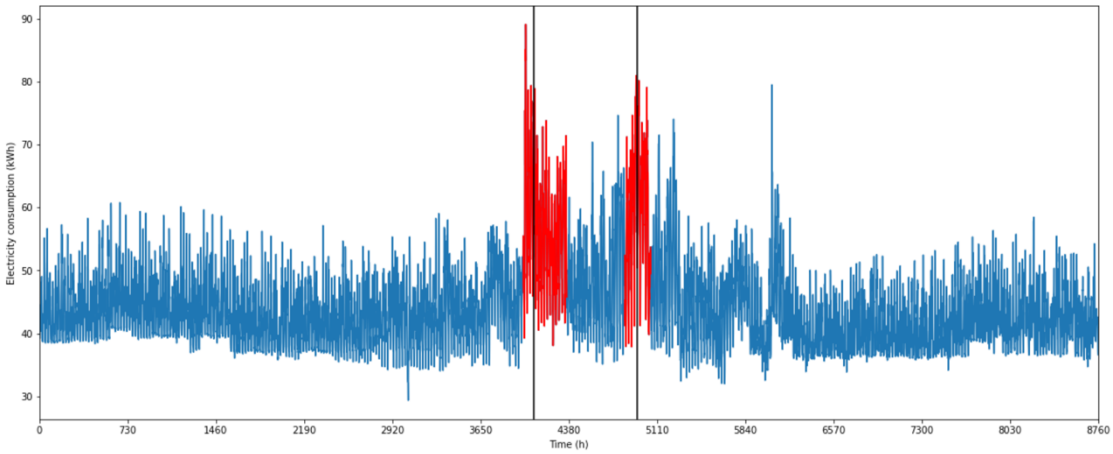
(a)



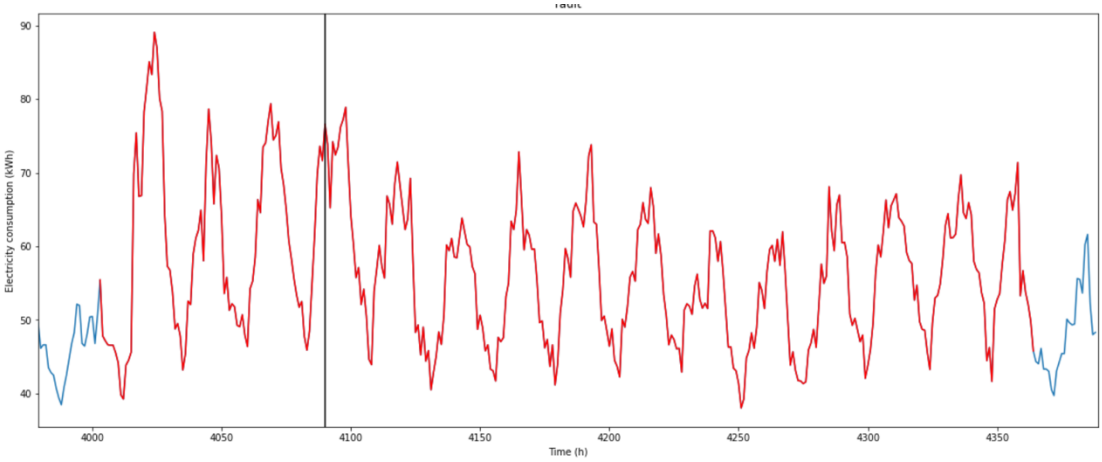
(b)

Figure 13(a), showing a substation with a single detected anomaly, and 13(b) displaying the area around the detected anomaly and the time of detection.

In figure 13(a) an anomaly has been identified beginning after hour 8000 and lasting until the end of the year. This anomaly is detected by the K-means fault detection model and also appears as an anomaly during December in Uppsalahem's reports. None of the models detect any other anomalies for the year evaluated. The electricity consumption for the rest of the period is fairly consistent which further confirms the detected interval actually being an anomaly, but it could represent an installation of new electricity consuming appliances. Figure 13(b) illustrates when the anomaly is found and the surrounding time interval.



(a)



(b)

Figure 14, a substation with two detected anomalies. Figure 14(a) displays the detected anomalies in the year profile and Figure 14(b) displays the consumption around the first of the detected anomalies.

Figure 14 displays the electricity load profile for a substation where the model has detected two anomalies. These two anomalies are detected around hour 4000 and 5000 respectively. The first anomaly is a sharp increase of the consumption in the middle of the summer. Tomas Nordqvist at Uppsalahem mentions that some of Uppsalahem's properties have air conditioning installed which might be the reason for the increase of electricity consumption (Nordqvist, 2020). However, there is a third peak after hour 6000 which has also been identified through the optical screening, but which is not detected as an anomaly by the model. This is due to the settings of the parameters, or possibly an increase in the model's average consumption during the autumn. The time from the beginning of the identified start of the first anomaly until the model detects it is visible in Figure 14(b).

Uppsalahem's model does not find any anomalies for this specific substation. This is due to the fact that the increased consumption during summer months is a recurring

pattern for this substation, which the K-means model is not specific enough to account for and thus, this pattern is detected as an anomaly. This indicates that there are also benefits to the current anomaly detection method when more specific patterns are annually occurring. It is also made apparent that expert analysis of the detected anomalies is still needed.

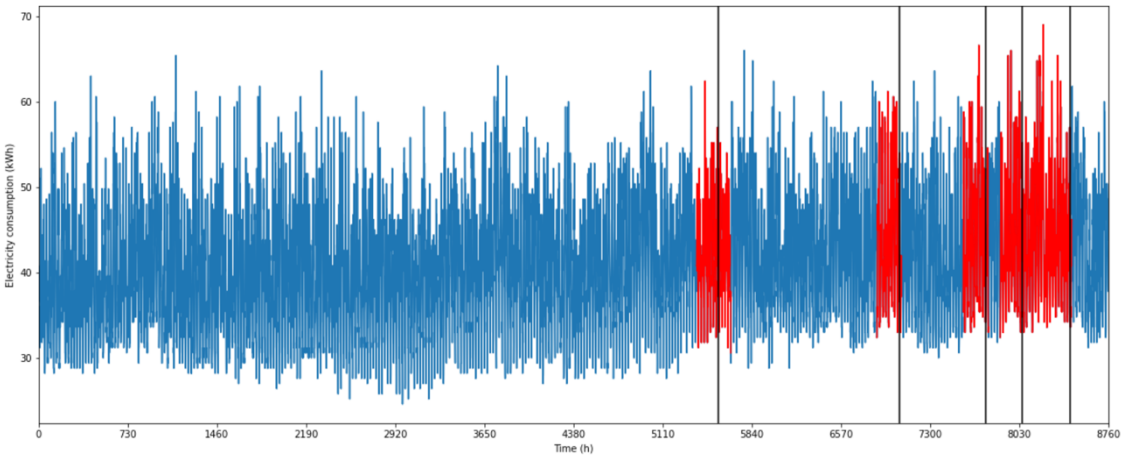


Figure 15, a substation with five detected anomalies in succession.

Figure 15 presents the consumption data for a substation where no distinct anomalies have been identified beforehand. Uppsalahem's existing anomaly detection model has not detected any anomalies during the year. The K-means anomaly detection model however finds an anomaly around hour 5500 and several additional anomalies are detected with an increasing frequency towards the end of the year. Uppsalahem's model does not detect any anomalies for this given substation. However, there seems to be a continuous slight increase of the base consumption from approximately hour 3000.

While this can be considered to be an example of the model capturing a *drift* type anomaly, the way it reports this as a series of discrete anomalies might be confusing to an eventual operator and thus less than ideal. The clustering for this substation is deemed to be correctly executed. An explanation regarding the detection of the separate anomalous intervals detected by the model is that they all have a relatively high minimum consumption, placing the consumption continuously outside the model's bounds during these specific intervals. Whether the detected anomalies are indeed a drift type anomaly or if the consumption increase is in fact within the bounds of what should be considered normal is arguable.

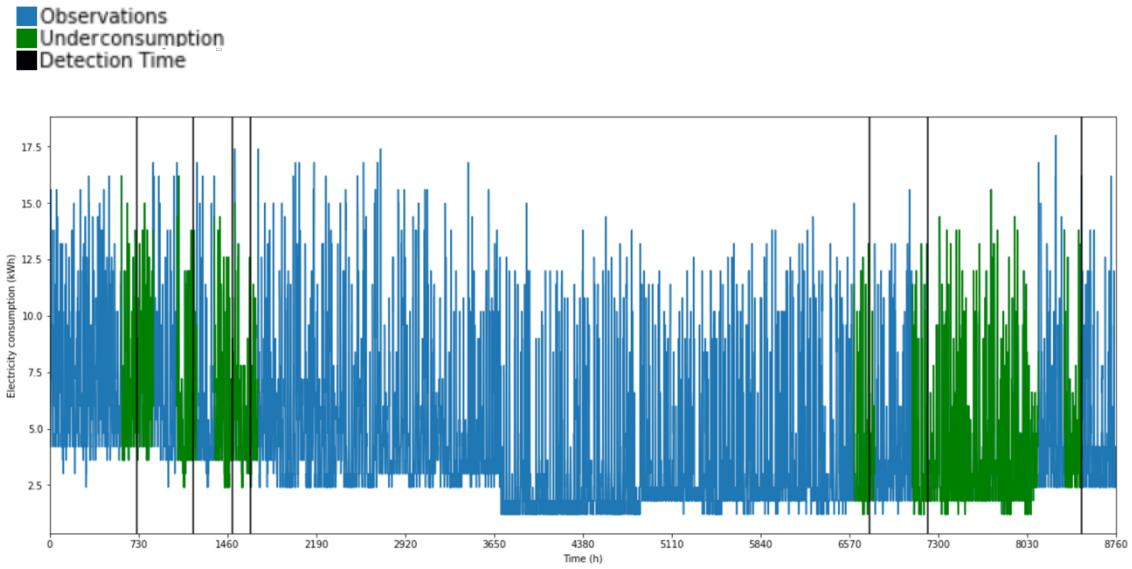


Figure 16, a substation where several intervals of abnormally low consumption have been detected.

Figure 16 displays a case where the detected anomalies are judged to be incorrectly determined. When analyzing the substation's consumption profile no anomalies are identified. The K-means model however detects two series of smaller anomalies, one at the beginning of the year and one near the end of the year. Uppsalahem's model also finds anomalies in the latter part of the year, for all months from August until December.

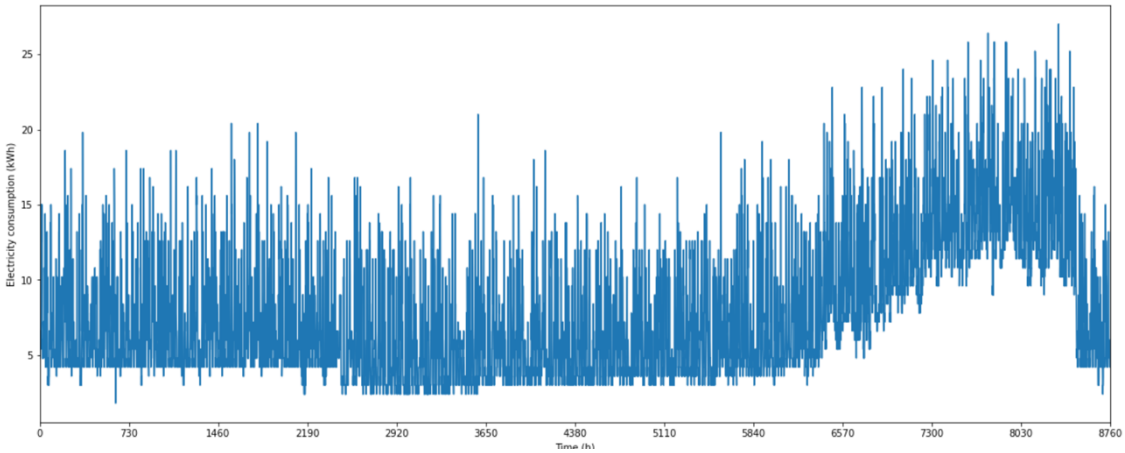


Figure 17, training data for the substation shown in Figure 16.

