Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Intelligent Computing

Master's Thesis

Ivan Osipov

FORECASTING OF ENERGY BALANCE IN GREEN CAMPUS

Examiners:	Prof. Lasse Lensu
	Associate Prof. Igor Anantchenko
Supervisors:	Adjunct Prof., Dr. Xiao-Zhi Gao
	M.Sc. (Tech.) Ville Tikka
	Associate Prof. Arto Kaarna
	Prof. Samuli Honkapuro
	Prof. Lasse Lensu
	Associate Prof. Igor Anantchenko

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Intelligent Computing

Ivan Osipov

Forecasting of energy balance in green campus

Master's Thesis

2017

59 pages, 32 figures, 13 tables.

Examiners: Prof. Lasse Lensu
Associate Prof. Igor Anantchenko

Keywords: data analysis, forecasting, energy balance, machine learning, time series

The focus of this thesis is the forecasting of the energy balance. Energy balance can be decomposed to the energy consumption forecasting and energy production forecasting, both of which can help users in making decisions related to the energy storage. It is difficult to store electricity so using forecasting the users can decide if they should plug in more devices or purchase additional electricity.

Time-series data analysis is a well-investigated field, so there are several methods that can be used to implement a predictor of electricity production and consumption. The most commonly used methods of forecasting have been reviewed in this work. Analysis and preprocessing of the input data provided has been developed. Linear Regression and Seasonal Autoregressive Integrated Moving Average and adjusting of settings have been used for the forecasting of electricity consumption and production. The accuracy measurement scores of the forecast have been evaluated.

According to the experiments, different methods are suitable for production and consumption data. This is likely to occur because of the differences in seasonality.

PREFACE

I would like to thank my supervisors from LUT. Lasse, Arto, Xiao-Zhi, with your support and everlasting faith you gave the power to move forward. Thanks to my parents and bro for everything what they gave me. Thanks to Kostya, Daria, Anna, Dmitry and other friends here in Lappeenranta for warm family-style parties.

I would like to say a few words. And here they are: Nitwit! Blubber! Oddment! Tweak!
Thank you.
Lappeenranta, May 25, 2018

Ivan Osipov

CONTENTS

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
ACF	Autocorrelation Function
ANN	Artificial Neural Network
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
CART	Classification and Regression Tree
CV	Coefficient of Variation
FI	Fixed Installation
FMI	Finnish Meteorological Institute
GP	Gaussian Process
GDP	Gross Domestic Product
LMBM	Levenberg-Marquardt Backpropagation Method
MA	Moving Average
MAD	Median Absolute Deviation
MAE	Mean Average Error
MAPE	Mean Absolute Percentage Error
MaxAE	Maximum Absolute Error
ML	Machine Learning
MRE	Mean Relative Error
MVG	Multi-Variate Gaussian
MSARIMA	Multiplicative Autoregressive Integrated Moving Average
NAN	Not a Number
NN	Neural Network
PACF	Partial Autocorrelation Function
PV	Photovoltaic
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with eXogenous Regressors
SP	Single Panel
SCGB	Scaled Conjugate Gradient Backpropagation
STLF	Short Term Load Forecast
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression

LIST OF SYMBOLS

a	Bias term
$a_0...a_n$	Coefficients of Regression
d	Degree of the First Differencing
D	Seasonal Degree of the First Differencing
n	Number of elements
p	Value of the Autoregressive Part
P	Seasonal Value of the Autoregressive Part
q	Moving Average Part
Q	Seasonal Moving Average Part
r^2	R-squared statistical measure value
u	Error in the Linear Regressor
$X_1...X_n$	Features of the Model
Y_i	Real Consumption or Production Value
$\hat{Y_i}$	Predicted Consumption or Production Value

1 INTRODUCTION

1.1 Background

Production and consumption has to balanced at every moment. The consumption and production of electricity is worth forecasting [1]. This is necessary, because inaccurate estimation of the production and consumption can cause purchasing excessive energy, that causes additional outcomes. Inaccurate production may also cause the lack of electricity if the forecast is much higher than the real production value [2].
49% of 35 investigated case studies consider the forecast for one hour. In 19% forecasting goes for 1 day [3].

If there is more demand than supply, frequency of the power system decrease. This cause generators to be disconnected from grid (to avoid their failures), which makes situation worse, and eventually, this will cause large black-out, if there are not any corrective actions. Hence, it is utmost important that power balance in whole power system is maintained. In small sub-system, as in the campus area, energy balance is more important from the economic reason. By better forecast, it is possible to use more own generation, and save money.

This thesis is related to the Green Campus project of LUT. The production of energy is worth forecasting also because campus not only consumes electricity but also produces it. Electricity production is based on innovative, unconventional methods: in the campus area there are 835 solar panels [4] which are united into six sets. Unfortunately, accurate forecasting is difficult because there are numerous factors involved, such as the weather condition and residence building types. The relationships of these factors between the electricity usage and weather factors can be highly nonlinear.

The production depends on a number of factors, so it is necessary to study them. This includes finding the correlation values between them and the produced electricity to include the relevant factors into the forecasting model. So one of the important tasks is to investigate influential and non-influential factors and take them into account while designing the forecasting model. Another main step is to study possible forecasting methods and find the best performing ones.

1.2 Objectives and Delimitations

The objective of this work is to develop an approach which gives the estimated value of consumed and produced electricity per hour for several hours ahead (at least one). Electricity consumption is considered as a time series with certain periodicity (for instance, seasonal, monthly, weekly).The following sub-objectives can be defined:

- Review the existing forecasting methods for the purpose and select suitable ones for further study.
- Investigate measurement data, seasonality and other factors, affecting the forecast.
- Design and implement model(s) for energy production and consumption forecasting for at least one hour ahead.
- Evaluate the prediction accuracy of each model.

1.3 Structure of the Thesis

This Master's Thesis consists of 7 chapters. Chapter 2 is an overview of common approaches which can be applied to forecasting. Chapter 3 describes methods with which the following work will be done and gives instructions related to the implementing of these forecasting models. Chapter 4 is related to the preprocessing of input data and how to evaluate the accuracy of forecasting. Chapter 5 is related to the complete solution of the approach based on the Linear Regression, it also describes the experiments and achieved results. Chapter 5 reflects the solution of the SARIMA model, it also describes the experiments and achieved results. Chapter 6 is related to discussions about research which has been made. Chapter 7 consists of the conclusions given with evaluation of the goals set.

2 METHODS FOR TIME-SERIES FORECASTING

There is a large variety of time-series prediction methods available for electricity consumption forecasting, all of these methods are built on different algorithms. Methods can be divided in different ways, for instance, into Machine Learning (ML) based methods and regression-based methods. However, performances of different methods can vary significantly and are often heavily case-dependent [3].

It is not necessary to rely on a single method - there is a way to implement several method and unite them into an ensemble. One of the methods has the primary role and others are supporting for instance main forecasting model is Neural Network (NN) and Gaussian Process (GP) is supporting it.

For most of the approaches, the workflow with the time-series prediction method can be divided into obtaining a data structure and model fitting [5]. Mathematical or statistical dependencies are determined between the value of load and input factors which affect it.

2.1 Machine Learning Based Methods

2.1.1 Neural Networks

Neural networks are often used in various applications [5]. The main steps in the neural network development are to choose the input data, features, select the architecture of the network, divide the input data and train the neural network. Accuracy in the considered cases of forecasting [2] is quite high - in one case the R squared measure (R_2) value is equal to 0.81, in another case the Mean Absolute Percentage Error (MAPE) is 2.88$\%$, in one more case, it is 7$\%$ for one day-ahead forecasting. The main advantage of Artificial Neural Network (ANNs) is their prominent performance for both data classification and approximation of the function [6]. An ANN is also able to detect dependencies in the given (training) data without additional factors (i.e., how it is done in regression models).

2.1.2 Support Vector Machine

In a common case, to ensure the forecasting accuracy different Support Vector Machine (SVMs) and parameters can be applied. SVM can perform a nonlinear classification with the kernel trick, implicitly mapping the inputs into high-dimensional feature spaces. For every type of load pattern, the corresponding dataset is selected to train the SVM model. For the training it is necessary to determine relevant values. So for each electricity consumer, a set of SVM models is developed. After the determination of the consumption pattern (for instance, with a decision tree) the SVM forecasting model with suitable parameters is used to ensure the forecasting accuracy [7]. The training dataset must cover a wide range of input patterns sufficient enough to train the model to recognize and predict the relationship between the input variables and target output. One of the methods of data selection is to calculate the distance between the forecasted input variable and its desired outcome. Then data which does not satisfy this condition is discarded [8], for instance, with the Kalman filter. SVMs which solve classification problems are named Support Vector Classification (SVC) methods [9] and SVM's suitable for modeling and prediction are named Support Vector Regression (SVR) methods.

2.2 Regression-Based Methods

2.2.1 Autoregressive Integrated Moving Average

Autoregressive Integrated Moving Average (ARIMA) models are based on transforming the time series to be stationary by the differencing process. The input consists of lags of the dependent variable along with lags of the forecast error.
A couple of cases with the ARIMA usage have been considered [10–12]. MAPE varies from 0.0384% [5] to 3.78% for monthly prediction.
Seasonal Autoregressive Integrated Moving Average (SARIMA) method contains the seasonal component(s), for instance, season length, [13] which is defined by the user and can be obtained by observing the time-series plot or empirically. SARIMA can be applied together with fuzzy methods [14], Support Vector Machines [13, 15] and Markov Processes [10].
It is possible to add exogenous components to the SARIMA model during the model building [12] to achieve better forecasting results. This model is called Seasonal Autoregressive Integrated Moving Average with eXogenous Regressors (SARIMAX).

2.2.2 Linear Regression Methods

The common linear regression algorithm can be used for forecasting. When the values of features is a new value (i.e., not part of the data that were used to estimate the model), the resulting value is a genuine forecast [5]. In a case related to forecasting of electricity consumption in New Zealand [16], the following model is given:

$$Y = a + b_1 * X_1 + b_2 * X_2 + u \tag{1}$$

where Y is the predicted electricity consumption, X_1 is the country Gross Domestic Product(GDP), X_2 is the cost of electricity, a is the bias, b_1 and b_2 are the regression parameters and u is the error [16]. This model has been thoroughly tested with statistical tests. Another way to calculate is the linear logarithmic form. Models from another case-study [17], take the form of a standart dynamic constant elasticity function:

$$\log(Y_{dom,t}) = a_0 + a_1 \log(X_{3,t}) + a_2 \log(PR_t) + a_3 \log(PR_{t-3}) + a_4 \log(Y_{dom,t-3}) \tag{2}$$

where the $Y_{dom,t}$ is the indoors electricity consumption, PR_t is the electricity price for domestic users, $X_{3,t}$ is the GDP per capita, a_0, a_1, a_2, a_3, a_4 are the coefficients of regression, and $t - i$ as subscript indicates the lag term (i.e. $t - 1$ indicates lag 1).

The outdoors model formula is:

$$\log(Y_{ndom,t}) = b_0 + b_1 \log(X_{1,t}) + b_2 \log(PRND_t) + b_3 \log(IND_{t-3}) + b_4 \log(Y_{ndom,t-3}) \tag{3}$$

where the $Y_{ndom,t}$ is the outdoor electricity consumption, $PRND_t$ is the electricity price for outdoor users, $X_{1,t}$ is the GDP, IND_t is a time trend, $b0$, $b1$, $b2$, $b3$, $b4$ are the coefficients of regression [17].

2.2.3 Fuzzy Linear Regression Methods

A function with fuzzy parameters is applied in various fields [5]. In a fuzzy linear function parameters are given by fuzzy sets and the prediction model might be introduced as a fuzzy system equation. Fuzziness should be taken into account in a case where human estimation influences the system [18]. The regression has two important questions which are required to be answered [19] :

1. What is the most appropriate mathematical model?
2. How to determine the best fitting model for the data?

A fuzzy linear regression model designed for load data of the previous three years as well, as the model coefficients, can be found after solving the mixed linear programming problem [18]. The MAPE can be equal to 2-3 % as it was during the prediction of the electricity consumption of households of the Bahia state, Brazil [20].

2.2.4 Additive Regression Prophet Model

The Prophet Model is a tool designed especially for time-series forecasting. This model is optimized for cases with seasonality and trend changes. It is built on the additive regression model with four components [21]:

- Curve trend
- Yearly seasonal component
- Weekly seasonal component
- List of deviations (for instance, holidays), which is given by the user

The Prophet Model has been built using its Python special library, it caughts seasonality pretty well. The disadvantage is that this model is specified for daily production only (not for hourly) [21]. This disadvantage conflicts with the objectives of the work (the hourly forecast is expected) so this method will not be used in this Thesis.

2.2.5 Random Forest

A Random Forest model is a method of generating and combining of Classification and Regression Trees (CART) with a subsequent aggregation [5, 22]. Every CART is made of a bootstrap sample which consists of a learning sample and a set of features which are chosen randomly at each node. Every tree in the forest is built on a set of learning points features considered at each node to split on.

Random Forest contains several parameters: number of trees (can be determined experimentally), number of input variables and depth of trees.

Trees are fitted separately with the bootstrap samples and then they are aggregating over the ensemble to obtain the final decision. This process is called "bagging" and it is necessary to improve the accuracy. The advantage of the Random Forest Classifier is that the fitting process could be performed in one sequence (with the cross-validation included) [22].

One of case study has been considered, related to the short-term forecasting of the electricity consumption. According to its results, Random Forest model can show better results than ARIMA or naive models, but they are not so accurate as the ANN based model [22].

2.3 Summary

Almost every method (expecting the Prophet model) from considered below is suitable for the forecasting of electricity production and consumption. Linear Regression will be chosen because of its simplicity and SARIMA because it is suitable for the initial data according to its seasonality. On plots of input data for prediction and consumption, different seasonal patterns are viewable so there could be a good decision to use SARIMA with different component values for the thesis work.