Figure 17 presents the training data for the substation from Figure 16. When analyzing the training data (2018 year's data) it presents an explanation to the detected anomalies. There is an increase of consumption during the last months of 2018. This might be the reason why it is assigned to cluster 1. When examining the electricity consumption for 2019 this substation no longer presents the type of pattern that cluster 1 represents. The substation might in fact rather belong to cluster 2 or cluster 4. This increase of

consumption during 2018 is likely connected to a renovation project where water pipes were changed during this time (Nordqvist, 2020).

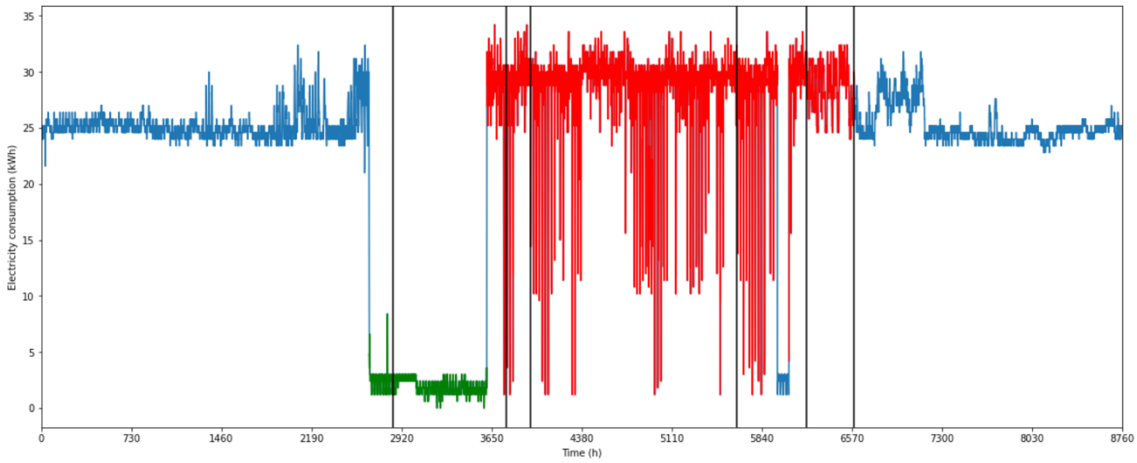


Figure 18, data and detected anomalies for a substation which is identified as irregular.

Figure 18 depicts one specific substation which shows an irregular behavior. The determination of anomalies for this case is difficult due to these irregularities. Uppsalahem's model detects anomalies in April and May due to the consumption being lower than expected. Uppsalahem's model also finds higher than expected consumption during all other months except for November and December. It is hard to draw any conclusions about whether the K-means model or Uppsalahem's model functions better for this given substation due to the irregularity. This is a substation that probably should have been discarded from the clustering and analysis as it is a clear outlier in the dataset, but no such cleaning procedure is executed for the dataset.

Overall, labeled data considering anomalies would have been immensely helpful in determining the two model parameters of the anomaly size limit L and bound B for the dataset in general.

As no labeled data of anomalies exists for the whole dataset, the optimization of the parameters of the anomaly size limit L and bound B has largely been done through an analysis of a small set of data, but correctly labeled data would enable a more specific determination of the optimal value for said parameters and clearer evaluation of the model performance. The fact that upgrades of the heating and facility management occur relatively frequently also poses problems. This raises the question if a reclustering is necessary and how often it is required. If a substation has a new installation of a heat pump or any other electricity demanding equipment it is very likely to have falsely detected anomalies. Furthermore, the number of substations that are well represented by the cluster centroid is highly dependent on the variety of patterns present within the dataset. The K-means method is first and foremost able to find consumption that varies from the seasonal pattern that is visible in the different clusters.

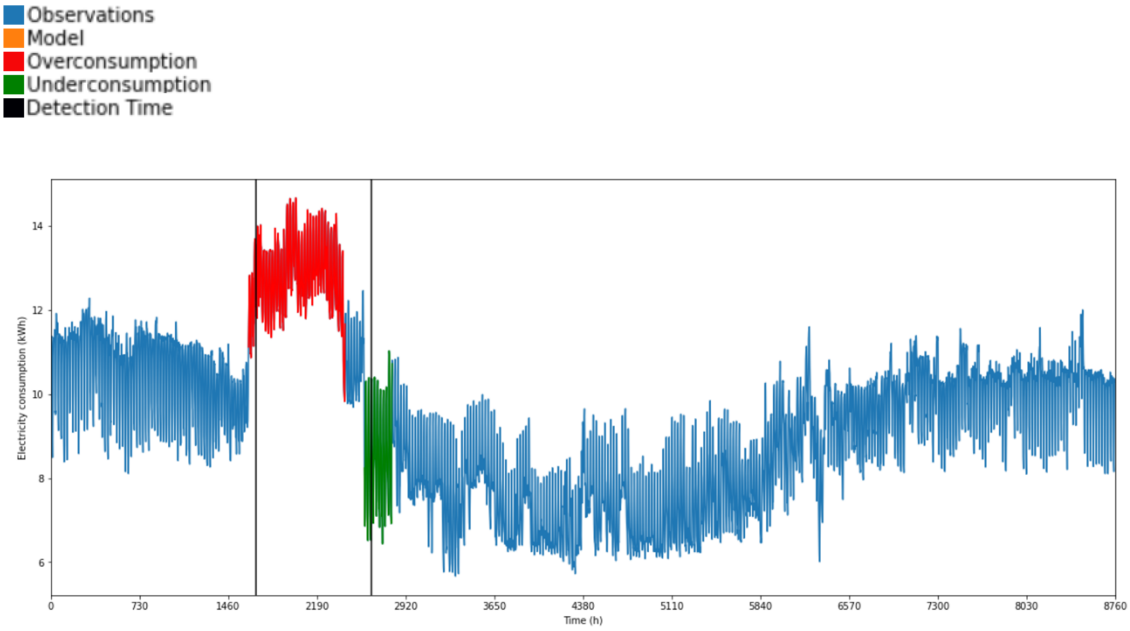
The cluster-then-compare approach utilized in this study comes with a mixed set of drawbacks and benefits when compared to directly learning and predicting the individual substation's measurement data. The main drawback which also limits the utility of the anomaly detection model is that the cluster centroid and the individual substation's behavior are relatively dissimilar. As illustrated in Figure 10 the mean R^2

between cluster centroids and individual substations for the interval of the hyper-parameter K determined to be of interest is around 0.4, meaning that the centroids only offer a somewhat accurate representation of the behavior of the clustered data.

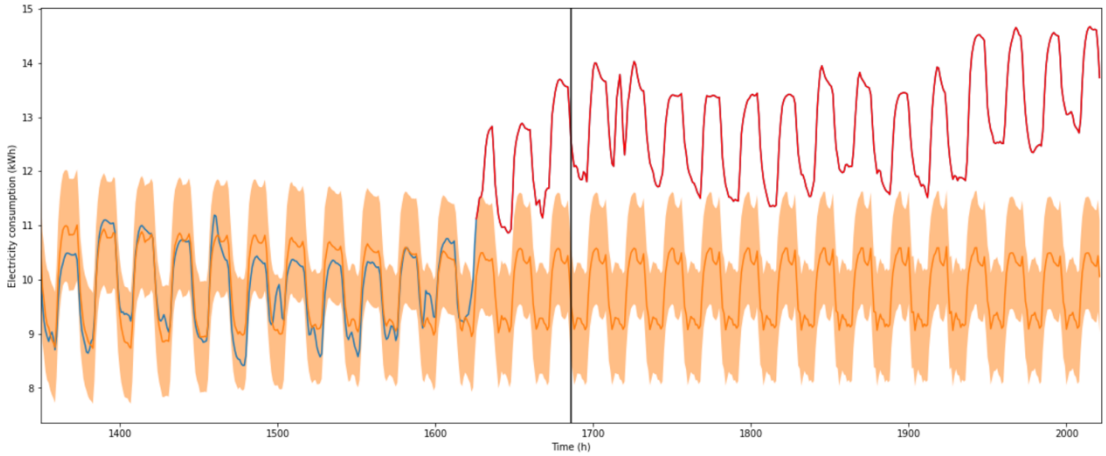
To summarize this analysis of the model, it is deemed to work well but its performance is dependent on the data. It is able to find most anomalies within a week, and unlike the existing model at Uppsalahem is able to do so independently of when during the month these anomalies appear. Furthermore, it would benefit from more extensive data of identified anomalies to enable an optimization of the anomaly detection parameters. Without such labeled data, extensive work would be required to tune these anomaly detection parameters individually to each substation.

4.3.2 Gaussian process regression anomaly detection

As for the K-means anomaly detection a subset of substations is examined to enable an analysis of the Gaussian process regression anomaly detection. As the possibilities of reaching accurate predictions differ between the data points in the dataset as established in Subsection 4.2, a focus is put on some of the data series for which the predictability is deemed to be sufficient to support such an analysis. The model parameters described in Subsection 3.5 are optimized in order to detect optically identified anomalies, similarly to the case of the K-mean anomaly detection model. A few examples of the model output are displayed and analyzed in sequence below.



(a)



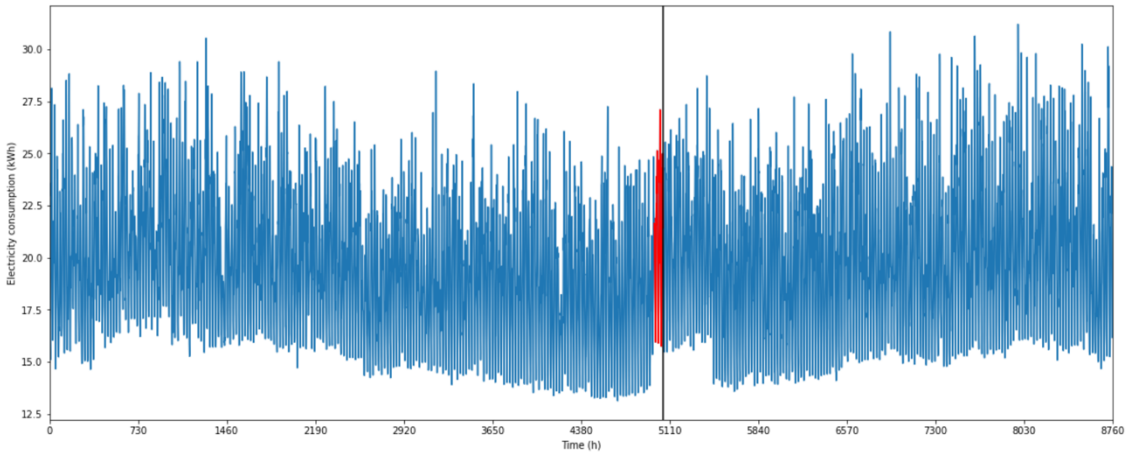
(b)

Figure 19, substation with the detected anomalies marked. Figure 19(b) is the interval surrounding the first anomaly detected.

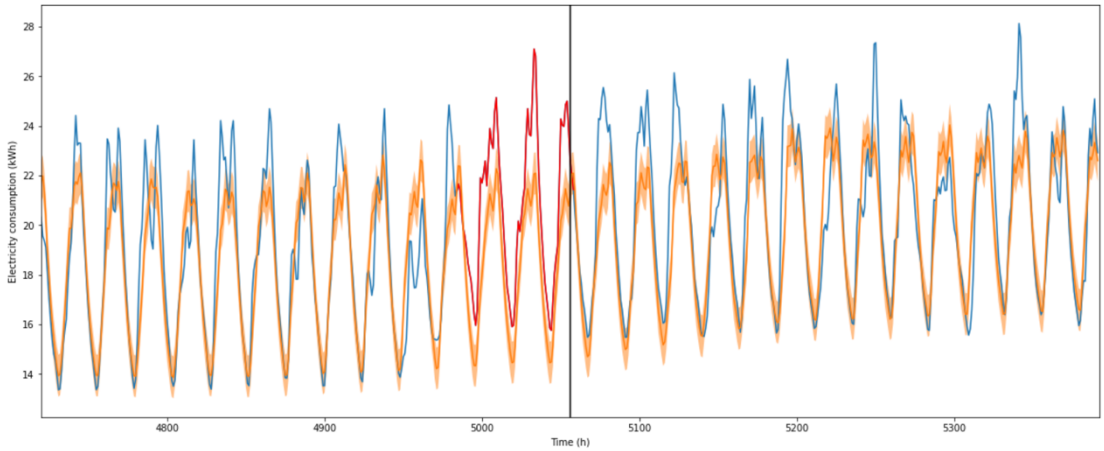
Figure 19 illustrates one of the cases where the Gaussian process regression model is able to predict the future consumption well ($R^2 = 0.83$ and MAPE 7.6%). Therefore the implementation of the anomaly detection algorithm operates well and locates two anomalies. The first anomaly detected as well as the model's prediction for the same time interval is illustrated in Figure 19(b).

The second anomaly that is found is a so-called secondary anomaly. It is only detected due to the deviating consumption returning to "normal" consumption. While testing and selecting model parameters, a neutral position is taken towards such output. The idea is that, in a real setting, operators would know whether an anomaly has occurred in the earlier data and thus may react accordingly. Additionally, the model theoretically

supports the option of manually selecting if and when training should begin again while retaining the previously learned hyperparameters. In this context, where anomaly detections are investigated and conclusions drawn about whether they resulted in actual anomalies which are subsequently amended or expected changes in behavior, choices about whether to retrain the model or not could be made based on this operator knowledge. However, as no clear information about the detected anomalies has been attained, this possibility is left untouched upon in these evaluations, and models are instead retrained whenever the observed consumption returns inside the chosen probability margin of the model expectations.



(a)

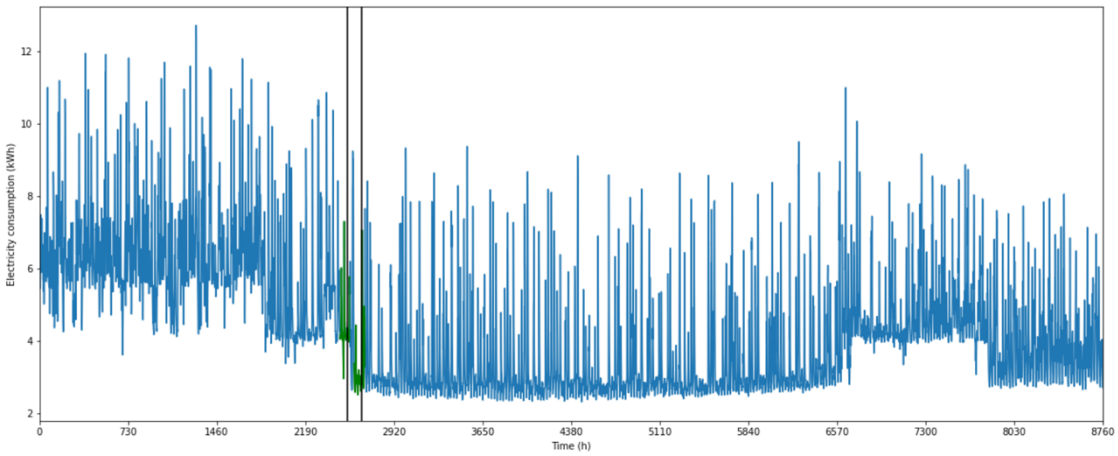


(b)

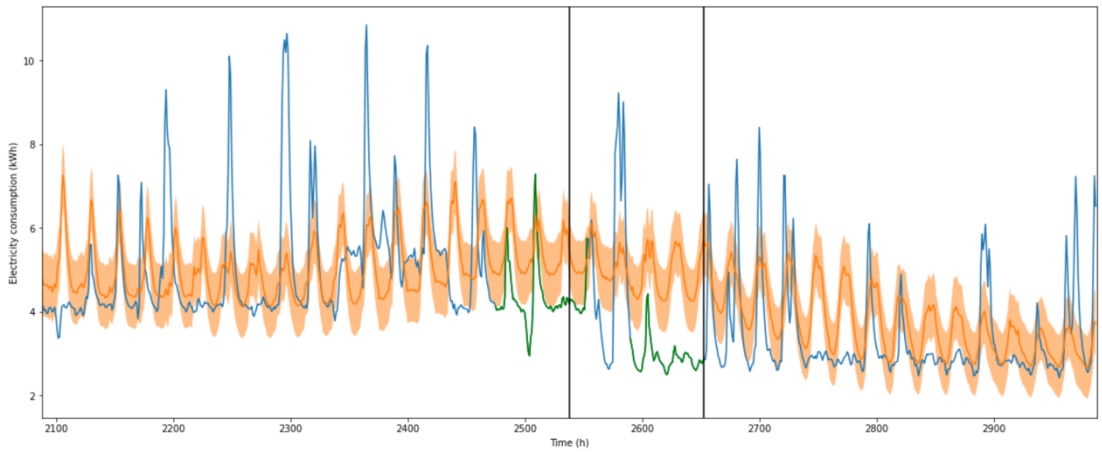
Figure 20, (a) consists of the consumption for the whole year whereas Figure 20(b) displays the surrounding interval of the detected anomaly.

Figure 20 displays the consumption for a substation where the anomaly is a lesser relative increase to the consumption than what is seen in figure 19. In this case the model interestingly enough does not cease to adapt the function to the new incoming data, but nevertheless detects the start of the abnormal interval as an anomaly. As the model is continuously updated, it ceases to give the anomaly signal as it adapts its

predictions to the new pattern. The chosen prediction model here achieved an R^2 of 0.74 and MAPE of 8.4 %, placing it slightly below the earlier example but retaining a high enough accuracy to keep similar parameter values for the anomaly detection. The detected anomaly is an increase of consumption by approximately 2 kWh per hour for every hour until the anomaly appears to be corrected. The parameter configuration is arguably very tight, as the model detects what is roughly an increase of 2 kW during only a few days as an anomaly, however it enables the model to detect only the interval which has been identified as abnormal. This example shows an issue which arises for the Gaussian process based model when anomalies are not sizeable enough for the model stop updating itself within the 24-hour prediction span. Similarly, the model is capable of detecting the anomaly within roughly 50 hours. As can be seen from Figure 20(b) the deviation is not as significant as in the previous case (Figure 19). This could however be detrimental to the model if the required length to detect an anomaly is defined to be longer. Furthermore, it does not detect any anomaly as the profile returns to the perceived normal. While this effect is achievable with a different set of parameters, only the initial detection of the anomaly is deemed to be of interest.



(a)



(b)

Figure 21, (a) consists of the consumption for the whole year whereas (b) is zoomed in on the detected anomalies.

Figure 21 displays a case where the attained predictions were far less accurate due to the wide variety in consumption peaks. Figure 21(b) displays the two anomalies detected, highlighting an issue which may arise in such a context. The model encounters an interval of high base consumption, and future predictions are influenced by this, causing the model to detect a perceived under-consumption from approximately 2450 to 2500. Arguably, the model should in fact detect the preceding interval of heightened base consumption as an anomaly, however, since it presents a large share of values above and beyond the predicted consumption, this does not happen, and the model instead predicts increased consumption in the subsequent interval based on this data. There is also a second anomaly detected as there is a second drop in base consumption, arguably presenting a correctly detected anomaly. Since the model detects the decreases in consumption and not the preceding increase of consumption it thus seems probable that the model is more likely to detect decreases rather than increases. This might be due to

the heightened prediction resulting from its inability to predict effect peaks. Additionally, there are multiple discontinuous changes in the base power consumption later in the year that are not detected. While it was possible to set a less strict bound for the anomaly detection model, that would also result in some inaccurate detections. Thus, this example speaks of increased difficulty of anomaly detection and uneven detections of heightened or lowered consumption when the prediction accuracy is low.

It is possible to implement the anomaly detection even in the scenarios where the predictions were even less accurate by setting much wider margins and longer thresholds for the anomaly detections. However, in these cases the model was not shown to learn much of daily patterns and thus they were not deemed to be of clear interest.

The method as such seems promising which is exemplified through the substations with regular behavior as they are able to detect anomalies on a weekly basis for the regression model. For irregular substations the prediction ability impairs the anomaly detection capability and the length of the interval the model must consider in order to detect an anomaly has to be weekly or longer rather than a few days, with anomaly detections that are deemed invalid still occurring under such parameter settings.

4.4 Comparison of the developed models

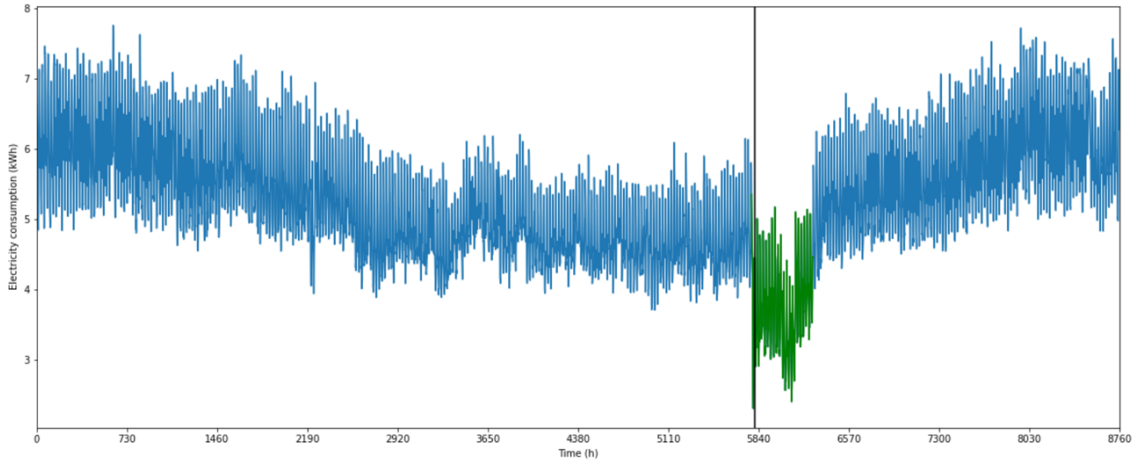
This project presents the implementation of two different models for detecting anomalies in electricity load profiles. The first of these models is based on unsupervised learning and finds anomalies through the grouping and comparison of similar profiles. The other is based on supervised learning and detects anomalies by comparing the current load to a forecast based on the history of the individual profile. A comparison of these models, their performance on the given dataset, and the anomalies found by the respective model is given in this subsection.