3 SELECTED METHODS USED FOR ELECTRICITY PRODUCTION AND CONSUMPTION FORECASTING

3.1 Autoregressive Integrated Moving Average

The term "integrated" in Autoregressive Integrated Moving Average means that it is necessary to ensure the time-series stationarity before the building of the model. The stationarity means that the time-series values vary around an unchanging mean, and the variance over time is constant, which is the requirement for the ARIMA model [12]. Examples of stationary and non-stationary time-series rows are presented in Fig. 1-3. Stationarity is important to ensure, that forecasted statistical values will be the same as values for historical data [23]. To check if the time-series is stationary or not, the Dickey-Fuller test is applied [24]. The term "Autoregressive" means that the forecasted values can be received from known data of the target value. "Moving Average" term means that the forecasted values are also predicted from the value of the error term [12].

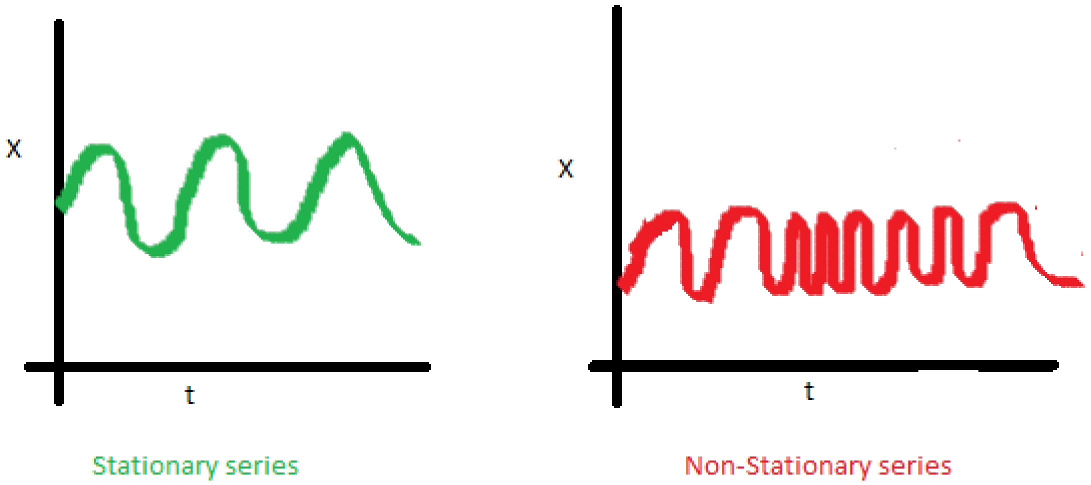

Figure 1. Non-stationary time-series when covariance is not constant [25]

ARIMA models are built by transforming the time series to its stationary form. The output contains a constant, the weighted sum of previous outputs and the weighted sum of error. The lags in terms of the stationary time series are referred to as "autoregressive", whereas the lags of the forecasted error terms are referred as "moving average". A time series which requires being differenced for the purpose of making it stationary is said to be an

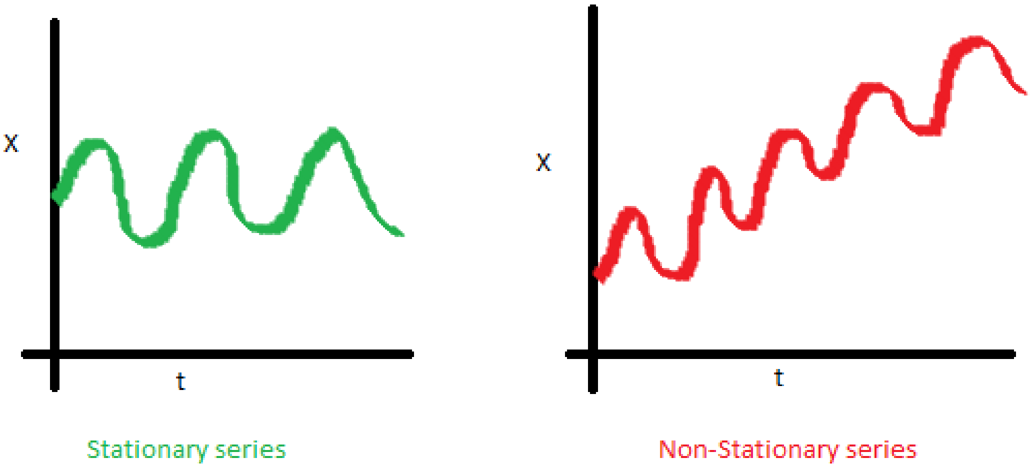

Figure 2. Non-stationary time-series when mean increases over time [25]

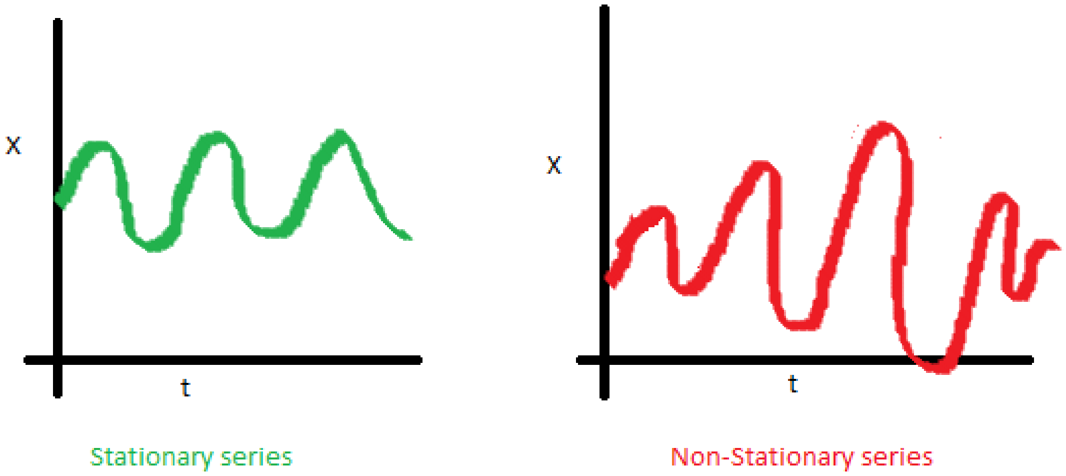

Figure 3. Non-stationary time-series with problems with dispersion [25]

"integrated" version of a stationary series. These models are denoted as:

$$ARIMA(p, d, q) \tag{4}$$

where p is showing the value of the Autoregressive part (AR), d denotes the degree of first differencing involved and q is for the Moving Average (MA) part [11]. In other words, the current value of the target value can be explained as a function of p past values, where p determines the number of steps back to the past needed to make a forecast about the current value. The MA of q assumes the white noise up to lags q are combined linearly to form the observed data [26].

To find the required parameters, two approaches can be applied. The first one is just to perform a model selection from a large number of models with different values of p, d, and q parameters and then choose a model with the minimal error or maximal Akaike Information Criterion (AIC). Obviously, this approach can take a lot of time so it can be unsuitable in practice. Another way is the visual inspection of the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) plots [26], which can be much quicker.

In SARIMA, the term "Seasonal" is related to the seasonality of initial data. SARIMA coefficients P, D, and Q reflect the same coefficients as p, d and q but for the seasonality [27]. The seasonal component indicates the seasonality length (for instance, for daily seasonality component length is equal to 24) and can be defined visually by observing the initial data plots.

A way to improve forecasting results relies on SARIMAX model. Additional features can be used as exogenous inputs, for instance temperature [12]. There have been a lot of case-studies [11–13, 15] based on the ARIMA models to forecast energy consumption and production. Some of them are based on pure ARIMA, some of them combine this approach with ANNs, SVMs, some of them use SARIMA or Multiplicative Autoregressive Integrated Moving Average (MSARIMA) [5, 20]. The results of the forecasting which have been achieved in these case studies are relatively good (MAPE no more than 9% for all of these cases) so it is worth to try to build the SARIMA-build model in this thesis work.

3.2 Linear Regression Method

Any X values can be taken as features, the question is how to find and choose them. Any time-series dataset consists of the date, time and the target value for the whole period of time. Date and time can be decomposed into various features, i.e., hour, the day of the week, the number of the week and others depending on the seasonality of the initial data. It is also possible to gather features from the train data by transforming the whole target variable dataset into lags vectors which contain information about the value for the previous hour [28]. The first, the last lag and their amount can be chosen according to the length of the initial data and its seasonality.

Different input factors can be important in different time periods [29]. It is important to identify factors with critical influence. The same factor (i.e., temperature or day type) for electricity consumption from different places can have a large impact or no have impact at all. To determine the necessary factors, the grey correlation analysis can be applied [29].

The Level of correlation is determined by analyzing the correlation between the target variable series and several influential factors. The less the two series differ, the more correlations they have [30]. Influential factors can be found with the grey association analysis. If the coefficient is high corresponding factors have a strong correlation. Also, the effect of a calendar day, such as day of the week, a month of the year, type of the day and type of the hour, [31] has proven to have an impact on a daily peak load [29]. It is also possible to choose suitable features based on their correlation. If some features correlate strongly between themselves then presumably these features contain the same information and there is no need to keep all of them [32].
Models built on the linear regression methods will be built with different combinations of features. Consumption and production forecasting results, achieved with models which do not use features related to the weather dataset are considered as baselines.

4 DATA PREPROCESSING AND MODEL PARAMETERS

4.1 Target Variable Dataset

4.1.1 Electricity Consumption

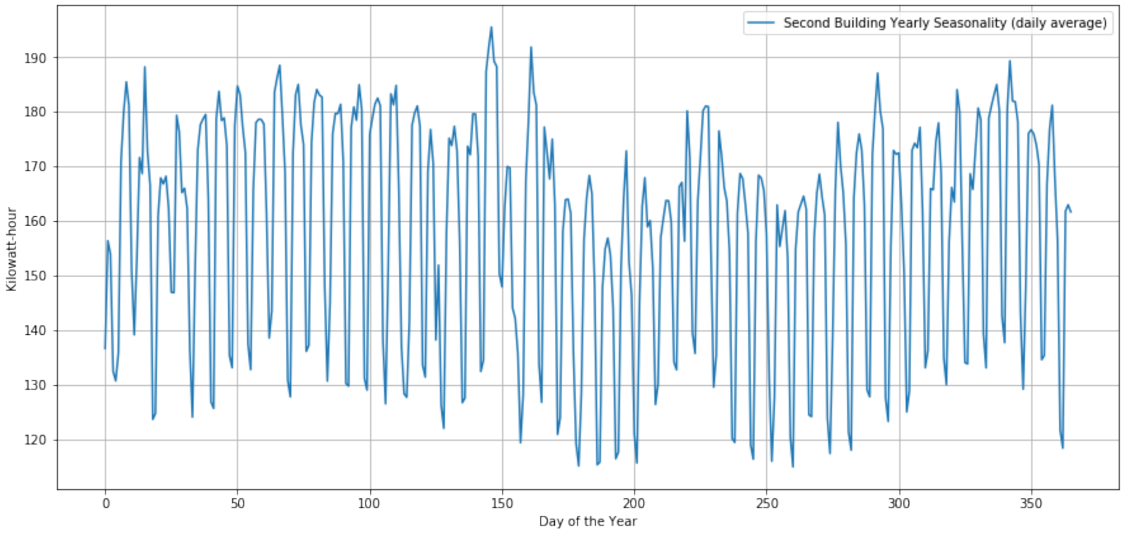

Figure 4. One year electricity consumption data for the 2nd building.

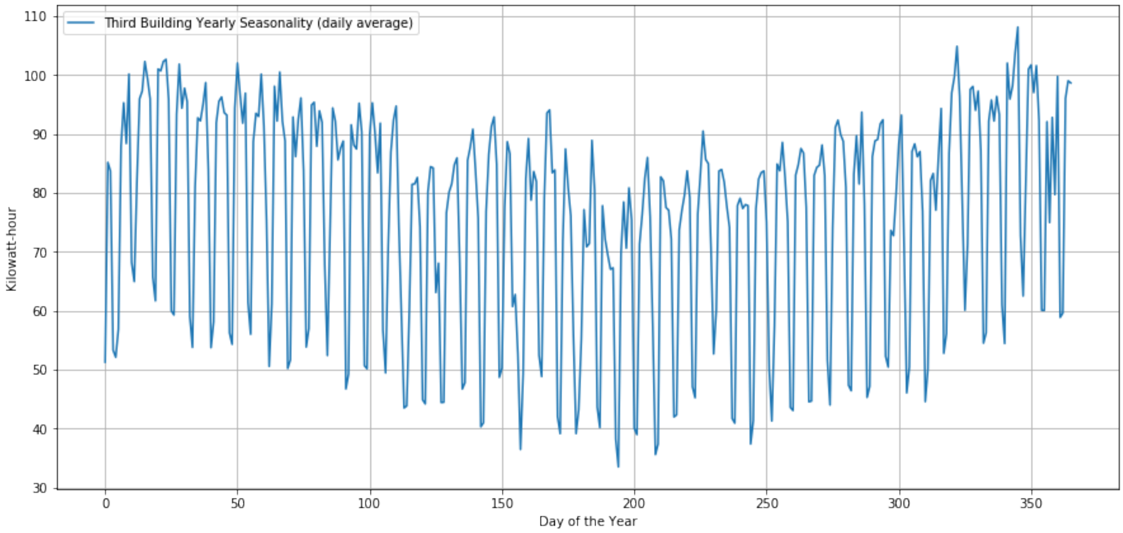

Figure 5. One year electricity consumption data for the 3rd building.