First, the K-means model, largely seems to cluster data based on seasonal trends. Additionally, it appears relatively insensitive to the presence of noise and shorter anomalies in the training data, however, it is sensitive to longer deviations which may cause the profile to be misclustered as illustrated in Figure 16. This also largely restricts the model's anomaly-detecting capabilities to finding deviations from the seasonally expected behavior. For the model to yield reasonable detections of anomalies, a lower limit of the detection time may exist at the daily time scale, while most anomalies deemed to be of interest may be detected within the week. This suggests anomalies may be detectable at a weekly or slightly shorter basis, which is a considerable improvement from the current detection time.

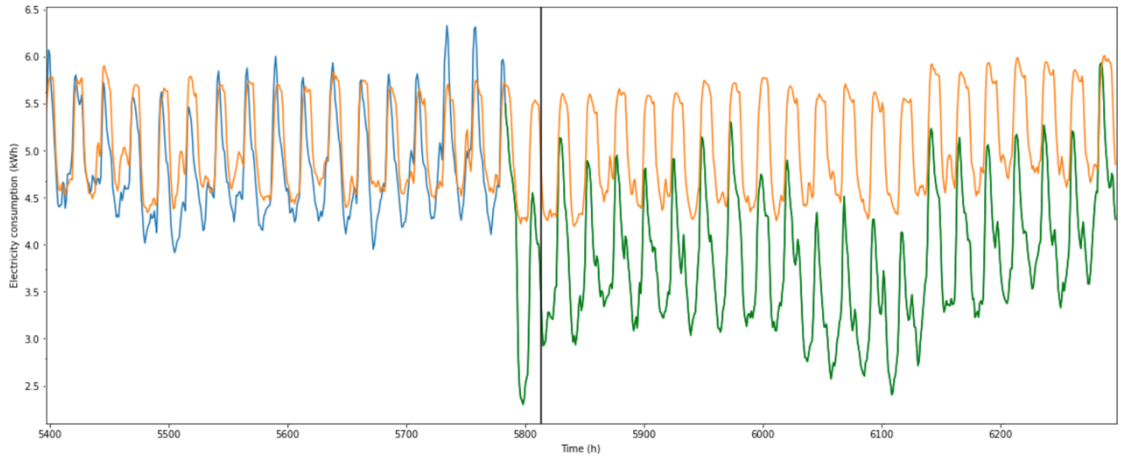
In contrast, the Gaussian process model is highly dependent on the most recent electricity consumption of the individual station and thus for the most part is less dependent on the overall seasonal patterns in the data. Rather, it treats the latest observed pattern as normal and detects any fast changes from that as anomalies. This, however, renders the model unable to detect slow drifts in consumption as it slowly learns the new pattern. For this reason, the model's anomaly-finding capabilities may actually be negatively impacted by demanding a longer time of deviation before detecting an anomaly, as it might learn abnormal patterns during the anomalous interval if they are close enough to the original. Therefore, the model sometimes balances on

finding criteria which are not too strict, to allow the detection of anomalies within a few days, but neither so loose as to trigger constant detections. The limitations of the Gaussian process based anomaly-finder are, thus, similarly placed around a few days to a week.

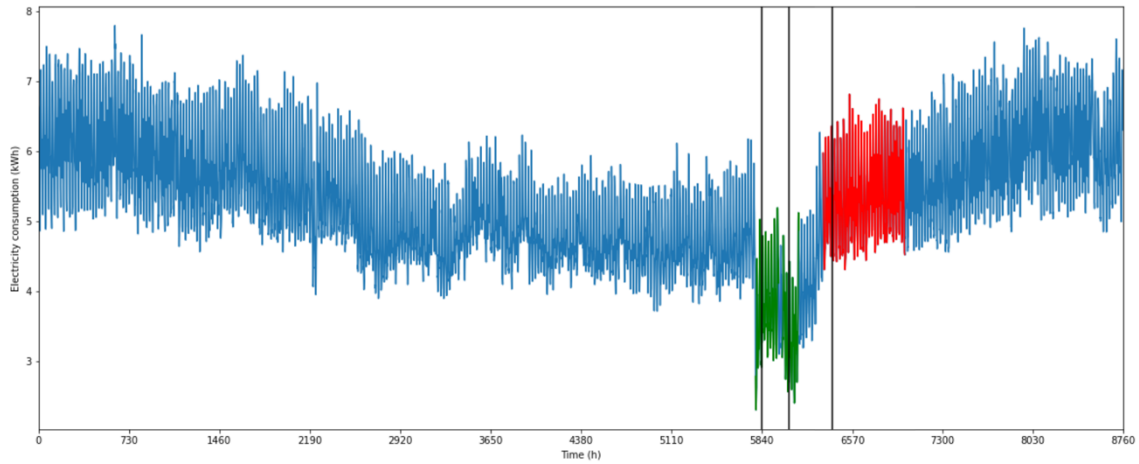
The shape and quality of the data are important to both models but in different ways. The Gaussian process regression model can be observed to perform well as long as the data retains a somewhat continuous pattern from day to day. Meanwhile, the K-means model is less dependent on these short-term patterns, but places more demand on similar seasonal profiles being present in the data. When the right prerequisites are present an accurate model can be created, which allows for much faster anomaly detection. For these profiles, reliable anomaly detections may be provided within the first 1-3 days. The illustrations below show one such case, the same profile with the same anomaly being detected for both the K-means and Gaussian process models.


Electricity consumption (kWh)
Time (h)
0
730
1460
2190
2920
3650
4380
5110
5840
6570
7300
8030
8760
3
4
5
6
7
8

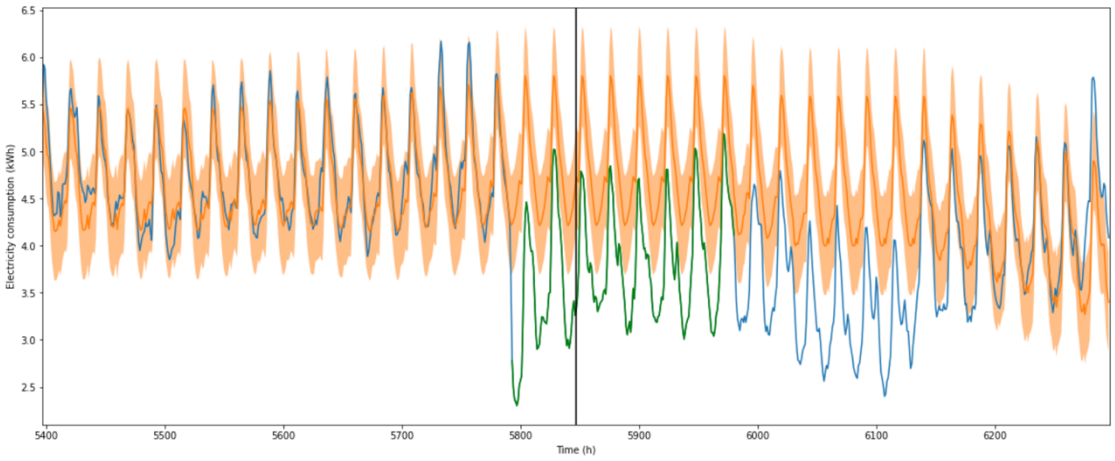
(a)


Electricity consumption (kWh)
Time (h)
5400
5500
5600
5700
5800
5900
6000
6100
6200
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5

(b)


Electricity consumption (kWh)
Time (h)
0
730
1460
2190
2920
3650
4380
5110
5840
6570
7300
8030
8760
3
4
5
6
7
8

(c)

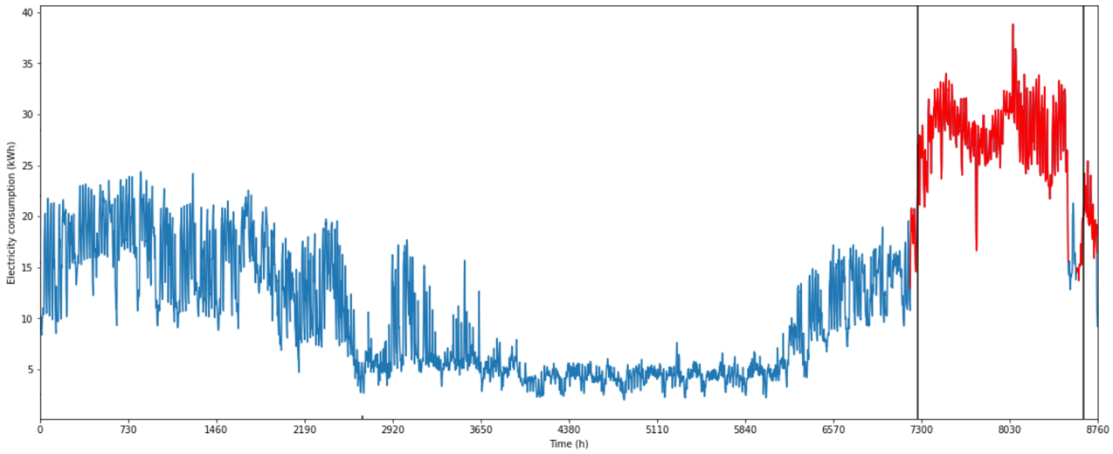


(d)

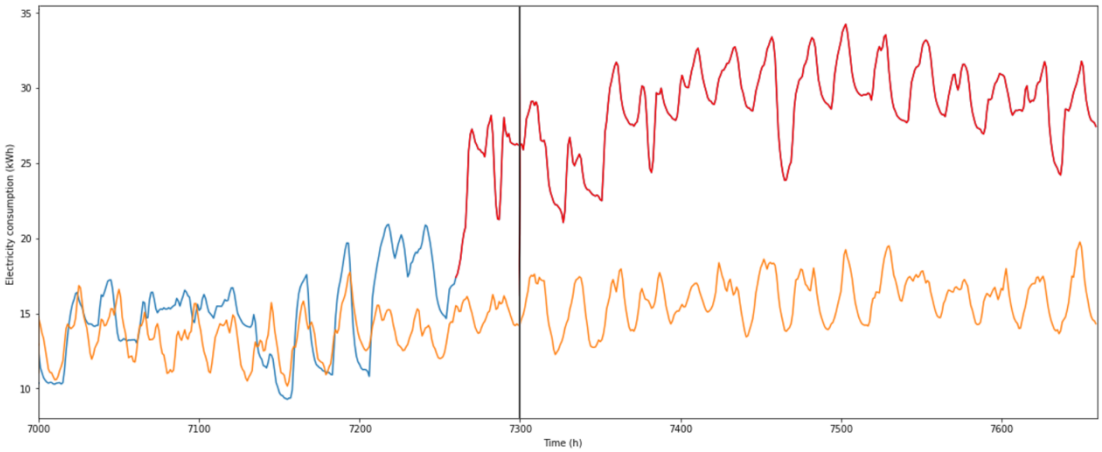
Figure 22, a substation with the two anomaly detection models implemented.