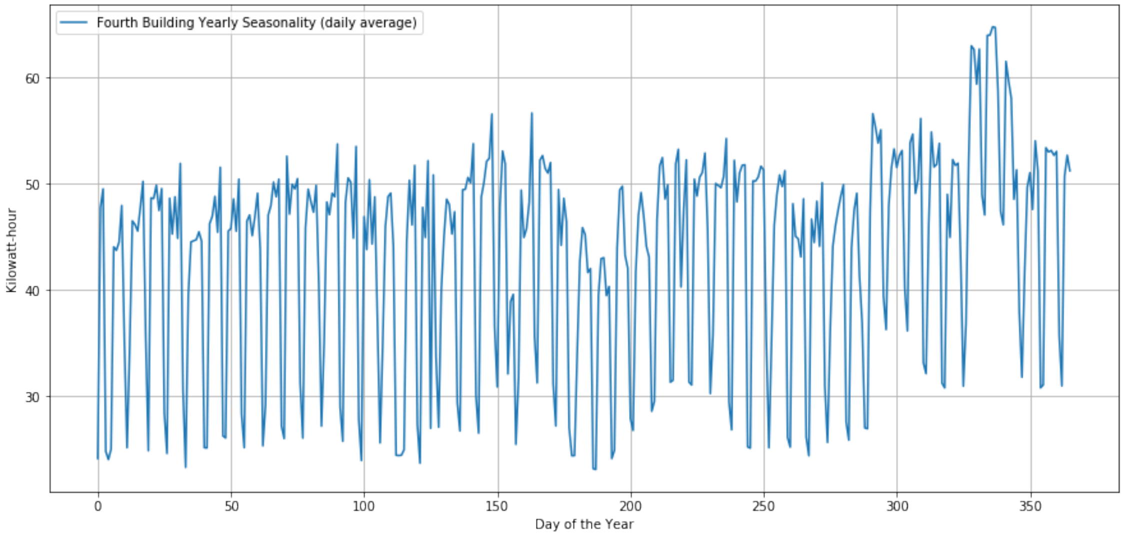

Figure 6. One year electricity consumption data for the 4th building.

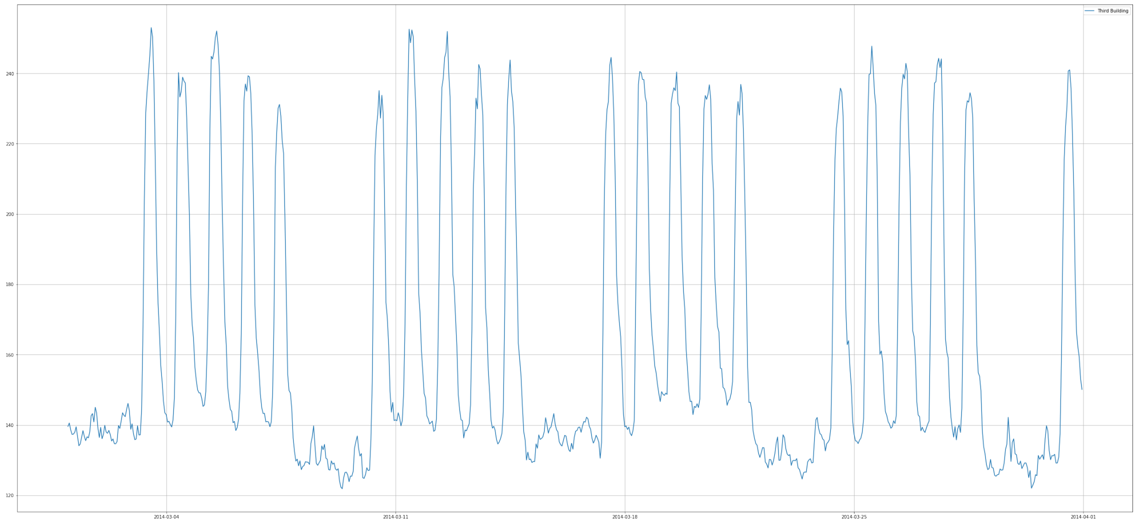

Figure 7. One Month Electricity Consumption Data for 4th Building.

Data provided by the LUT School of Energy Systems contains information about the electricity consumption and electricity production. 5 buildings of the University are considered. Dataset related to the consumption contains hourly information about 4 years (2014-2017) of electricity consumption in kWh and separately - of district heating energy consumption in MWh. As well as in example cases considered below, data have been measured each hour.

The following presumptions related to the seasonal patterns of the energy consumption data have been made:

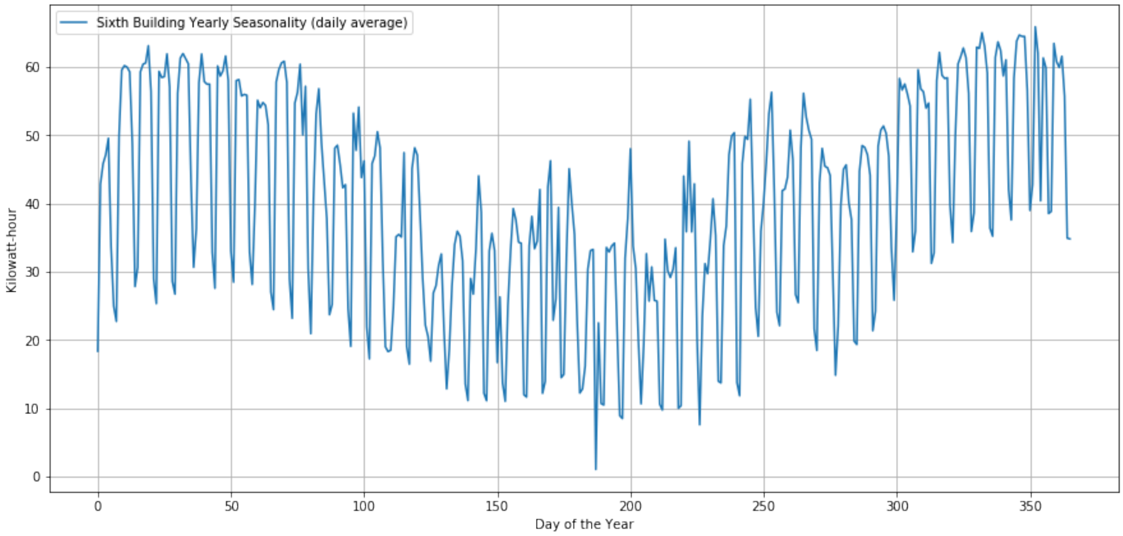

Figure 8. One year electricity consumption data for the 6th building.

- Although university buildings consume different amounts of electricity (for instance, the second building contains laboratories which consume a vast amount of power), they should have the same seasonality patterns.
- During the night-time, the university consumes less electricity compared to the daytime. There is no teaching in spite of the university is accessible day and night.
- During the weekend, the university consumes less electricity than during the working days because there is no teaching. The same can be said about national holidays and gaps between teaching periods.
- Seasons of the year do not influence electricity consumption because of the university staff is at work throughout the year.

All these assumptions can be confirmed (or denied) by using the data visualization. Yearly plots of the electricity consumption have been prepared for all buildings of the university (Fig. 4, 5, 6, 7, 8). Weekly averages have been calculated beforehand. The second and fourth buildings consume less energy during July. The sixth building shows a recession during spring. The seventh building shows a load peak in June and August. Third building consumes the same amount of electricity during the year. So it seems that the monthly seasonality does not have any common behaviour, presumably it is related to the aims of equipment which is placed in different buildings.

The monthly plot (March of the 2014th) is shown for the third building in Fig. 9. It indicates hourly and daily seasonality, which should be taken into account during forecasting.

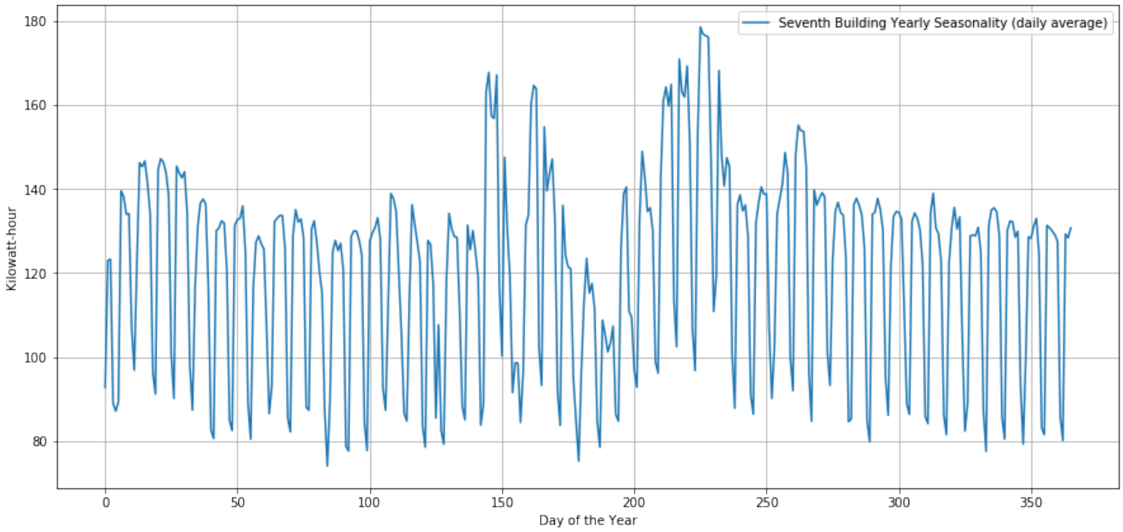

Figure 9. One year electricity consumption data for the 7th building.

4.1.2 Electricity Production

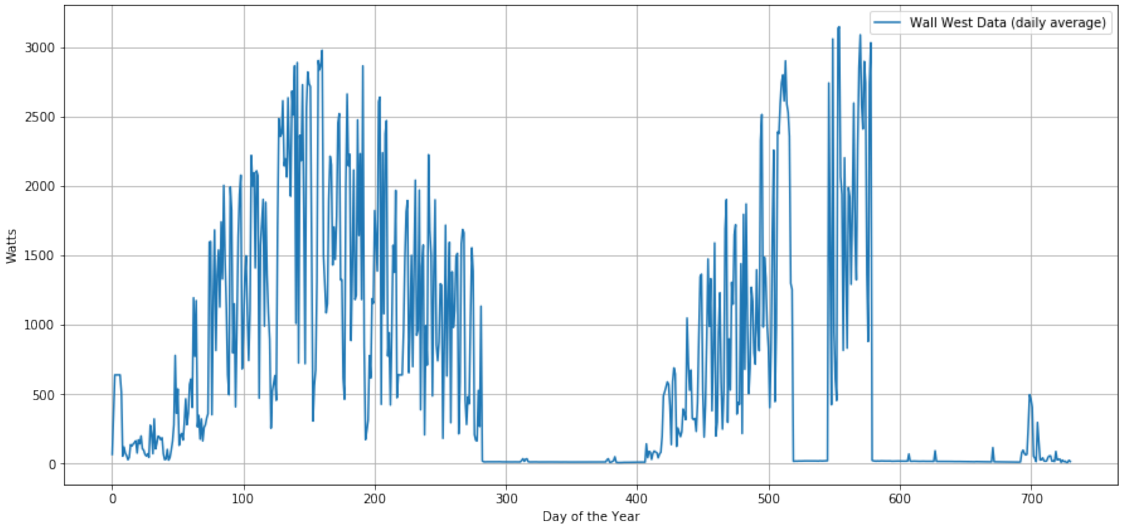

Figure 10. Production of the wall west PV group for 2 years.

Part of the data related to the production describes the last three (2015-2017) years. The data from year 2015 was not taken into account because it contains a lot of Not a Number (NaN) values (presumably there was no measurement). It contains information about the electricity gathered from the sets of PhotoVoltaic (PV) plants. Data has been merged by years and divided by panel groups - Fixed Installation (FI), Carport, SP, Tracker, Wall South and Wall West. Remained data also contains NAN values, which are appearing randomly. These NaN values have been replaced with the mean values of previous and

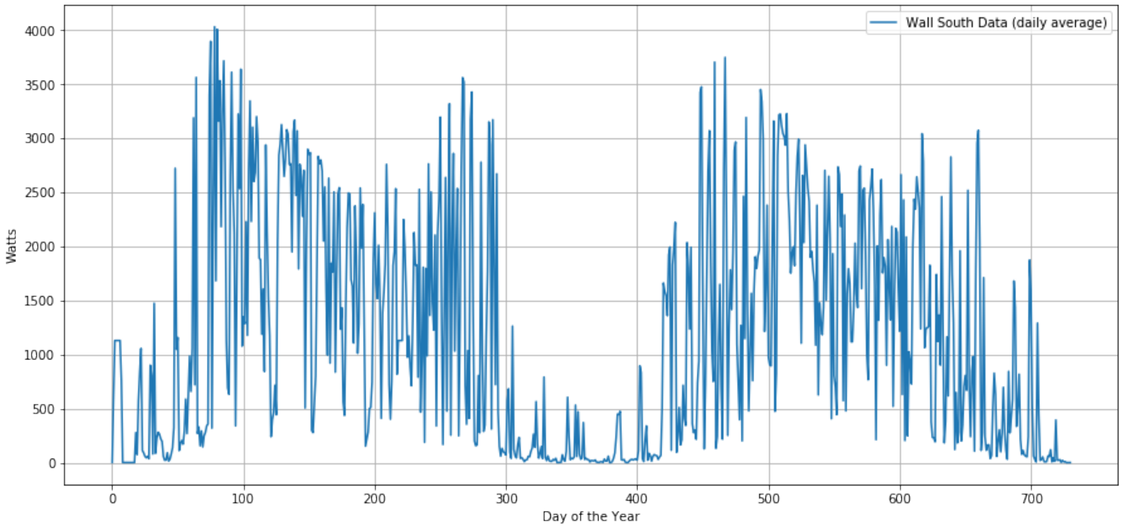

Figure 11. Production of the wall south PV group for 2 years.

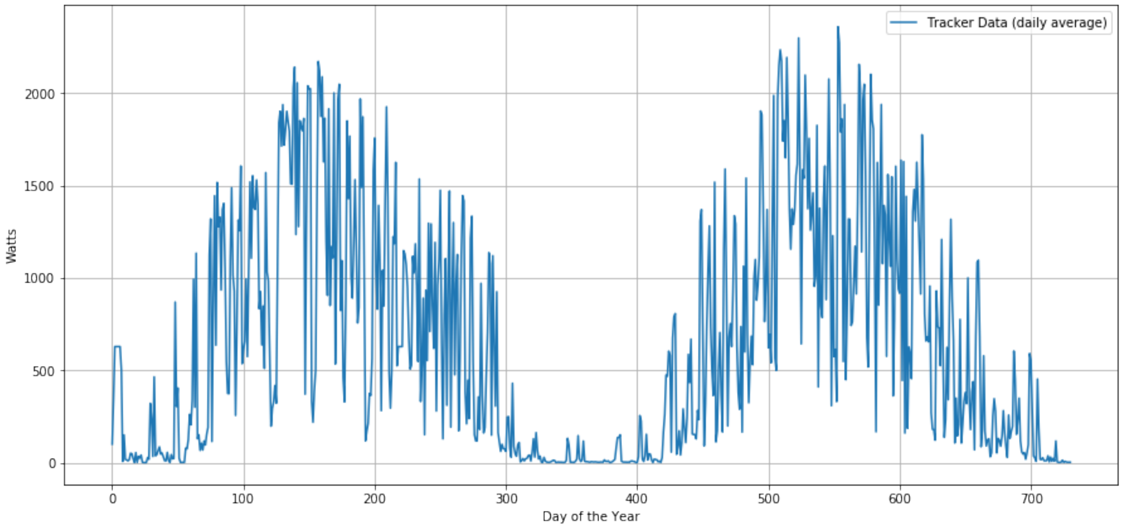

Figure 12. Production of the tracker PV group for 2 years.

next values.

For the experiments it will be necessary to study seasonalities. The following assumptions can be made:

- There is no seasonal trend related to the day of the week. The working schedule does not affect the electricity production.
- There is seasonal trend related to month and it is expressed for every solar panel in a similar way.

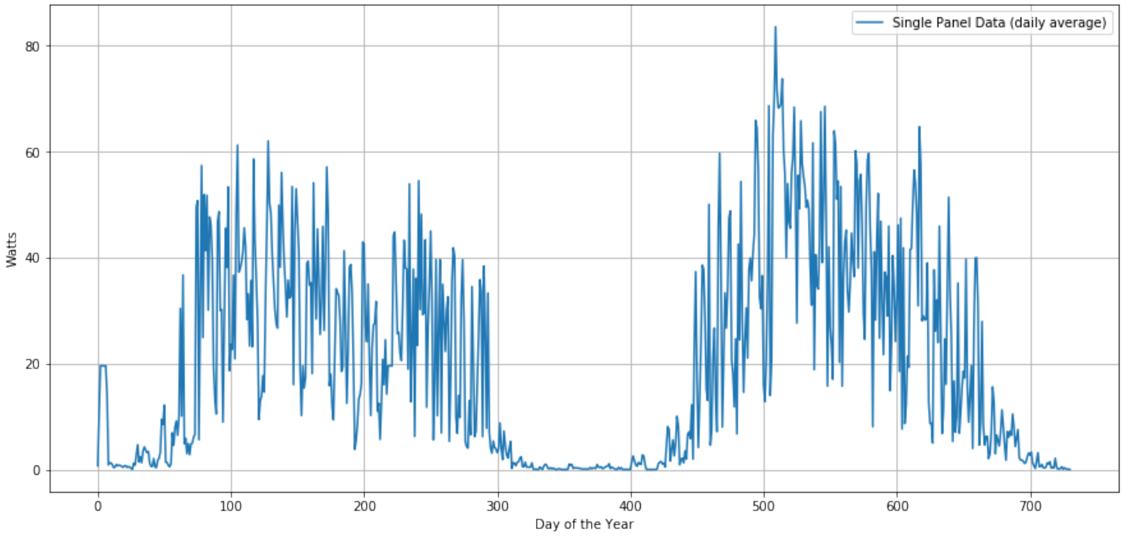

Figure 13. Production of the Single Panel (SP) group for 2 years.

- There is seasonal trend related to an hour (obviously sun shines less during night hours).
- There are dependencies between the weather data related to the solar information.

Plots related to production in different time periods are shown (see Fig. 10, 11, 12, 13, 14, 15). The following conclusions can be made:

- Seasonal trend related to an hour does not always exist because of the different day length.
- There is no seasonal trend related to the day of the week but month seasonality is present.
- There is a large period of missing data in June related to the Wall West data (see Fig. 12). Probably this fact will influence the make the forecasting a bit worse.

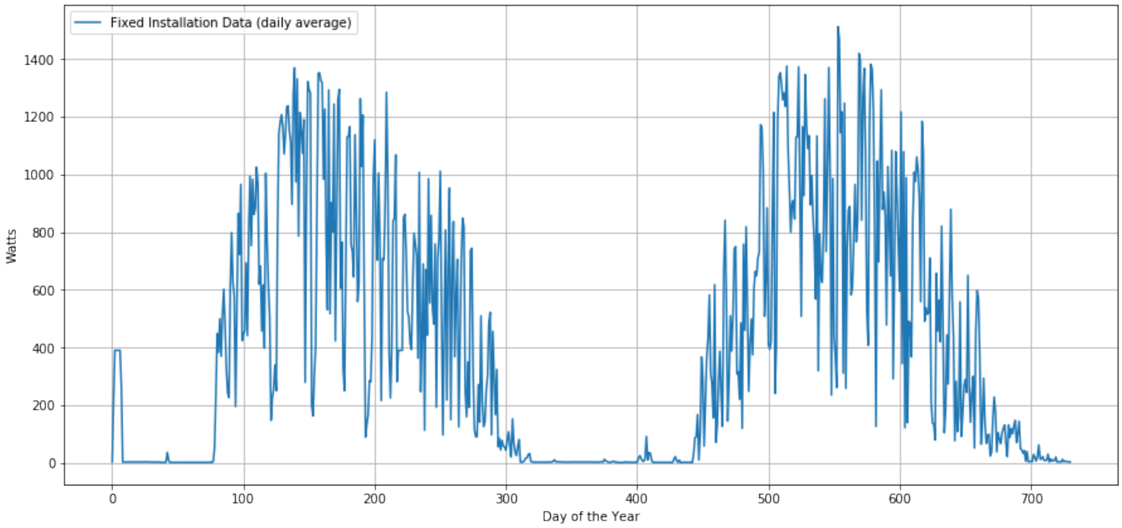

Figure 14. Production of the FI PV group for 2 years.

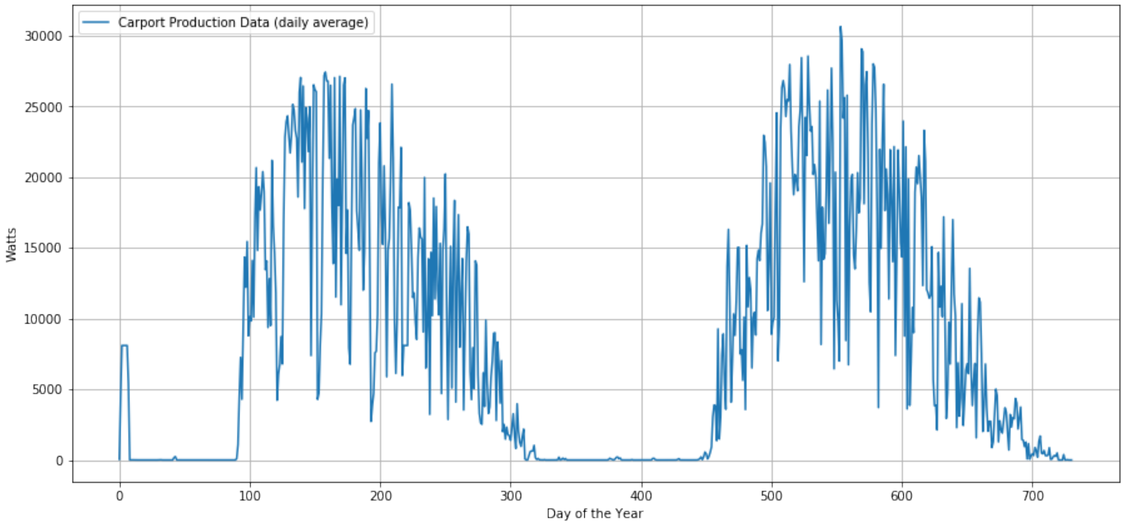

Figure 15. Production of the carport PV group for 2 years.

4.2 Outlier Detection

During the data preparation, it is necessary to process the outliers. An outlier is a data point which has a large distance from the vast majority of other observations mostly they appear because of different errors, for instance, a temporary hardware failure or a human mistake [33].

There are various ways for the outlier detection, for instance, by viewing a plot or a boxplot diagram but unfortunately, it is not suitable for the considered case because the dataset is too large. Computational methods include the extreme value analysis [34], iterative methods (during every iteration the group of suspicious objects gets deleted), local outlier factor (the idea is that outliers usually do not have a lot of neighbors in contradistinction to non-outliers), and the quartile method [35].

Matlab built-in function "isoutlier" offers the following ways for the outlier detection [36]:

- Median: elements which are more than three scaled Median Absolute Deviation (MAD) from the median are considered as outliers.
- Mean: elements which are more than three standard deviations from the mean are considered as outliers.
- Quartiles: elements more than 1.5 interquartile range above the upper quartile or below the lower quartile. Useful in case of the non-normal distribution.
- Grubbs: applies Grubbs test for outliers, which iteratively removes one outlier per iteration based on hypothesis testing. Useful in case of normal distribution.
- Gesd: applies the generalized extreme Studentized deviate test for outliers.

As seen from the list, one of the steps to select the outlier detection method is to check if the distribution is normal. For instance, the Anderson-Darling method can be applied [37]. By using this method, it seems, that the energy production and consumption data are not normally distributed. So the quartile method can be used for the outliers detection. Results of the outlier percentage for the electricity consumption dataset are presented in Table 1.

Table 1. Percentage of outliers for the electricity consumption data.

Type	2nd	3rd	4th	6th	7th
Electricity	1.87%	4.6%	4.2%	0.65%	0.29%
Heating	1.21%	1.89%	1.8%	1.3%	2%

And the electricity production percentage of outliers is presented in the Table 2.

The energy production dataset contains a lot of values which are considered as outliers. Probably some of them are just real rare values. The visual data analysis of Fig. 12-15 helps to understand that the majority of the, for instance, zero values reflect real zero values related to the darkness. Several possible solutions can be proposed:

Table 2. Outliers measurement for the electricity production data.

Panel Set	Carport	FI	Flatroof	Tracker	Wall South	Wall West
Percentage	16.9%	17%	16.3%	16%	13.6%	13.5%

- Select the k value. If k (or more) values in a row are equal to zero, do consider them as real values.
- Take the date and time features (season, hour, month, weather) into account. If zero values have been registered during a dark hour or cloudy season, do not consider them as outliers.
- Check data about all production panels. If every of them produces nearly the same amount of electricity, do not consider production values as outliers.

4.3 Outlier Processing

The next step after the outlier detection is the outlier processing. It can be done with the linear interpolation obased on the neighbourings [38]. After the outlier replacement, it is worth to check the amount of outliers again. It was made for the electricity consumption (see Table 3) and production (see Table 4):

Table 3. Outliers measurement for the electricity consumption data after the outlier processing.

Type	2nd	3rd	4th	6th	7th
Electricity	0%	0.2%	0.02%	0%	0%
Heating	0%	0%	0%	0%	0%

Table 4. Outliers measurement for the electricity production data after the outlier processing.

Panel Set	Carport	FI	Flatroof	Tracker	Wall South	Wall West
Percentage	2.8%	7.4%	0.8%	7.3%	3.5%	8.2%

4.4 Weather Data

Data about the weather and solar radiation for last 4 years (from 2014 until 2017) can be gathered from the Finnish Meteorological Institute (FMI) website. Lepola weather station and Airport weather station have been used to get the most complete information because they are the closest to LUT. The datasets contain the following features related to weather and solar values. Another preprocessing step is the search of the NaN values. NAN values percentage for each feature has been calculated for two weather stations (see Table 5).

Table 5. Amount of NaN values for the weather features.

Station	Lepola	Airport
Temperature	41.6%	0.00047%
Wind Speed	45%	0.35%
Wind Gusts	45%	0.34%
Wind Direction	45%	0.35%
Humidity	41.6%	0.006%
Dew Point	41.6%	0.006%
Rain	94%	84%%
Rain10min	42%	0.26%
Snow	53%	30.6%
Pressure Sea Level	52.7%	0.006%
Visibility	41%	1.16%
Cloud Amount	41%	0.007%

According to the Table 5, Snow and Rain consist of NaN values for more than 40%, so these values can be removed because of the uselessness. Another conclusion is that it is better to take the Airport station because a huge (for 20 months) part of the data from the Lepola station is absent from the beginning.

The correlation matrix has been plotted for the features from the FMI station and the heatmap has been plotted for visualization (see Fig. 16). Correlation heatmap helps to make conclusions about the dependency and independency between the features.

According to the heatmap (see Fig. 16), it seems that Visibility and Wind Direction are the unique values which do not influence other values and Temperature, Wind Speed, Wind Gusts, Humidity, Dew point, Rain10min strongly correlate between themselves. Probably it is not necessary to include all of them to the model.