Figure 22(a) and Figure 22(c) illustrates the K-means model and the Gaussian process regression model on a yearly basis. From these figures it can be seen that the two models detect the same anomaly. Figure 22(b) and Figure 22(d) illustrates the anomaly detection model in the interval near the anomaly for the K-means model and the Gaussian process regression model respectively.

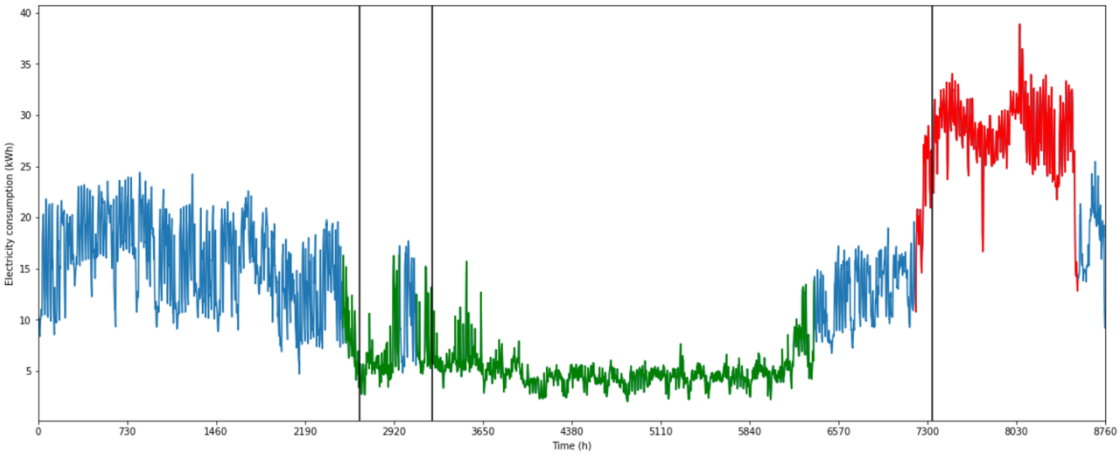
In this specific instance, the K-means model could detect the anomaly within 30 hours while the Gaussian process model required around 50 hours. Figure 22(c) shows the Gaussian process re-learning and detecting an additional anomaly within the region that the K-means model considers abnormal, the second anomaly is the consumption being too low at approximately hour 6000. Similarly, it detects the return to normal after the abnormal interval. Whether such a behavior is desirable or not may be argued, however it showcases the Gaussian process regression model's ability to learn new patterns relatively fast, unlike the K-means model. Uppsalahem's existing model also finds a deviation in September which corresponds to the anomaly that these models are able to detect.


Electricity consumption (kWh)
Time (h)
0
730
1460
2190
2920
3650
4380
5110
5840
6570
7300
8030
8760
5
10
15
20
25
30
35
40

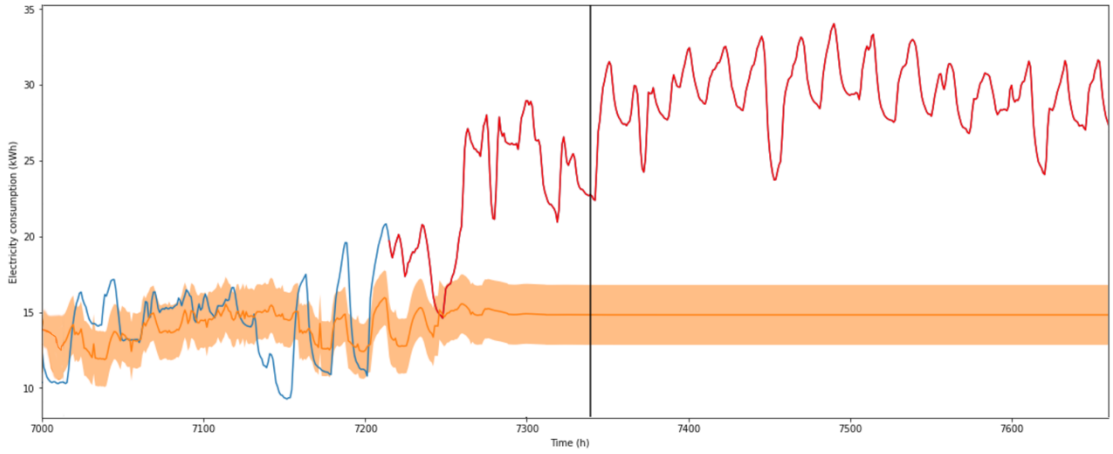
(a)


Electricity consumption (kWh)
Time (h)
7000
7100
7200
7300
7400
7500
7600
10
15
20
25
30
35

(b)


Electricity consumption (kWh)
Time (h)
0
730
1460
2190
2920
3650
4380
5110
5840
6570
7300
8030
8760
5
10
15
20
25
30
35
40

(c)



(d)

Figure 23, a substation containing anomalies displayed for the two models.

Figure 23 displays the two models implemented on another substation. Figure 23(a) and Figure 23(c) displays the whole year for the K-means model and the Gaussian process regression model respectively. Figure 23(b) and Figure 23(d) illustrate the detected anomaly at approximately hour 7200 for the K-means model and the Gaussian process regression model respectively. For this case the K-means model detects the anomaly circa 24 hours before the Gaussian process model.

The Gaussian process regression model further detects two more anomalies that the K-means model does not detect. Both the anomalies that the Gaussian process regression detects are cases where there is a sudden drop in consumption. The reason that the second anomaly is so long is due to the fact that the consumption does not return to what the model considers to be "normal" consumption until late summer (approximately hour 6400). Whether these events are in fact interesting to flag as anomalies is arguable, as a similar consumption pattern was present the year before, as seen in Figure 24, suggesting it is a recurring pattern.

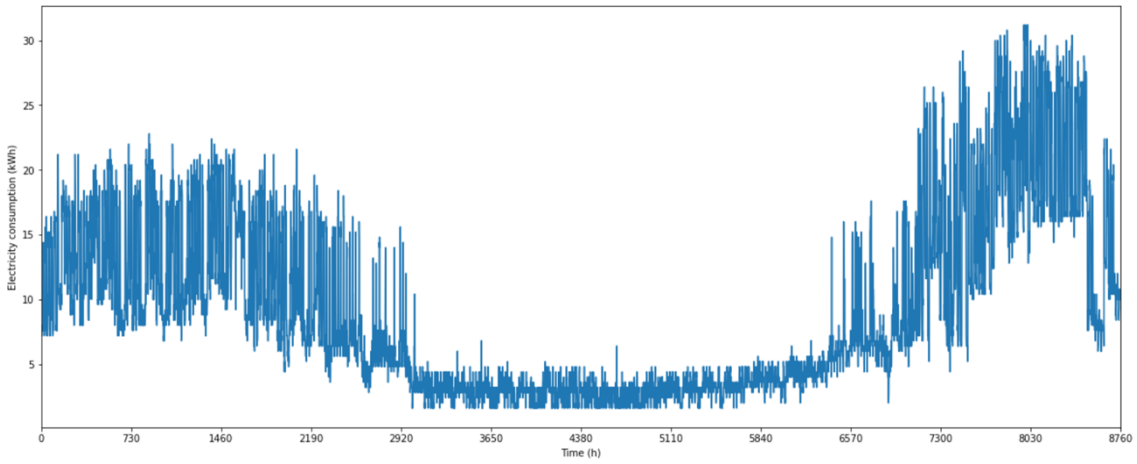


Figure 24, the investigated substation with corresponding data for 2018 year's consumption

Uppsalahem's model detects deviations as overconsumption compared to the last year for every month except for February through April for the data series seen in Figure 24. This contrasts with the two models generated in this thesis, for which shorter anomalies are found. The diagram above shows the corresponding data for 2018 which the current comparisons are based on. It shows no clear difference in profile shape but the average monthly electricity consumption for 2019 was 25 % higher or more for most months.

Despite their marked differences, the two models are arguably more similar to each other than any of them is to the current anomaly report system at Uppsalahem, due to the fact that they both operate on hourly data and that they both base their expected behavior on something which happens in or near the current time, rather than making comparisons a full year back.

As can be seen from the figures above the K-means based anomaly detection model detected both anomalies faster for the cases studied. This is largely rooted in the fact the K-means model considers the size of the anomaly and not only the time. A similar approach was considered for the Gaussian process regression model, but it was considered unfeasible to combine it with the necessity of allowing some values of the anomalous intervals to exist within the bounds of the model.

It is noteworthy that due to the absence of any validation data for the anomaly-finders, the timeframes established above are only preliminary, and it is expected that in any real implementation if sufficient testing data could be acquired these timeframes could change.

To summarize, while the Gaussian process model performs well situationally, and may be implemented using less training data than the K-means model, it is more sensitive to data quality and has a hard time learning behaviors of many of the more irregular profiles in the set. Meanwhile the tuning process for the Gaussian process is more complex than that of the K-means model while also yielding less reliable results.

5 Discussion

The section begins with Subsection 5.1, Reflections, which aims to highlight some reflections from the authors. The following subsection, Choice of modelling techniques, aims to evaluate the modelling techniques chosen in this project and their adequacy for the studied dataset. Possible changes which may be of interest for further improvement of the models as well as other modelling techniques which may yield more accurate results are also brought up within this section. The section is further split between the analysis of cluster models and regression models due to the significant differences in functionality between the two. This is followed by a critical look at the dataset on which the study is based, and offers suggestions of additional data which might be of interest to collect and other ways of gathering high resolution electricity data which may be beneficial to pave the way for the development of more accurate algorithms for prediction and anomaly detection for the concerned residential stock. Finally, the discussion ends with a short commentary on the possibilities of an online implementation of the developed anomaly detection models and the challenges which may be encountered during such an implementation.

5.1 Reflections

The goal defined at the beginning of this thesis work was to create a model or models which can aid in the detection of anomalies affecting the electricity consumption at a residential property owner. As is often pointed out in the paragraphs above, the lack of labelled data of anomalies has been a major obstacle in this process, for which reason the developed models are often described as finding *anomalies* rather than faults, where an anomaly is a behavior which is deemed by the authors as irregular, and thus possibly indicative of a fault affecting the consumption. However as there exists no complete validation data on which anomalies are desirable to find, the project is restricted to showing the possibilities of detecting behaviors which have optically been identified as points of interest by the authors, but it has not been possible to optimize anomaly detection bounds or to provide any overall measurement of accuracy of the anomalies detected. The optimization of parameters is otherwise highly desirable due to the differing degree of irregularity between the different substations. Throughout the development of the anomaly detection models, the authors have had to identify "reasonable" anomalies optically in order to determine objectives for the tests of the two models.

The eventual usefulness and interpretation of the output generated from these models is however uncertain, and an experienced facilitator may in fact come to different conclusions about the profiles at display and which consumption patterns may be of interest to identify. It should also be noted that the behavioral patterns which are considered as anomalies within this project are often far from exceptional, but that they appear rather frequently within the studied dataset, suggesting that most instances may in fact not be indicative of faults. This belief has been further corroborated by contacts with the project supervisor Tomas Nordqvist, who often knew of plausible explanations to consumption patterns which the project authors identified as abnormal, such as electrical heating or cooling equipment linked to the different substations. Some of the detected anomalies, such as the examples seen in Figure 12 and Figure 13, where observed data suddenly shifts to a new base level, are also likely to represent new

installations of such equipment, which happens regularly for the studied interval. Such changes to the profiles create further difficulties for the task of creating models. It stands clear that while Uppsalahem currently have a rather extensive system for anomaly detection, much of the information which may be needed to draw conclusions from the detection of anomalies is not catalogued in the database but rather tied up in tacit knowledge within the company.