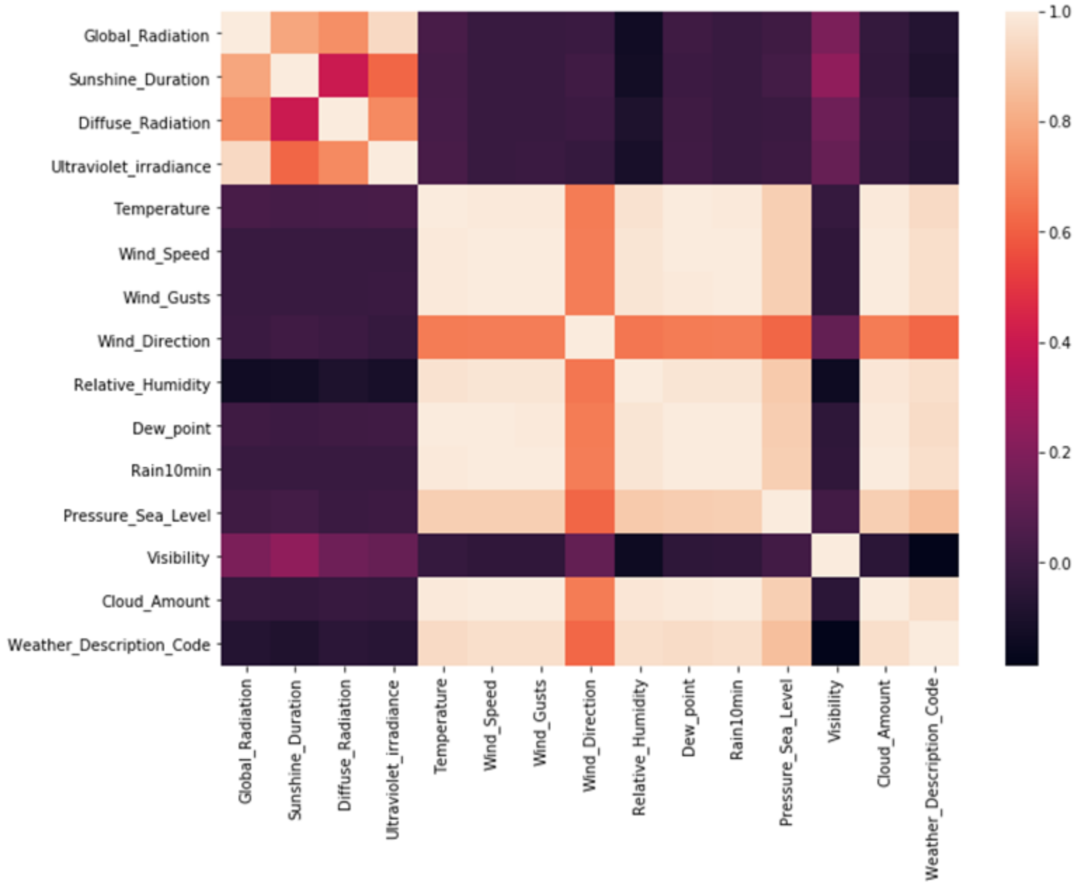

Figure 16. Heatmap of the correlation of the weather features.

Similar correlation matrice heatmaps with the target variables, related to consumption and production also have been built (Fig. 17-20).

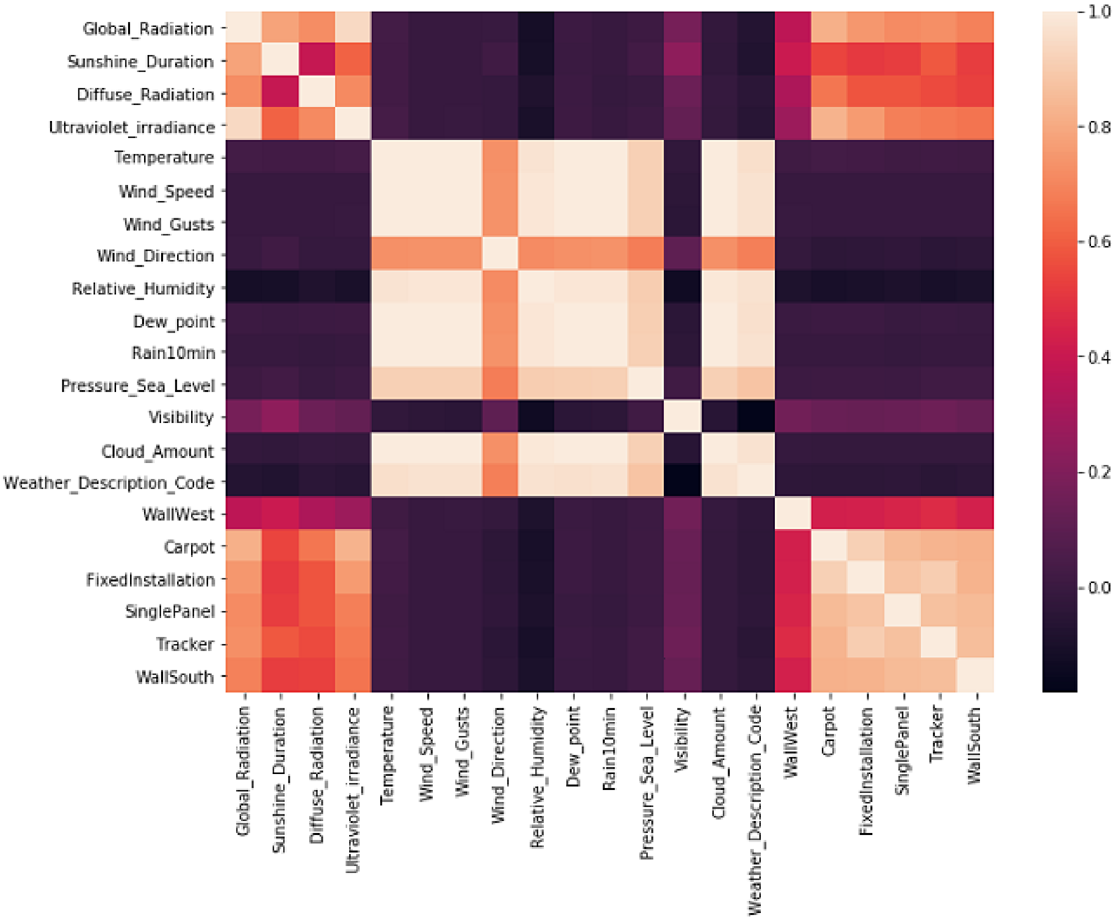

Figure 17. Correlation heatmap between the production and values from the weather stations.

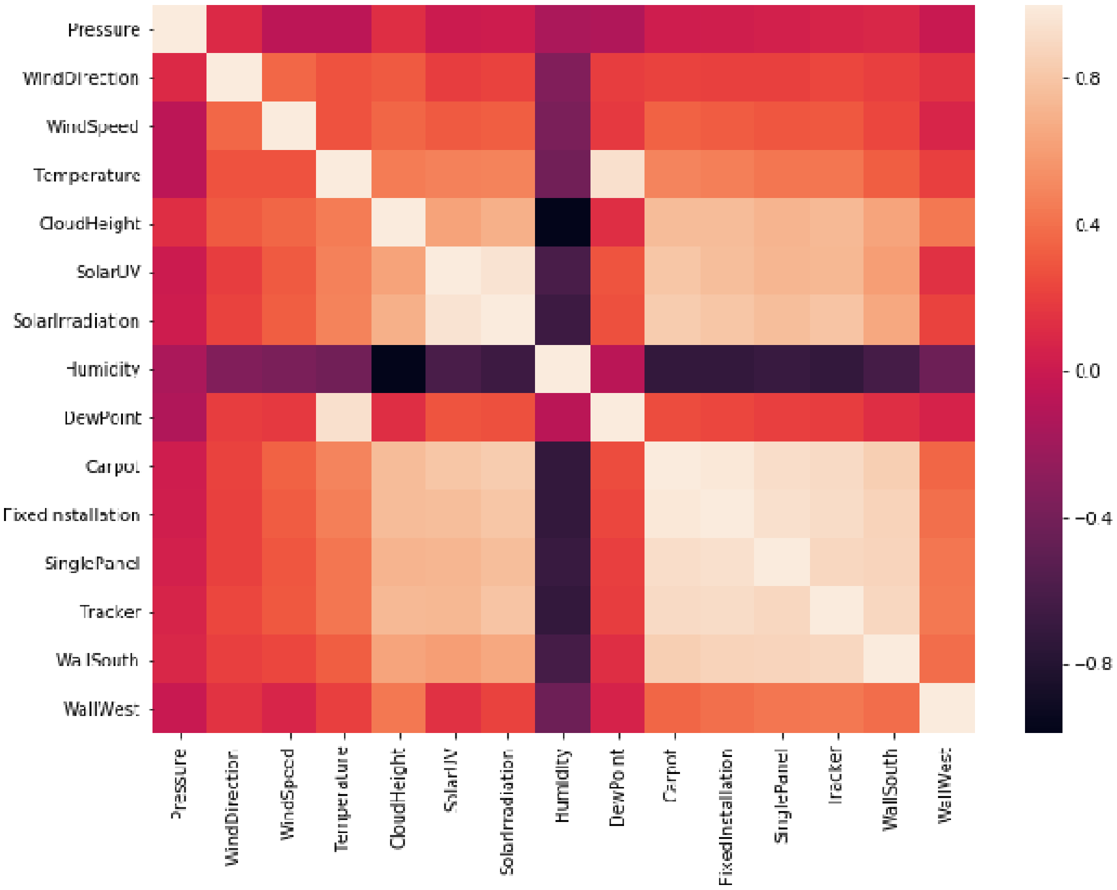

Figure 18. Correlation heatmap between the production and values from the weather stations.

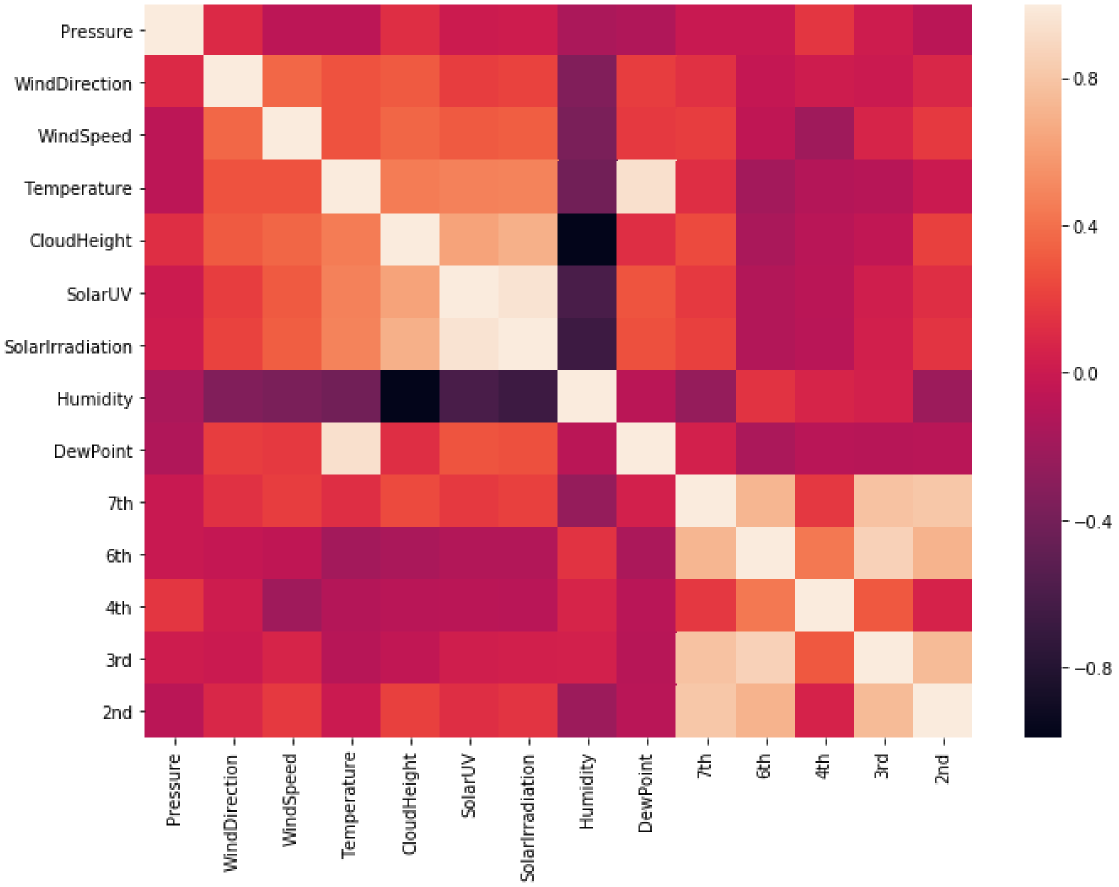

Figure 19. Correlation matrix heatmap between the consumption and values from the local weather and FMI stations.

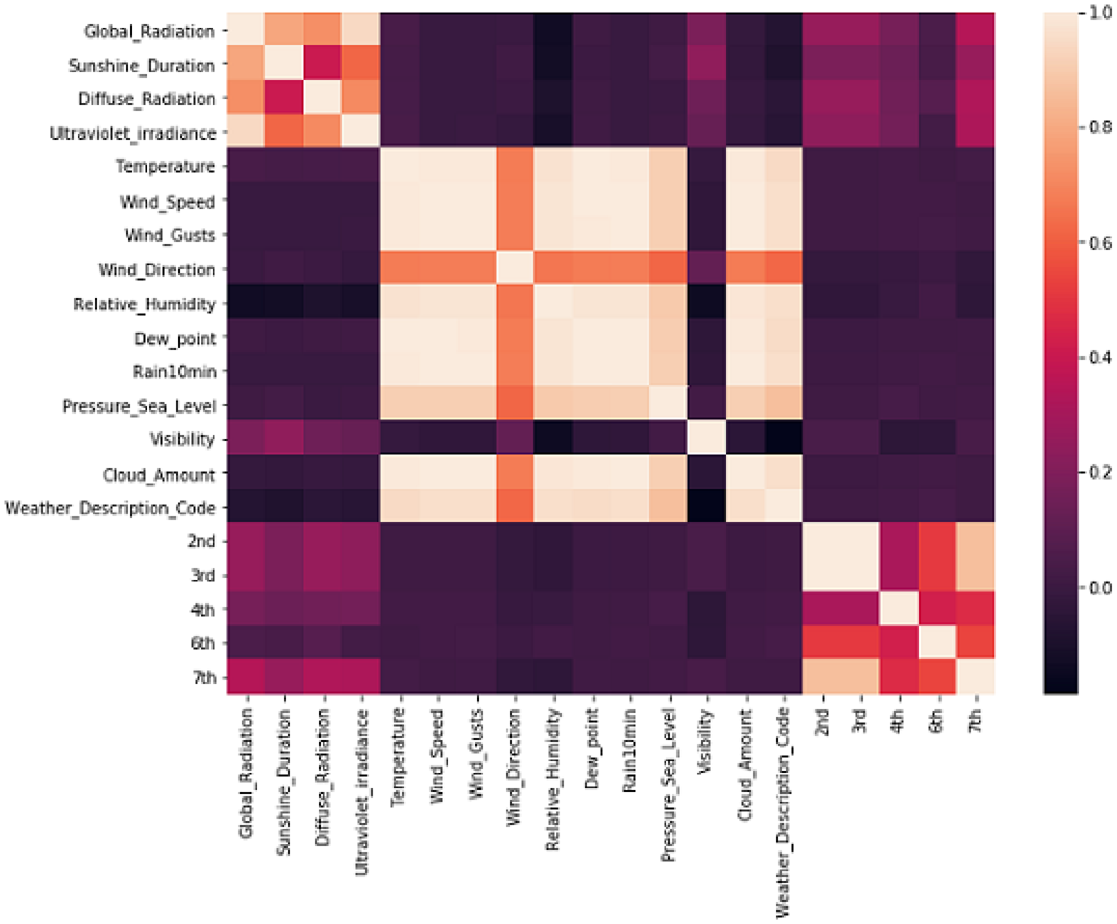

Figure 20. Correlation matrix heatmap between the consumption and values from the local weather and FMI stations.

Their results are quite logical and expected: it shows that solar data and weather does not influence to electricity consumption at all. It also shows that for the electricity production ultraviolet irradiance and solar irradiation are the most correlated for each panel group excepted the Wall West.

Initial data requires preprocessing because it has different time resolution: from 3 to 6 times per hour. The mean values have been taken to calculate feature values per one hour. Another way to get the information related to the weather and solar data is to use the local weather station which is located at the LUT. It is located on the lawn, works wirelessly and is served by solar panels of the university [39]. Assumingly, weather features values make more precise results for the electricity production forecast.

Access to the local weather station values was gained by using the Grafana system (Fig. 21). Grafana keeps and shows information related to consumption, production, and weather in real time as well as archived. The advantage of Grafana is the ability to choose necessary time step and fill the NaN values with various methods (with zeros, previous values or by linear method). The only disadvantage is that the service provides recorded weather information values about only for the 1.5 years.

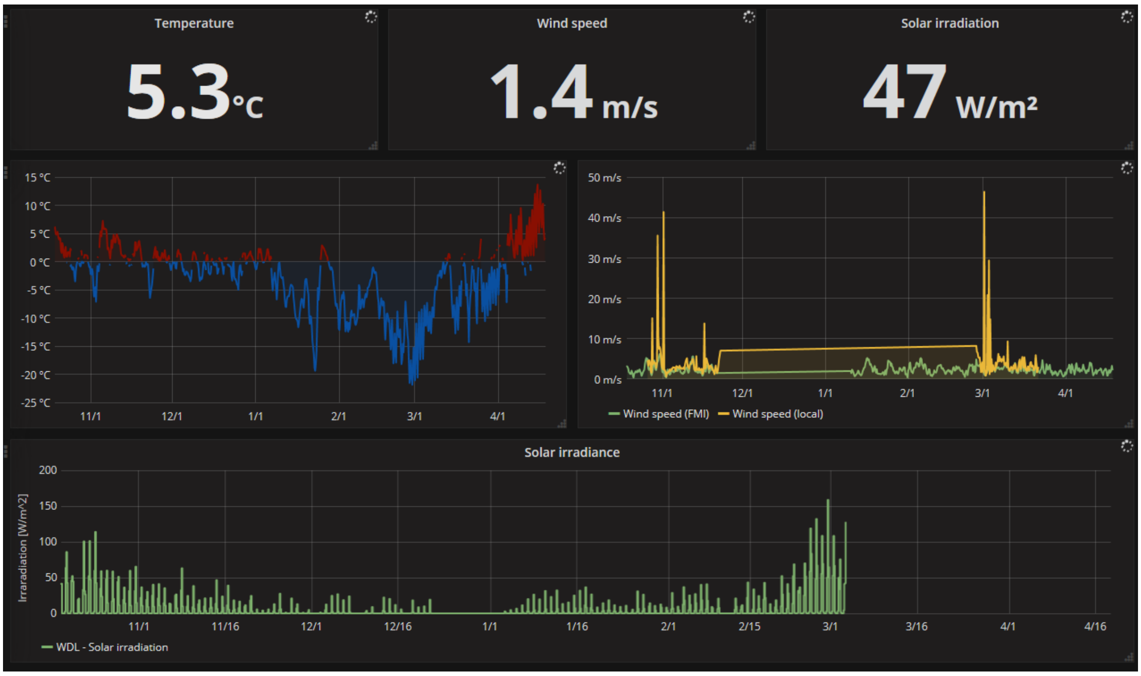

Figure 21. Grafana Dashboard.

Correlation matrice heatmap for the features from the local weather station also has been plotted (see Fig. 22). The following conclusions are made:

- Solar UV and Solar Irradiation strongly correlate amongst themselves.
- Humidity value has a negative correlation with almost each value.
- The majority of parameters do not have a high correlation amongst each other.

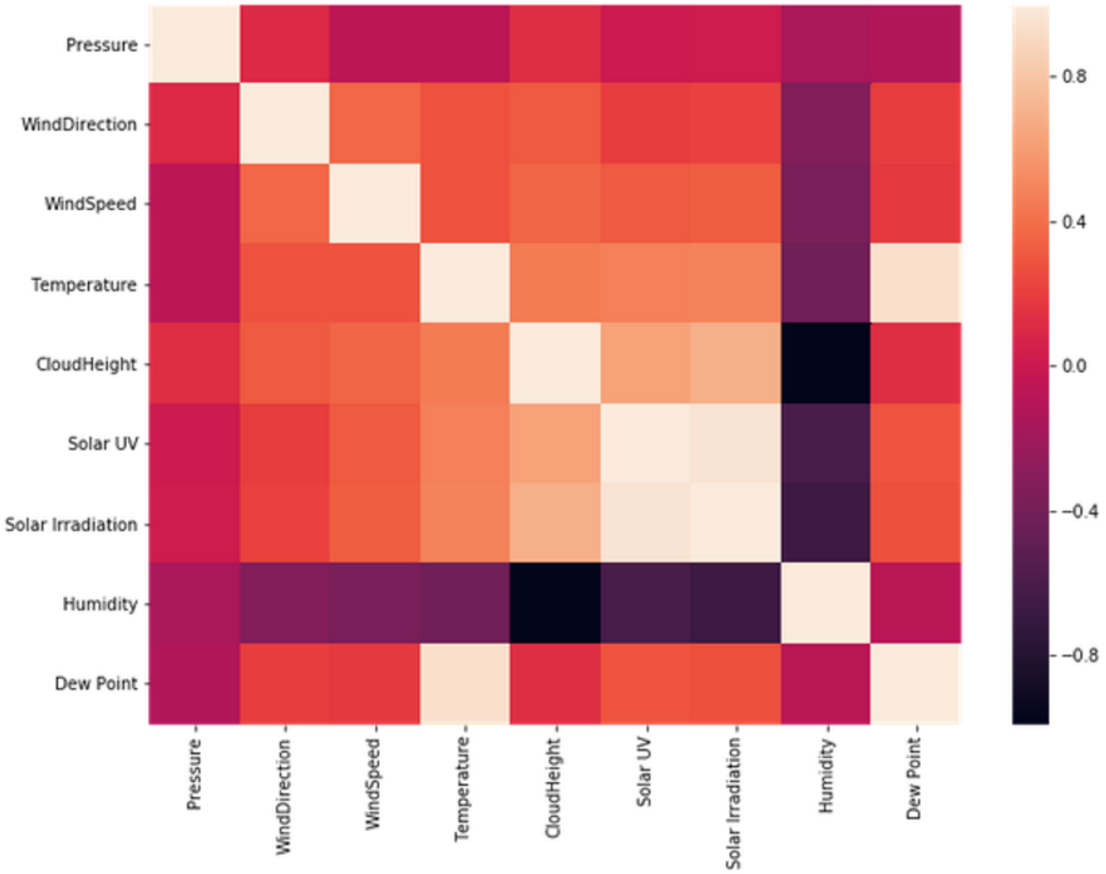

Figure 22. Local weather station features correlation heatmap.

Grafana also makes possible to access data from FMI (unfortunately from the Lepola station only) and provides great visualization abilities. For instance, it is allowed to compare the same features from different stations (Fig. 23-25). Comparing of different sources says that for the common, i.e., temperature or pressure, characteristic values are exactly equal. But for the non-common characteristics, for instance, wind speed values are different. It can be explained by particular qualities of the landscape, for instance, difference between the wind speed values presumably takes place because the local weather station is surrounded by forest tracks stronger.

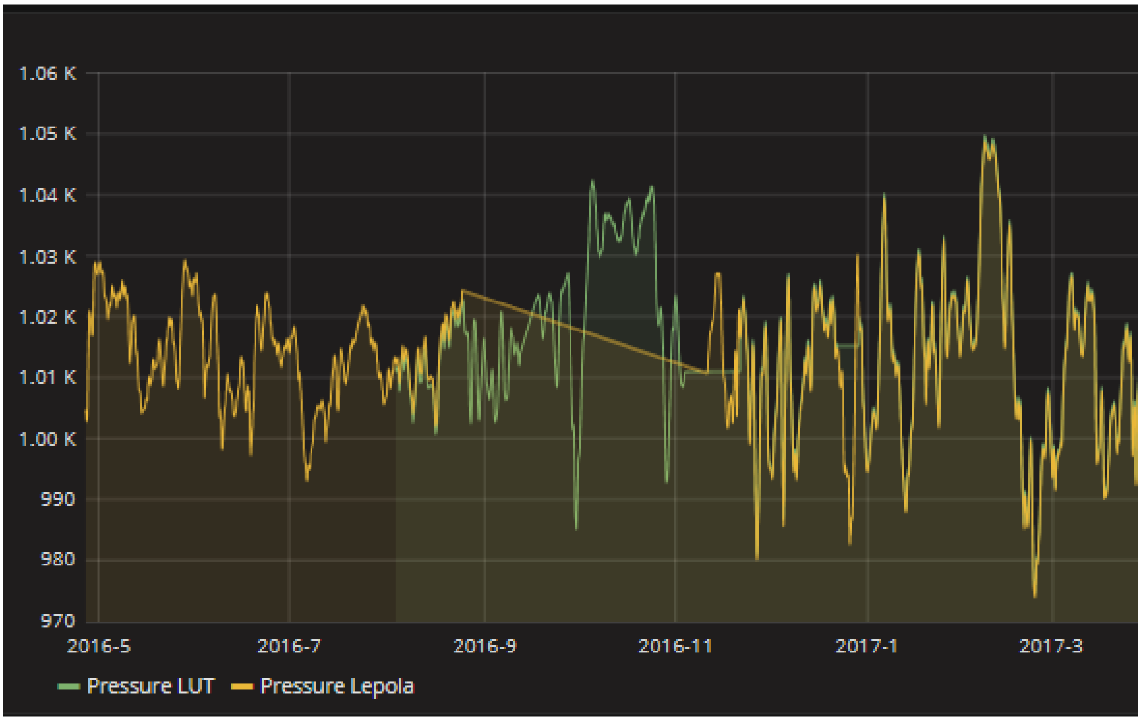

Figure 23. Pressure values from the local weather station and Lepola.

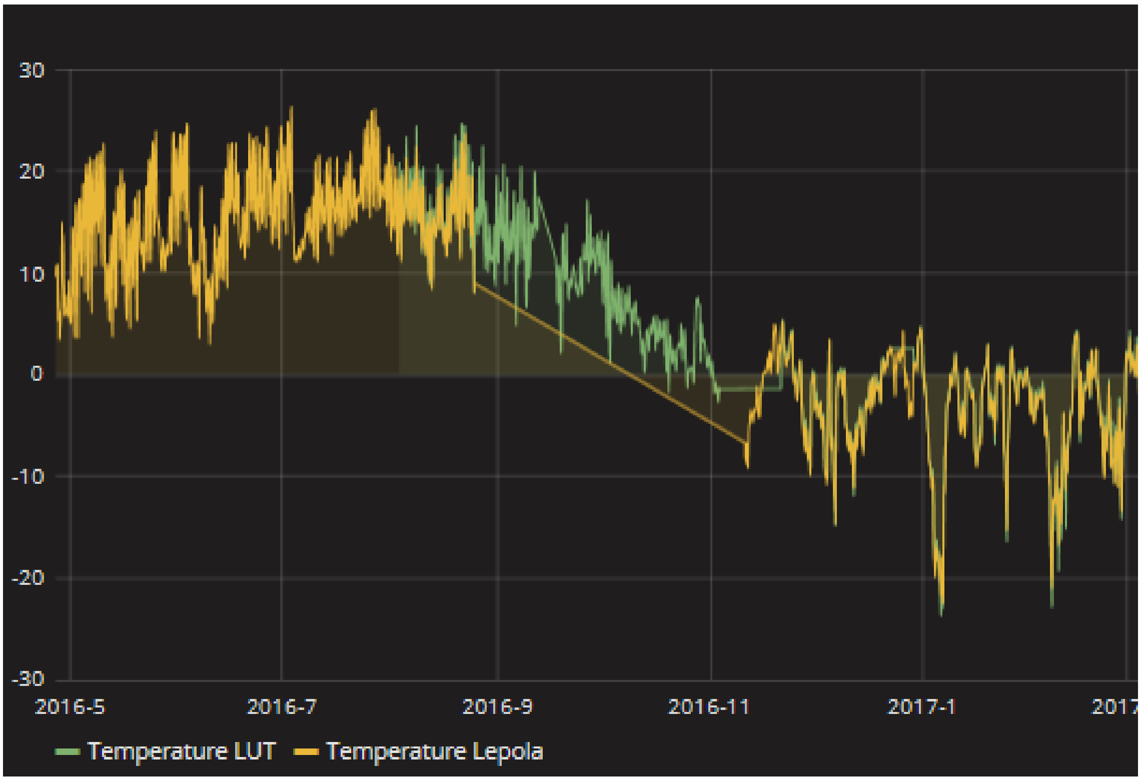

Figure 24. Temperature values from the local weather station and Lepola.