Furthermore, the large presence of such hard to explain consumption patterns in the data has likely been detrimental to the performance of the models developed. At the start of the project there was an implicit assumption that the electricity data studied should, at least for the most part, amount to an aggregation of the electricity consumption of households, and models were developed from this assumption. Later, as model performance turned out poor for some of the cases, explanations were gradually sought in the data, from which many patterns which contradicted such an assumption were found. The project thus serves as an example of the importance of doing a pre-study of data. As of now, knowledge of that nature has rather been gathered by first trying to model the dataset. While the results from this project do not show what might be done with other modelling techniques, the wide differences in the regularity of data and presence or absence of diurnal patterns strongly suggest that developing models for this dataset is a hard task and that regression models may largely have to be adapted to the individual substations. This also impacts the possible identification of abnormal behavior, and the results of this study suggest the timeframe for identifying such behaviors might be restricted to a few days or even weeks for some of the less predictable profiles. Reliable anomaly detection on a sub-daily timeframe was not observed to be possible for either model for any part of the studied dataset.

5.2 Choice of modelling techniques

This subsection discusses the two different models developed within the project, compares their performance to what has been observed in other previous studies and suggests some possible improvements. It is divided between Subsection 5.2.1, in which the K-means model and other clustering approaches are discussed, and Subsection 5.2.2 which focuses on the Gaussian process regression model and also suggests other potentially interesting regression models.

5.2.1 Clustering

The silhouette index and the elbow method for the metrics of WSS and within-cluster R^2 are utilized to determine the optimal number of clusters. The silhouette index points towards the optimal number of clusters being two, however it is not deemed relevant as the low value also points to a lack of separated clusters. Therefore, the elbow method was chosen as the method for determining the optimal number of clusters. This decision is, however, arguable, and different researchers handle the silhouette score metrics for electricity data differently. Yilmaz et al. (2019) clusters 300 households using the K-means method and arrives at an optimum of $K=3$ with a silhouette index of 0.28. This can be compared to the silhouette score of 0.12 for $K=2$ established in this study, implying that the data clustered in Yilmaz et al (2019) displays slightly more separated clusters, but that there is still a significant overlap between the clusters. Other researchers similarly to this thesis calculate the silhouette index but choose other metrics to determine the optimum.

Wen et al. (2019) study a dataset of electricity consumption for 5000 Irish homes and businesses, and calculate a silhouette index of 0.8 indicating an optimum of $K=2$ after a Principal Component Analysis of the data series. However, they choose to select their optimum from the metrics of Mean Index Adequacy and Clustering Dispersion Indicator, where the first is an average of distances between cluster centers and data points, similar to the WSS measurement utilized in this report and the second measures a between-cluster compatibility. From these metrics, they conclude that the optimal number of clusters is instead 7. (Wen et al., 2019) These results are very similar to those found in this thesis, namely that silhouette index is optimized at 2 clusters, but the optimal number of clusters is higher. Sun et al. (2020) cluster approximately 4000 households to generate load profiles and calculate the silhouette index, indicating that the optimal number of clusters are 2. There are three points of the silhouette score that are analyzed in the article with regards to the silhouette score, 2, 4 and 6 clusters. When utilizing another method, the optimal number of clusters are determined to be 6. This means that also according to Sun et al. (2020) the "optimal" silhouette score should not be utilized.

The low silhouette score attained seems reasonable when the dataset is considered. The silhouette score can be described as a quota between cluster compactness and cluster separation, and when the dataset clustered is that of normalized electricity profiles of residential buildings, it seems intuitive that the degree of separation is low. Where there is a clear distinction between data points, e.g., homes and businesses, a higher degree of separation may be expected, but when the set is entirely composed of residential buildings the data is less likely to present clearly separated clusters. The low silhouette score presented thus seems reasonable, although the clustering may also benefit techniques like Principal Component Analysis to bring out features where the data is more clearly separated.

The WSS is arguably the more relevant metric for the anomaly detection algorithm developed, as it describes the expected error between cluster centroids and individual data series. The number of 6 clusters functioned well for the available dataset, but the optimal number is likely to differ depending on the amount of data clustered. For the K-means anomaly detection algorithm to perform well, there are two features that are identified to be of high importance. The first is that clusters are compact enough for the centroid to function as a good representation of the individual elements in the cluster. The second is that clusters should contain a high enough number of elements so that an anomaly in a single individual consumption profile does not affect the cluster centroid noticeably, so that it may trigger anomalies in other normal profiles.

There exist several K-means-based clustering algorithms such as Fuzzy C-means and K-medoids which have also been tried for electricity data. An additional family of clustering algorithms considered for the purpose of this study was that of hierarchical clustering, with criteria such as single-link, complete-link and Ward's method. The results of changing clustering algorithms between these alternatives were, however, not deemed to be different enough to support further evaluation. Another option which was considered was the possibility of changing distance metrics. Euclidean distance may not be ideal for representing similarities between time series. Dynamic Time Warping, which is a measurement specifically for time series, was considered, but dropped due to

high computational demand for the full year profiles. It may however be an attractive alternative for clustering time series when the amount of measurements is smaller. One of the main opportunities of improvement for this model is that it does not recluster itself, which would likely be necessary for a real implementation. Ideally, an interval for retaining good cluster representativity while also enabling the capturing of anomalies should be established, however this is similarly prevented by the lack of external definitions of what an anomaly is.

The conclusion from the comparisons to other researchers is that the data that is studied is important for the outcome of the study. However, the articles discussed in this subsection corroborate the method and the results attained in this thesis. Furthermore, it is also implied that the silhouette index is sometimes a poor indicator of the optimal number of clusters in terms of electricity consumption.

5.2.2 Regression

The Gaussian process regression model developed within this project is useful for a share of the dataset, where it attains accurate predictions, but for the norm case within this dataset it does not perform as well as expected, which is also a hindrance for an effective full-scale anomaly detection. The high degree of variance in model performance between the different data series however suggests that the model is able to perform well in certain contexts, and that the less favorable performance in other cases may have explanations in the variety of the data. The utilization of Gaussian processes for predictions of electricity consumption, seems reasonable based on previous literature. Predictions are also shown to be possible and a general improvement is observed compared to the persistence model, even without an exhaustive cross validation procedure being conducted on an individual basis.

Two different explanations of the low performance for some part of the dataset have been identified within this study. The first is a lack of clear diurnal patterns in a subset of the data. This is observed during the cross-validation procedure when analyzing the performance of the persistence model. The second is the presence of what has been identified as anomalies, discontinuous changes in the daily patterns of electricity consumption, which likely disturb the model training process. Discontinuous training data easily offsets the learned hyper-parameters of the Gaussian process as they become based on the small but sharp fluctuation in the training data. Duvenaud (n.d.) describes this as a common pitfall of the common stationary kernels like RBF. Many of the examples showcased are, therefore, selected from the more smooth and regular profiles in the dataset. An option for handling this would have been to edit out these identified anomalies from the training data, but once again, this is a hard task without proper labelling of anomalies, and thus it was not considered doable for the complete dataset.

For these smooth profiles the Gaussian process shows promising results for the purpose of anomaly detection, and when focusing on such profiles most identified anomalies may be detected within a matter of days. It was also possible to utilize the standard deviation of the predictions as a bound for the fault detection algorithm, which is likely to provide the model with some flexibility as it may establish different standard deviations during different periods of the year based on temporary variances in the predictability of the data. It is however no complete solution for the establishment of

bounds for the anomaly detection, as the results imply the optimal confidence interval should be determined individually for the different substations.

It should also be noted that the Gaussian process regression model was capable of achieving much more accurate predictions for the dataset in general when predicting on a 1-hour time horizon instead of a full day. However, for the scope of this thesis it was necessary to utilize a longer prediction horizon. The anomalies which were identified to be of interest largely consisted of intervals spanning several weeks, and a recurring issue with using a regression model which updates itself continuously based on observed data was that the model would learn the anomalous consumption pattern and thus become biased towards anomaly detection. This issue was particularly prevalent at the 1-hour prediction horizon, as the model would base predictions largely on the observations of the previous few hours. So, while the Gaussian process could attain a high accuracy on a short prediction horizon, the objective of finding long-lasting anomalies demanded that a less accurate model be chosen. The 24-hour prediction horizon chosen for this project may be seen as an attempt to compromise between these two goals, as the resulting predictions retain relatively high accuracy, but the model would also occasionally present the problem of learning anomalous behaviors as was shown in Figure 20.

The Matern3+RBF kernel is chosen as the best performing of the evaluated kernels for the dataset as a whole. The results speak of sum-kernels generally being beneficial due to their increased flexibility, and it might be possible to further improve predictions through the construction of increasingly complex sum-kernels, although it is not explored deeper in this thesis. Van der Meer et al. (2018b) also predict electricity consumption using both dynamic and static Gaussian processes, and concludes that the same combination of the Matern3 and RBF kernel is ideal for their data for this reason. They predict electricity consumption for the upcoming half hour and report a MAPE in the 3-5 % range for their studied buildings. It should also be noted that van der Meer et al. (2018b) use a very different configuration of dependent variables, consisting of the past consumption values.

The variables considered in this study are time and outside temperature. Outside temperature was however discarded from most of the models created as it would sometimes lead to less consistent results for the hyper-parameter optimization and showed slightly reduced prediction capability in many cases. While weather variables are found to be relevant for electricity consumption forecasting by among others Yang et al. (2018), removing it allows this thesis to consider less complex models and eases the task of optimizing hyper-parameters. Additional variables which might be of interest include electricity prices and a more advanced selection of calendar variables, e.g. introducing a variable for weekends or weekdays. (Yang et al., 2018) Another variable that could be of interest is the number of daylight hours since it affects the lightning, however, it is likely closely related to the date.

It should be noted that in many otherwise comparable studies the modeled data is selected after a more discriminatory preprocessing than what is the case in this thesis. For instance, van der Meer (2018b) studies a set of 300 residential customers from the Sydney metropolitan area. After what is described as “a thorough data cleaning process” 54 of these remain. The chosen approach in this thesis is different; While preprocessing is conducted, it only concerns the cleansing of data which does not meet the criteria of

having a hourly resolution, and the goal has rather been to salvage as much of the data as possible through means like interpolation in order to see what might be done for the entire set of substations. Additionally, the differences between working with individual customer data and substation data should not be understated, as one of the core insights of this project has been that the substations contain more than simply residential electricity usage, which often adds to the challenge of prediction.

A possible improvement might be brought about by the technique of Gaussian process quantile regression, which is a modification of the standard Gaussian process regression popularly used for electricity consumption forecasting, as seen in Yang et al. (2018) and van der Meer (2018c). Quantile regression methods aim at estimating quantiles of the dependent variable. Quantile regression estimates are according to Yang et al. (2018) more robust to outlier observations in the dependent variable, and are, therefore, most relevant when the response displays high variability or randomness in its behavior. Alternatively, as the observed variability is highest in the effect peaks, which may in turn be linked to certain times of day, suggesting that a model which can fit varying amounts of white noise to different observations may improve performance. Heteroscedastic Gaussian processes have been used with success to model this phenomenon in electricity prices (Kou et al., 2014), and may have similar applications for energy demand data.