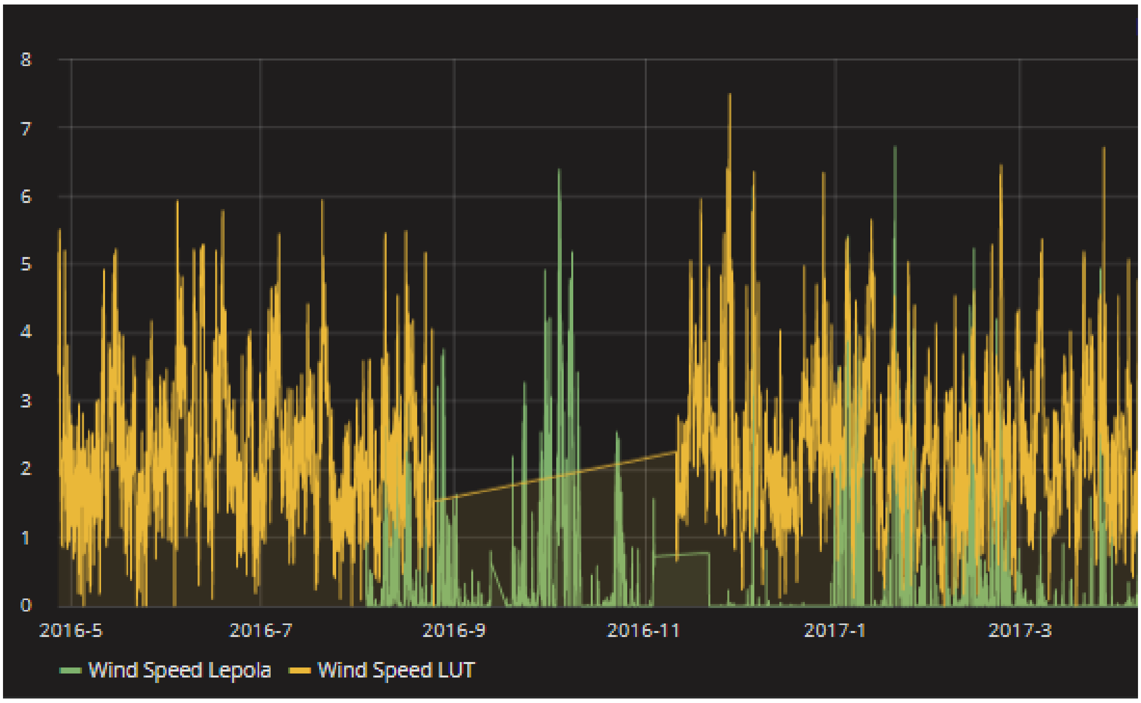

Figure 25. Wind Speed Values from the local Station and Lepola.

4.5 Forecasting Length

The suitable length of historical data needs to be estimated, it can be done with, for instance, PACF analysis [29]. It is considered that the hourly values of parameters related to the weather capture the most conservative conditions which can occur in this hour. It is a common approach [8, 40] which is considered in many articles. Forecasting can be divided into short-term (up to 24 hours), medium-term (up to one week) and long-term (up to one year data) forecasting [41]. 49% of investigated case studies consider the forecast for one hour. In 19% forecasting goes for 1 day [3].
The length of prediction can be 48 hours ahead of information issuing every 12 hours [3]. The majority of the conventional ANN-based load forecasting methods work with 24-hour-ahead load forecasting with the forecasted temperature. The drawback of this method is that when rapid changes in temperature of the forecasted day occur, the power of the load changes significantly, which leads to the higher forecast error [42]. In this work, the short-term forecasting is used to achieve better accuracy.

4.6 Measurement of the Accuracy of the Prediction

4.6.1 Cross-Validation

Cross-Validation is a method which applies to measure accuracy. Firstly the initial dataset is randomly split into K folds of the target variable, then the model is trained and tested K times using different fold as the testing subset [43]. Cross-Validation is an important step [44] for a wide range of the machine learning tasks to estimate the prediction accuracy.

Unfortunately, the cross-validation in its pure form is not suitable for the time-series forecasting [45], because during the forecasting data from the future will be caught. The reason is ignoring of the temporal components and assumption that there is no relationship between the observation [45].

So the modification of the cross-validation intended for the time-series forecasting can be named "cross-validation on a rolling basis". The training set consists of a set of observations which occur before the observation which contains the test set. Efficiency is calculated by averaging the results of all sets [46]. The approximate algorithm is presented in Fig. 26, where the blue dots represent the training set and the red dots represent the test sets. So the training set becomes larger and larger, the Mean Average Error (MAE) is calculated as the mean value of errors for each fold.

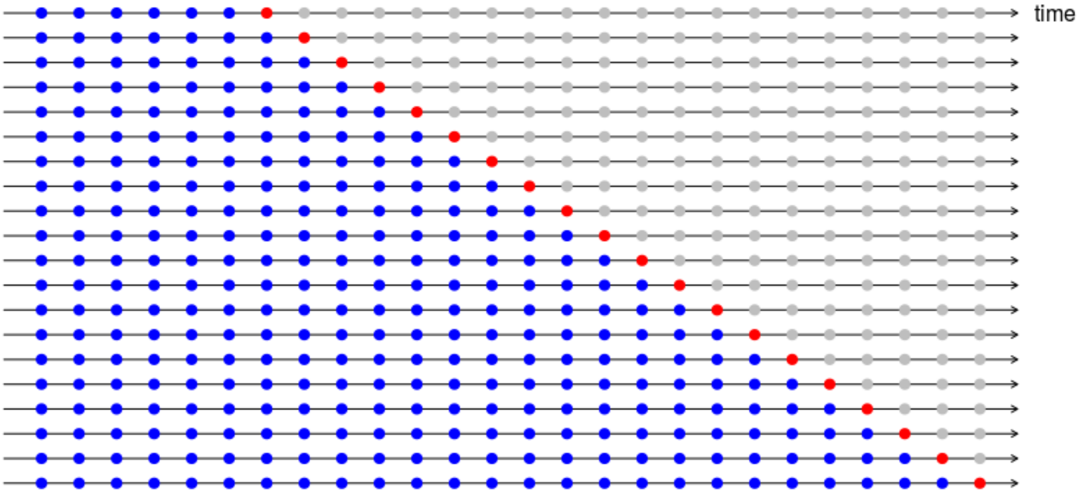

Figure 26. Example of the evaluation on a rolling forecasting origin (red dots represent the test sets and blue dots represent the training set).

4.6.2 Measurement Scores

Evaluation of results is always an important step. To measure the efficiency of forecasting, there are several ways. It is worth to mention that the method of measuring does not influence the method of forecasting, the same measuring method can be applied to the neural network model as well as to the model based on, for instance, linear regression [8]. On the other hand, the best way of choosing the measurement way depends on specific conditions [47]. In practical case-studies, the following characteristics have been used as measurement scores :

- MAE [48–50].
- MAPE [49, 50].
- Root Mean Square Error (RMSE) [41, 51].
- Maximum Absolute Error (MaxAE) [8].
- R^2 (R-squared) value [48] .

- Coefficient of Variation (CV) [41,51].
- Least Squares Approach [52].

MAPE is being recommended for the common use [53]. But it has some disadvantages. Firstly data should be suitable for the considered case of energy balance. Secondly MAPE puts heavier penalty on forecasts which are more than the actual values [47], because according to the formula the division by the actual value occurs. One more disadvantage of MAPE is that it is not included to the scikit-learn library, which will be used during the model building. Another problem is related to the formula of MAPE.

$$MAPE = \frac{1}{n}(\sum_{i=1}^{n})\frac{|Y_i - \hat{Y}_i|}{Y_i} \tag{5}$$

If the real value Y_i is equal to zero then this formula is unusable. The way to avoid it is to add some value (for instance 1) to all forecasted and real values. So MAE, MAPE and R^2 value will be used. Typical R^2 value is between 0 and 1, but it can also be negative. For the absolutely equal values R_2 is equal to 1.

5 EXPERIMENTS AND RESULTS

5.1 Linear Regression Model

Designing of the model, learning, fitting and evaluating of the results can be considered for the electricity consumption and production separately because of the differences related to seasonality and influential features.

5.1.1 Electricity Consumption Modeling and Results

For the first baseline data the following features (gathered from the date and time information) have been taken into account:

- Time of the day (hour)
- Week of the year
- Number of the weekday

Training data is divided into the time lags, which are using as features. 168 lags (one week) have been taken according to the seasonality, Moving window has been realized (one fold per every month). Results can be considered as a baseline which should be improve, for instance, by using weather or other features. Results can be obtained in Table 6.

The obvious way to improve the forecasting result for the model based on a linear regression is to add useful features. The features have been taken from the weather datasets provided by FMI and Local Weather Stations. The forecast has been built for 1-year data. Presumably and empirically it was discovered that the only helpful feature for this case is the cloud amount, adding of this feature reduces the MAPE for each University building. MAPE has been reduced (see Table 6). Example of the consumption prediction is shown in a plot (see Fig. 27).

Table 6. Electricity Consumption Forecast Improved Result.

Building	2nd	3rd	4th	6th	7th
MAE, no weather	4.379 kWh	4.581 kWh	0.33 kWh	4.04 kWh	4.239 kWh
MAPE, no weather	2.399%	5.835%	8.4%	10.15%	3.113%
R^2, no weather	0.97566	0.9394	-2	0.893	0.96
MAE, weather	4.373 kWh	4.5784 kWh	0.33 kWh	4.01 kWh	4.238 kWh
MAPE, weather	2.395%	5.83%	8.17%	10.14%	3.11%
R^2, weather	0.97	0.9389	0.966	0.94	0.937

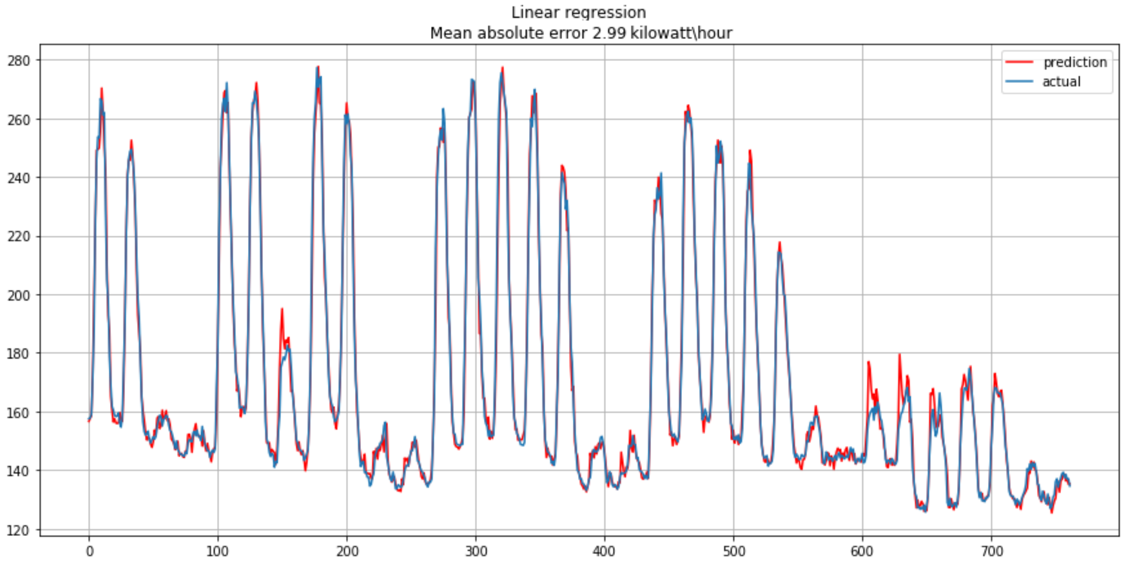

Figure 27. Electricity consumption forecast for the 2nd building.

5.1.2 Electricity Production Modeling and Results

The following features (gathered from the date and time information) have been taken into account:

- Time of the day (hour)
- Week of the year

Unfortunately, the local weather station provides information only about 18 past months. As was discovered before, production data has a yearly seasonality, so it is better to take 1 year. Results can be obtained in Table 7. The plot of the production is shown for better understanding (see Fig. 28).

Table 7. Electricity production forecast based on 1 year data.

Panel	Carport	FI	SP	Tracker	Wall West	Wall South
MAE	2.2 kWh	0.1 kWh	0.007 kWh	0.2 kWh	0.353 kWh	5.3 kWh
R_2	0.9	0.88	0.4	0.85	0.78	0.83
MAPE	131828%	3954%	91.49%	6346%	6176%	6942%

The problem of MAPE is that for instance, if forecast equal to 0.011 Kilowatt-hour and real value equal to 0.002 Kilowatt-hour, Percentage Error is 450% in spite of this error is small. It causes to fabulous prediction forecasting values up to 500000% and makes MAPE inconvenient. So the R_2 is also included.