Alternatively, there are many other regression models which have been tried for prediction of electricity consumption. Some examples which have been mentioned before in this thesis are Support Vector Machines and Artificial Neural Networks. Zeng et al. (2019) compares these two, Gaussian process regression and Multivariate linear regression, and arrives at the conclusion that Support Vector Machines and Multivariate linear regression perform best for the buildings considered. An additional family of models worth consideration is ARIMA models. Van der Meer et al. (2018b) cite ARIMA as a state-of-the-art for predicting electricity load profiles and report lower error metrics for ARIMA models than they do for Gaussian process regression. This conclusion is however arguable, as Fan et al. (2014) on the contrary find that the ARIMA model does not predict building energy use very well. This difference is important to highlight since it clearly shows that the applicability of techniques differs from case to case depending on the approach and the available data. It is thus difficult to in advance determine a technique that will perform well for a given data set.

5.3 Potential for additional data collection at Uppsalahem

As of now, Uppsalahem only store data of electricity consumption on a monthly basis, however hourly resolution data is available from electricity service providers and should likely be attainable through cooperation with service providers or automatic measurement equipment. The results of this thesis, however, show that this dataset is very diverse and shows widely different patterns. To allow for reliable anomaly detection, it is, as mentioned, recommended to only consider a share of the substations where the models perform well. If the goal is to allow for a fast and reliable anomaly detection system for the building stock as a whole, it is suggested to record data with a higher resolution and accuracy. Additionally, it might be beneficial to collect data at a higher spatiotemporal resolution, e.g. connected to specific appliances.

Issues which have affected this project is an inability to fully explain the set of more irregular electricity profiles. As previously mentioned, many of the detected anomalies may in fact have reasonable explanations, and although some information may be gathered from the follow-ups of Uppsalahem's existing anomaly reports, commentary is sometimes short and difficult to interpret without previous knowledge of the facilities. Particularly, much of the knowledge required to interpret the model output, such as the installation of electricity intensive appliances, does not exist in the database, but rather as tacit knowledge within the company. A clearer documentation of this could aid not only in the development of this model but also in interpretation of the existing model.

However, when the properties of the data such as the varied patterns and frequent changes to the base consumption, and the issues these qualities present to the models developed in this project are considered, the merit of working on the current substation data may be argued, and it may be interesting to consider other options for gathering residential energy data. The vast majority of studies referenced for comparison earlier in the report conduct their studies on either *building* or *customer* data, where a customer normally represents a single household. Thus, they consider data on a more detailed level than what is done in this study. As mentioned, substations often contain more than an aggregate of residential customers, and the lack of a clear understanding of what appliances might impact electricity consumption at the individual substations is a hindrance in achieving accurate predictions. Thus, more detailed electricity consumption measurements, separating for instance single buildings or even appliances such as heat pumps, if possible, may allow for much more detailed analysis of the data. Gathering energy data on a building level may allow for analysis of interesting combinations of energy consumption data and existing data about the individual buildings such as year of construction, renovations, the floor area and number of floors etc., which may allow for detailed evaluation of variables which likely affect the energy performance of the individual buildings (Cai et al., 2019). Uppsalahem also already stores a relatively vast amount of this type of information about their buildings, which could be useful in such an endeavor.

An example of the opportunities provided by more fine-grained data is that it can be utilized to enable balance services such as frequency regulation, which are of interest for the grid owner. Collecting data every other second allows for such services. High frequency gathering of data is a key aspect of most adaptive regulations of electricity, and the hourly data available in relation to the current substations does not provide support for such frequency regulations. Collecting data at an even less aggregate level, for instance charging stations for electric cars or at grid inverters for solar photovoltaic installation and allow for a more detailed analysis of the specific appliances and a deeper understanding of their behavior. There are thus additional benefits of collecting more specific and more high-resolution data, which may also aid in the detection of anomalies. (Tapia, 2020)

5.4 Possibility of online implementation

The models developed within this study have been implemented for a set of 2-year long series of historical data, extracted from excel files downloaded from an electricity service prover's database. As the goal is to create models which may be utilized for the detection of anomalies in real time, it is highly relevant to provide some perspective of the technical and social prerequisites for an implementation of these algorithms.

The K-means anomaly detection model is dependent on the availability of data from the different substations in the cluster to determine the model. A problem which may arise from this is the occasional loss of individual measurements for different reasons, and a real implementation of the model may have to be changed to support model calculations while missing occasional single measurement data points.

The Gaussian process-based model in this project is subjected to a few changes to cope with computational demand, which is a potential issue for online implementations. It is however noteworthy that the scope of the testing process requires the handling of a rather large amount of data at a single time as predictions are generated for a full two-year scope at a single time, which is not the objective for a real-time implementation. It is, therefore, the author's decision that through effective allocation of resources (e.g. not re-training all individual models at the same time each month), the algorithm could be implemented for the entirety of the concerned property base. This is also backed up by other researchers like Zeng et al. (2019) who see no obstacle for an online implementation of a Gaussian process prediction model. The high computational complexity of the training process is however still an issue to be mindful of in any real implementation.

The question of how property facilitators may react to the implementation of these models is less certain. As has been established before the interpretations of the anomalies found by this model are not always clear. This similarly holds true for the anomaly report system currently in place at Uppsalahem. It should however be noted that facilitators have existing experience with the current system, and have attained knowledge of how to interpret the generated reports. For example, there may be known behaviors such as the system generating anomaly reports during cold months due to the use of car engine heaters. In an implementation of a model like the ones described in this thesis, it might take time for facilitators to attain similar knowledge of how to interpret the generated output. New technology can disrupt existing work routines and the organization must go through a learning process and make adjustments in order to bring it into practice, and failure to adapt new innovations is not uncommon, even when benefits are visible (Edmondson et al., 1999). The somewhat uncertain nature of anomaly detection algorithms, and the fact that many detected anomalies may often not be linked to faults affecting energy consumption, or in some cases not even fully explained, may entail that benefits of these models are not visible to begin with. Thus, the eventual attractiveness of the models developed in this project is largely dependent on how they are perceived by facilitators.

6 Conclusion

The aim of this thesis is to investigate whether the available data from measurements of electricity consumption at a residential property owner may be utilized to create machine learning models to enable a fast and reliable anomaly detection system. Two models are developed to investigate this, one based on K-means clustering and the other on Gaussian process regression.

The dataset obtained for this thesis is shown to be very diverse and it is concluded that the developments of machine learning models for anomaly detection should probably be restricted to a subset of the substations to enable more accurate modelling. Regarding the possibility of additional data gathering, several suggestions are provided. One opportunity exists in the gathering of more spatially specific data, such as electricity consumption connected to individual buildings, which could then be combined with the rather extensive existing data about the buildings to for instance make assessments of building energy performance. Additionally, regularly storing more data about anomalies in the energy consumption, such as if they are linked to specific faults affecting the energy consumption or if normal explanations for deviating energy consumption are found, would likely be of assistance in future attempts to create more accurate models.

Both developed models are shown to be capable of detecting what has been identified as anomalous patterns in electricity consumption, but the timeframe required for reliably doing so varies between a couple of days to weeks depending on how accurately the individual substations are modeled. The models are, however, assumed to be faster than the current anomaly detection system at Uppsalahem, which operates on monthly data. Both models present with strengths and weaknesses, but the K-means model is chosen as the most preferable due to the relative ease of implementation and higher capability of handling irregular data series, which are frequent in the studied dataset. The K-means model also shows slightly faster detection times in a comparison between the two models. It also displays a higher ability to detect deviations in the form of long-term drifts and seasonal anomalies, while the Gaussian process regression model may largely be restricted to finding rapid short-term changes in electricity consumption. The results of this study show that it is more straightforward to proceed with the K-means model.

References

Allouhi, A., El Fouih, Y., Kousksou, T., Jamil, A., Zeraouli Y., Mourad, Y., "Energy consumption and efficiency in buildings: current status and future trends ", *Journal of cleaner production*, 109, pp. 118-130, 2015

Amasyali, K., El-Gohary, N. M., "A review of data-driven building energy consumption prediction studies", *Renewable and sustainable energy reviews*, 81, pp. 1192-1205, 2018

Auffhammer, M., Baylis, P., Hausman, C. H., 2017. "Climate change is projected to have severe impacts on the frequency and intensity of peak electricity demand across the United States", *Proc. Natl. Acad. Sci. U. S. A*. 114 (8), 1886–1891.

Bang, M., Skals Engelsgaard, S., Kjøller Alexandersen, K., Riber Skydt, M., Reza Shaker, H., Jradi, M., "Novel real-time model-based fault detection method for automatic identification of abnormal energy performance in building ventilation units", *Energy & buildings*, 183 pp. 238-251, 2019.

Bedi, J., Toshniwal, D., "Deep learning framework to forecast electricity demand", *Applied energy*, 238, pp. 1312-1326, 2019

Bedingfield, S., Alahakoon, D., Genegedera, H., Chilamkurti, N., "Multi-granular electricity consumer load profiling for smart homes using a scalable big data algorithm", *Sustainable cities and society*, 40, pp. 611-624, 2018

Beveridge, S. Least squares estimation of missing values in time series. Commun. Stat. Theory Methods 1992, 21, 3479–3496.

Bilj, H. Schön, T. B., vanWinderden, J-W. Verhaegen, M. "Online Sparse Gaussian Process Training with Input Noise". 2016. ArXiv.:1601.08068v1 [stat.ML] available at: https://arxiv.org/pdf/1601.08068v1.pdf

Brusaferri, A., Matteuci, M., Portolani, P. Vitali, A. Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices. 2019. Applied Energy 250, pp. 1158-1175.

Bourdeau. M., Zhai, X. Q., Nefzaoui, E., Guo, X., Chatellier, P., "Modeling and forecasting building energy consumption: A review of data-driven techniques", *Sustainable cities and society*, 48, 101533, 2019

Burkov, A., "The hundred page machine learning book", January 2019.

Available 20200303 at http://themlbook.com/wiki/doku.php

Bynum, J. D., Claridge, D. E., Curtin J. M., "Development and testing of an automated building commissioning analysis tool (ABCAT)", *Energy and Buildings*, vol. 55, December 2012, pp. 607-617.

Byrd, R. H., Lu, P., Nocedal J., "A limited memory algorithm for bound-constrained optimization". 1996. SIAM J. Sci. Comput. 16 (5): 1190–1208.

Cai, H., Shen, S., Lin, Q., LI, X., Xiao, H., "Predicting the Energy Consumption of Residential Buildings for Regional Electricity Supply-Side and Demand-Side Management". 2019. IEEE Access, VOLUME 7.

Cheng, C. L., Shalabh, Garg, G., "Coefficients of determination for multiple measurement error models", *Journal of multivariate analysis*, 126, pp. 137-152, 2014

Cheng, H., Ding, X., Zhou, W., Ding, R., "A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange", *Electrical power and energy systems*, 110, pp. 653-666, 2019

Chicco, G., "Overview and performance assessment of the clustering methods for electrical load pattern grouping", *Energy*, 42, 68–80, 2012

De Jaeger, I., Reynders, G., Callebaut, C., Saelens, D., "A building clustering approach for urban energy simulations", *Energy and buildings*, 208, 109671, 2020

Duvenaud, D. "Automatic Model Construction with Gaussian Processes". 2014.

Duvenaud, D. "The Kernel Cookbook: Advice on Covariance functions", n.d. available 20200607 at: https://www.cs.toronto.edu/~duvenaud/cookbook/.

Elsheikh, A. H., Sharshirc, S. W. , Elazizd, M. A., Kabeelf, A. E., Guilang, W., Haioub, Z.,"Modeling of solar energy systems using artificial neural network: A comprehensive review", *Solar energy*, 180, pp. 622-639, 2019

Energimyndigheten, "Energiläget i siffror 2019", 2019, available 20200303 at: https://www.energimyndigheten.se/globalassets/statistik/energilaget/energilaget-i-siffror-2019.xlsx)

Esteri, H., Omran, B. A., Murphy, S. N., "kluster: An Efficient Scalable Procedure for Approximating the Number of Clusters in Unsupervised Learning ", *Big data research*, 13, pp. 38-51, 2018

Fan, C., Xiao, F., Wang, S., "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques", *Applied energy*, 127, pp. 1-10, 2014

Farshad, M., "Detection and classification of internal faults in bipolar HVDC transmission lines based on K-means data description method", *Electrical power and energy systems*, 104, pp. 615-625, 2019

Gasser, P., "A review on energy security indices to compare country performances", *Energy policy*, 139, 111339, 2020

Girard A., Rasmussen C. E., Candela J. Q., Murray-Smith R., "Gaussian process priors

with uncertain inputs-application to multiple-step ahead time series forecasting", *Advances in Neural Information Processing Systems*, pp. 545-552, 2003

Govender, P., Sivakumar, V., "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)", *Atmospheric pollution research*, 11, pp. 40-56, 2020

Horrigan, M., Turner W. J. N., O'Donnell J., "A statistically-based fault detection approach for environmental and energy management in buildings", *Energy and buildings*, 158, pp. 1499-1509, 2018

Hyndman, R. J., Athanasopoulos, G., Monash university Australia, Forecasting: Principles and practise, April 2018, available 20200407 at: https://otexts.com/fpp2/

Ingenjörsvetenskapsakademin, “Energieffektivisering av Sveriges bebyggelse”,2012a, available online 20200607 at: https://issuu.com/iva-publikationer/docs/energieffekt-rapport2

Ingenjörsvetenskapsakademin, “Energieffektivisering av Sveriges flerbostadshus”, 2012b, available 20200303 at: https://www.iva.se/globalassets/rapporter/ett-energieffektivt-samhalle/201206-iva-energieffektivisering-rapport1-f1.pdf

Katipamula, S., Brambley M. R., “Review Article: Methods for fault detection, diagnostics, and prognostics for building systems - A review, part 1”, *Hvac&R Research*, 11(1), pp. 3-25, 2005

Kim, W., Katipamula. S., “ A review of fault detection and diagnostics methods for building systems”, *Science and Technology for the built environment*, 24(1), pp. 3-21, 2017

Kjøller Alexandersen, K., Riber Skydt, M., Skals Engelsgaard, S., Bang, M., Reza Shaker, H., Jradi, M., “A stair-step probabilistic approach for automatic anomaly detection in building ventilation system operation” *Building and environment,* 157, pp. 165-171(2019)

Kou, P. Liang, D. Lou, J. “Probabilistic electricity price forecasting with variationalheteroscedastic Gaussian process and active learning”. 2014. Energy conversion and management 89. pp. 298-308.

Lepot, M., Aubin, J., Clemens, F. H. L. R., “Interpolation in Time Series: An Introductive Overview of Existing Methods, Their Performance Criteria and Uncertainty Assessment”, Water 2017, 9, 796

Lu, X., Zhou, K., Yang, S., “Multi-objective optimal dispatch of microgrid containing electric vehicles”, *Journal of cleaner production*, 165, pp. 1572-1581, 2017

Marszal-Pomianowska, A., Heiselberg, P., Kalyanova Larsen, O., “Household electricity demand profiles e A high-resolution load model to facilitate modelling of energy flexible buildings”, *Energy*, 103, pp 487-501, 2016

Masud, M. A., Huang, J. Z., Wei, C., Wang, J., Khan, I., Zhong, M., “I-nice: A new approach for identifying the number of clusters and initial cluster centres”, *Information sciences*, 466, pp. 129-151, 2018

McNeil, M. A., Karali, N., Letschert, V., “Forecasting Indonesia's electricity load through 2030 and peak demand reductions from appliance and lighting efficiency”, *Energy for sustainable development*, 49, pp. 65-77, 2019

Miller, C., Nagy, Z., Schlueter, A., “Automated daily pattern filtering of measured building performance data”, *Autom. Construct*, 49, pp. 1–17, 2015

Miller, C., Schluter, A., “Forensically discovering simulation feedback knowledge from a campus energy information system”, *Proceedings of the symposium on simulation for architecture & urban design*, pp.136-143, 2015

Musial, J. P., Verstraete, M. M., Gobron, N., Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. Atmos. Chem. Phys. 2011, 11, 7905–7923.

Ovoenergy, “What is energy efficiency”, available 20200606 at: https://www.ovoenergy.com/guides/energy-guides/what-is-energy-efficiency.html

Pham, A., Ngo, N., Truong, T., Huynh, N., Truong, N., "Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability", *Journal of cleaner production*, 260, 121082,2020

Pimentel, B. A., de Carvalho, A. C. P. L. F., "A Meta-learning approach for recommending the number of clusters for clustering algorithms", *Knowledge-based systems*, 195, 105682, 2020

Rasmussen, C. E., & Williams C. K. I., Gaussian Processes for Machine Learning, the MIT Press, 2006,ISBN 026218253X. c 2006 Massachusetts Institute of Technology.*

Scikit-learn, n.d. "1.7 Gaussian processes", available 20200608 at: https://scikit-learn.org/stable/modules/gaussian_process.html

Seem, J. E., "Using intelligent data analysis to detect abnormal energy consumption in buildings", *Energy and buildings*, 39, pp. 52-58, 2007

Seyedzadeh, S., Pour Rahimian, F., Glesk, I., Roper, M., "Machine learning for estimation of building energy consumption and performance: a review", *Visualization on engineering*, 6:5 2018.

SMHI, "Ladda ner meteorologiska observationer", n.d., available 20200612 at: https://www.smhi.se/data/meteorologi/ladda-ner-meteorologiska-observationer#param=airtemperatureInstant,stations=all

Stein, M. L., "Interpolation of Spatial Data", Springer-Verlag, New York., 1999

Singh, S., Yassine, A., "Big Data Mining of Energy Time Series for Behavioral Analytics and Energy Consumption Forecasting", *Energies*, 11, 452, 2018.

Sun, L., Zhou, K., Yang, S., "An ensemble clustering based framework for household load profiling and driven factors identification", *Sustainable cities and society*, 53, 101958, 2020

Tan, P., Steinbach, M., Kumar, V., *"Introduction to data mining, global edition"*, 2018-04-01.

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., P. Finn, D., " Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach", *Building and environment,* vol. 140, pp. 90-106, August 2018.

Towers, S., "K-means clustering", 20131024, available 20200406 at: http://sherrytowers.com/2013/10/24/k-means-clustering/

Uppsalahem, "Hållbarhet", n.d.a, available 20200303 at: https://www.uppsalahem.se/om-oss/hallbarhet/

Uppsalahem, "Om oss", n.d.b, available 20200303 at: https://www.uppsalahem.se/om-oss/

van der Meer, D. "Spatio-temporal probabilistic forecasting of solar power, electricity consumption and net load". 2018a.

van der Meer, D. W., Shepero, M., Svensson, A., Widén, J., Munkhammar, J., "Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes", *Applied energy*, 213, pp 195-207, 2018b

van der Meer, D. W., Widén, J., Munkhammar, J., "Review on probabilistic forecasting of photovoltaic power production and electricity consumption", *Renewable and sustainable energy reviews*, 81, pp. 1484-1512, 2018c

van Every, P. M., Rodriguez, M., Jones, C. B., Mammoli, A. A., Martínez- Ramón, M., "Advanced detection of HVAC faults using unsupervised SVM novelty detection and Gaussian process models", *Energy build.*, 149, pp. 216–224, 2017

Viegas, J. L., Vieria, S. M., Melício, R., Mendes, V. M. F., Sousa, J. M. C., "Classification of new electricity customers based on surveys and smart metering data", *Energy*, 107, pp. 804-817, 2016

Wen, L., Zhou, K., Yang, S., "A shape-based clustering method for pattern recognition of residential electricity consumption", *Journal of cleaner production*, 212. pp. 475-488, 2019

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., Zhao, X., "A review of data-driven approaches for prediction and classification of building energy consumption", *Renewable and sustainable energy reviews*, 82, pp. 1027-1047, 2018

Widén, J., Lundh, M., Vassileva, I., Dahlquist, E., Ellegård, K., Wäckelgård, E., "Constructing load profiles for household electricity and hot water from time-use data—Modelling approach and validation", *Energy and buildings*, 41, pp. 753-768, 2009

Yilmaz, S., Weber, S., Patel, M. K., "Who is sensitive to DSM? Understanding the determinants of the shape of electricity load curves and demand shifting: Socio-demographic characteristics, appliance use and attitudes", *Energy policy*, 133, 110909, 2019

Yildiz, B., Bilbao, J. I., Sproul, A. B., "A review and analysis of regression and machine learning models on commercial building electricity load forecasting", *Renewable and sustainable energy reviews*, 73, pp. 1104-1122, 2017

Zeng, A., Liu, S., Yu, Y., "Comparative study of data driven methods in building electricity use prediction", *Energy & buildings*, 194, pp. 289-300, 2019

Zhang, Y., Huang, T., Bompard, E., F., "Big data analytics in smart grids: a review", *Energy informatics,* 1:8, 2018

Zhang, Z., "Understand data normalization in machine learning", created 20190327, available 20200417 at: https://towardsdatascience.com/understand-data-normalization-in-machine-learning-8ff3062101f0

Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., Li, J., "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis", *Energy and built environment*, 1, pp. 149-164, 2020

Zhao, Y., Li, T., Zhang, X., Zhang C., "Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future", *Renewable and Sustainable Energy Reviews*, 109, pp. 85-101, 2019

Zheng, K., Wang, Y., Chen, Q., Li, Y., "Electricity theft detecting based on density-clustering method", *2017 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*. Available 200302 at: https://ieeexplore.ieee.org/abstract/document/8378347/authors#authors

Zytkow, J., Rauch, J., “Principles of data mining and knowledge discovery”, Third european conference PKDD, 1999, available 20200408 at: https://books.google.se/books?id=2Q1qCQAAQBAJ&pg=PA28&lpg=PA28&dq=moving+average+preprocessing&source=bl&ots=mARDljOA0O&sig=ACfU3U0ouIRT_9s8QBEZUVKZgKlOjrAi3g&hl=sv&sa=X&ved=2ahUKEwjzxrmbrtjoAhXhAhAIHX6nDx8Q6AEwCXoECAwQLA#v=onepage&q=moving%20average%20preprocessing&f=false

Interviews

(Tapia, Camilo; Customer project manager, Power2U, 2020)

(Nordqvist, Tomas; Head of energy department, Uppsalahem, 2020)