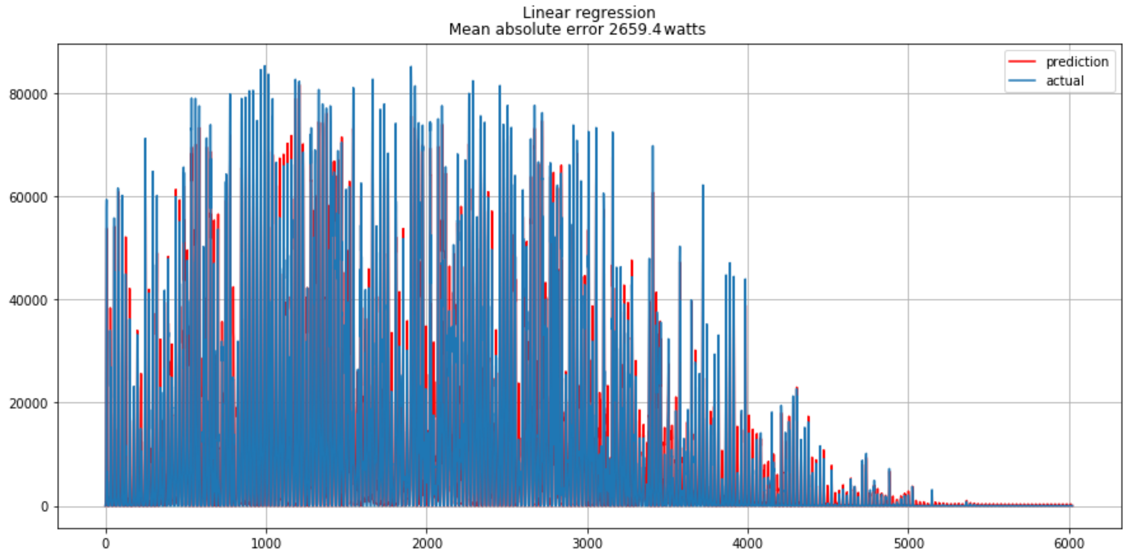

Figure 28. Electricity production forecast for the carport plot.

Another step is to get useful features empirically. The following conclusions have been made:

- Including the Temperature, Dew Point, Rain10min, Pressure Sea Level, Visibility, Cloud Amount, Wind Speed, Sunshine Duration, Humidity features spoils the prediction result for all solar panels. So these features will not be included at all.

- Adding any features to the Wall West solar reduces the prediction result accuracy. Maybe this panel does not have any dependencies.

- Including the Pressure and Diffuse Radiation features reduces the prediction result accuracy or does not change it.

- Including the Wind Direction (from the Local Weather Station), UV Irradiance (from the Airport) improves the forecasting result for all weather stations as well as the combination of this features.
- Including the Cloud Height (Local), Global Radiation, Relative Humidity improves forecasting result for Carport and Tracker. There should be a logical connection.
- Including the Solar Irradiation (Local) vastly improves forecasting result for Carport, FI and Tracker.

A lot of combinations have been tested to find optimal combination of features. According to the Table 8, it seems that in spite of common patterns and improvements each solar panel has its unique best set of features.

Table 8. Electricity consumption results based on different feature sets.

Feature Set	Carport	FI	SP	Tracker	Wall West	Wall South
UV + Wind Direction	3416.3	183.78W	8.9 W	277.42 W	376.4 W	458.8 W
UV + Wind Direction + Cloud Height + Solar Irradiation	3118 W	176 W	9.2 W	257 W	428 W	496 W
UV + Wind Direction + Solar Irradiation + Global Radiation	3125 W	176.9 W	9 W	260 W	363 W	482 W
UV + Wind Direction + Cloud Height + Solar Irradiation + Global Radiation + Relative Humidity	3135 W	179.3 W	9.16 W	254.28 W	381.6 W	492.4 W
Cloud Height	3408 W	186.1 W	9.2 W	271 W	337.75 W	487 W
Cloud Height + Solar Irradiation	3126 W	175.3 W	9.4 W	263.3 W	343.8 W	515 W
Solar Irradiation	3115 W	176 W	9.34 W	265.5 W	289.4 W	506 W
Global Radiation	3431 W	186.2 W	9 W	277.15 W	381.2 W	462.9 W
Solar Irradiation + Global Radiation	3132 W	177.8 W	9.36 W	266.78 W	393.6 W	501.7 W
Humidity + Solar Irradiation	3121 W	175.2 W	9.32 W	262.5 W	308 W	507 W
UV + Wind Direction + Solar Irradiation	3115 W	177.8 W	9.17 W	261.7 W	410.1 W	486.2 W
Diffuse Radiation	3447 W	185.8 W	8.9 W	278.8 W	243.3 W	463.1 W
Diffuse Radiation + UV + Wind Direction	3394 W	182.2 W	8.87 W	274.85 W	290 W	515.6 W

To sum up, there should be the following feature adding (see Table 9). Using the same sets of features for other models is worth trying in the following work.

Table 9. Features influence to the prediction forecast.

Panel	Best Weather Features	MAE	MAPE	R^2
Carport	UV + Wind Direction + Solar Irradiation	3.115 kWh	36797 %	0.88
FI	Humidity + Solar Irradiation	0.175 kWh	1054%	0.85
SP	UV + Wind Direction + Diffuse Radiation	0.008 kWh	212.3%	0.84
Tracker	UV + Wind Direction + Cloud Height + Solar Irradiation + Global Radiation + Relative Humidity	0.254 kWh	2581%	0.82
Wall West	-	0.221 kWh	496%	0.65
Wall South	UV + Wind Direction	0.458 kWh	1324%	0.81

5.2 Seasonal Autoregressive Integrated Moving Average

To build a model based on the ARIMA, it is necessary to commit the following steps:

1. Preprocess the data
2. Estimate the parameters (p, d, q, P, D, Q).
3. Perform tests to measure the goodness of fit.
4. Perform a forecasting.

As was mentioned before, the autoregressive component consists of p, q and d components. To fit a data to the model, it is necessary to determine values related to these components.

5.2.1 Consumption Forecasting Modeling

The Dickey-Fuller test was made to estimate the stationarity of time-series. Augmented Dickey-Fuller value is close to 0, so time-series data is stationary, so there is no need to implement additional operations and d and D parameters are equal to 0. Otherwise these parameters would be equal to the amount of operations which was made to turn the time-series into the stationary view.

There can be 2 ways of parameters determination, as was said before. Unfortunately, building of several models with each combination of parameters requires a lot of resources, so the remained possible way is to find parameters with PACF and ACF plots.

So PACF and ACF for each separate building have been drawn. For instance, for the 2nd building, plots are displayed in Fig. 29.

To get the p and P values, it is necessary to estimate the number of lags on the PACF plot which strongly differs from 0. There are 3 values which are $> |0.5|$ so it seems that p and P values are equal to 3. To get the q and Q values, it is necessary to find the lag from the ACF plot which has the largest difference with the next lag. According to the Fig. 30, values are equal to 2. The last parameter is seasonality and it is equal to 24. In a similar way, parameters for each building to forecast the consumption have been found, they are presented on Table 10.

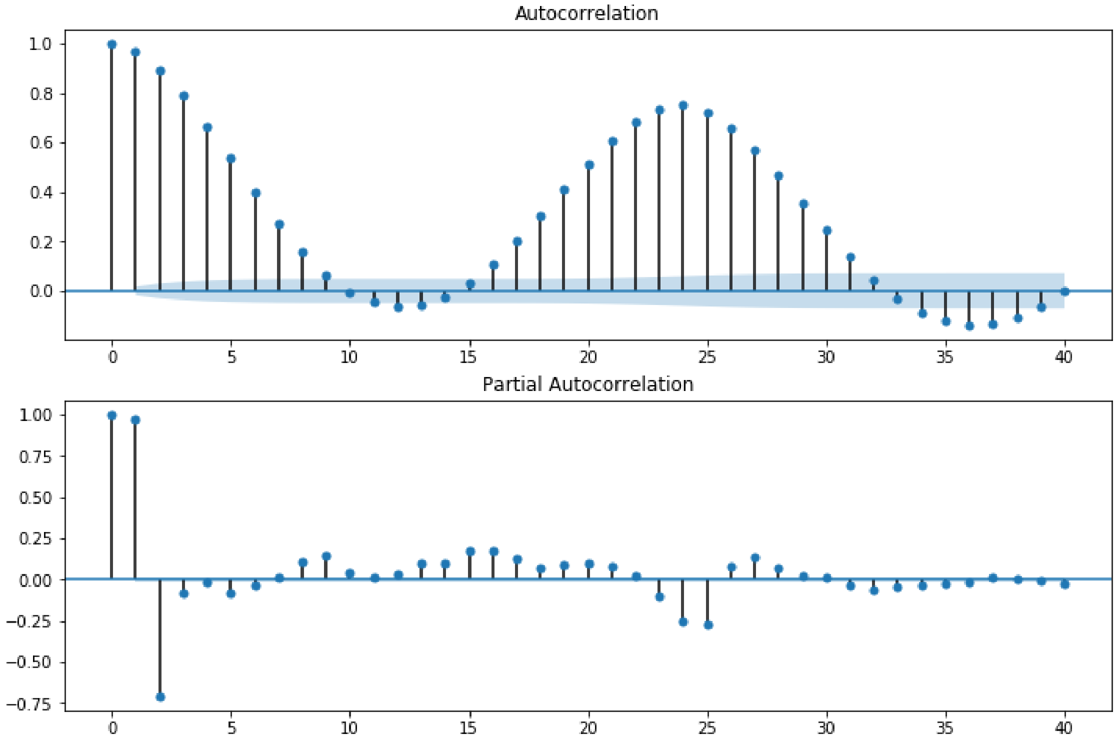

Figure 29. PACF and ACF plot for the 2nd building consumption data.

Table 10. Parameters of SARIMA for the electricity consumption prediction.

Building	Parameters (p, d, q); (P, D, Q)
2nd	(3, 0, 2); (3, 0, 2, 24)
3rd	(2, 0, 1); (2, 0, 1, 24)
4th	(2, 0, 1); (2, 0, 1, 24)
6th	(4, 0, 2); (4, 0, 2, 24)
7th	(3, 0, 2); (3, 0, 2, 24)

According to the parameters, models for each building have been built, results are received (see Table 11). Prediction accuracy is approximately the same as the accuracy of the model built on the linear regression. Plot related to the consumption of the seventh building is shown in Fig. 30.

Table 11. Forecast result for the consumption.

Building	2nd	3rd	4th	6th	7th
MAE	4.19 kWh	4.9714 kWh	2.392 kWh	4.282 kWh	4.9 kWh
MAPE	2.91%	7.65%	6.59%	16.5%	5.2%
R^2	0.97	0.939	0.966	0.94	0.937

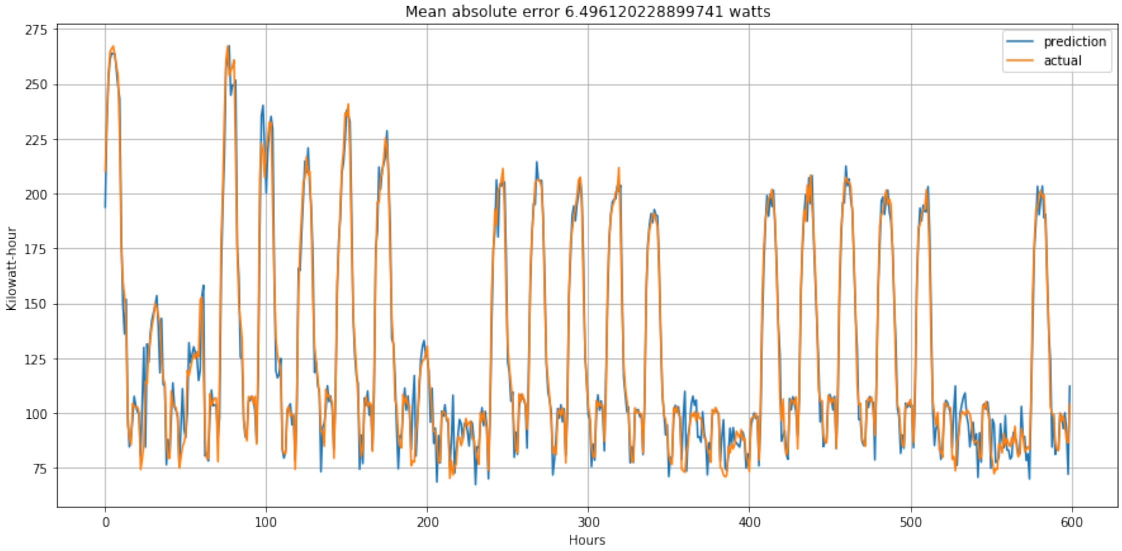

Figure 30. Plot of the electricity consumption of the 7th building.

5.2.2 Production Forecasting Modeling

Parameters for the 1 year data have been found in the similar way as before (Table 12), model has been built for 1 year production, results received (see Table 13).

Table 12. Parameters of SARIMA for the electricity production forecasting.

Building	Parameters (p, d, q); (P, D, Q)
WallWest	(2, 0, 2); (2, 0, 2, 24)
WallSouth	(3, 0, 3); (3, 0, 3, 24)
Tracker	(2, 0, 2); (2, 0, 2, 24)
SP	(3, 0, 3); (3, 0, 3, 24)
FI	(2, 0, 3); (2, 0, 3, 24)
Carport	(4, 0, 2); (4, 0, 2)

The same has been done for 1 year to build the SARIMAX model. One more problem is that it is allowed only to choose one feature, so Solar Irradiation for Carport, Tracker and FI, Diffuse Radiation for SP and Global Radiation for Wall South has been chosen (see Table 13).

For the carport, SP and wall south panel sets SARIMAX model shows the highest accuracy than SARIMA. For the FI, tracker and wall west panel sets SARIMAX shows a bit

Table 13. The forecasting result for the production.

Panel	Carport	FI	SP	Tracker	Wall West	Wall South
MAE (no ftrs)	1.85 kWh	0.1 kWh	0.005 kWh	0.015 kWh	0.05 kWh	0.32 kWh
MAPE (no ftrs)	7295%	306.1%	59%	461%	142.7%	628%
R^2 (no ftrs)	0.91	0.88	0.88	0.87	0.86	0.83
MAE (ftrs)	1.87 kWh	0.1 kWh	0.004 kWh	0.161 kWh	-	0.3 kWh
MAPE (ftrs)	6226%	362.9%	58%	537%	-	586%
R^2 (ftrs)	0.9	0.89	0.88	0.87	-	0.84

lower accuracy than SARIMA. But for all the panel sets anyway SARIMA model shows better results for the production than the linear regression based model (see Table 8). Plots related to the Carport electricity production are shown in Fig. 31 and Fig. 32 for intuition understanding. One of them shows the production prediction during August (see Fig. 31) and the second one - during November (see Fig. 32).

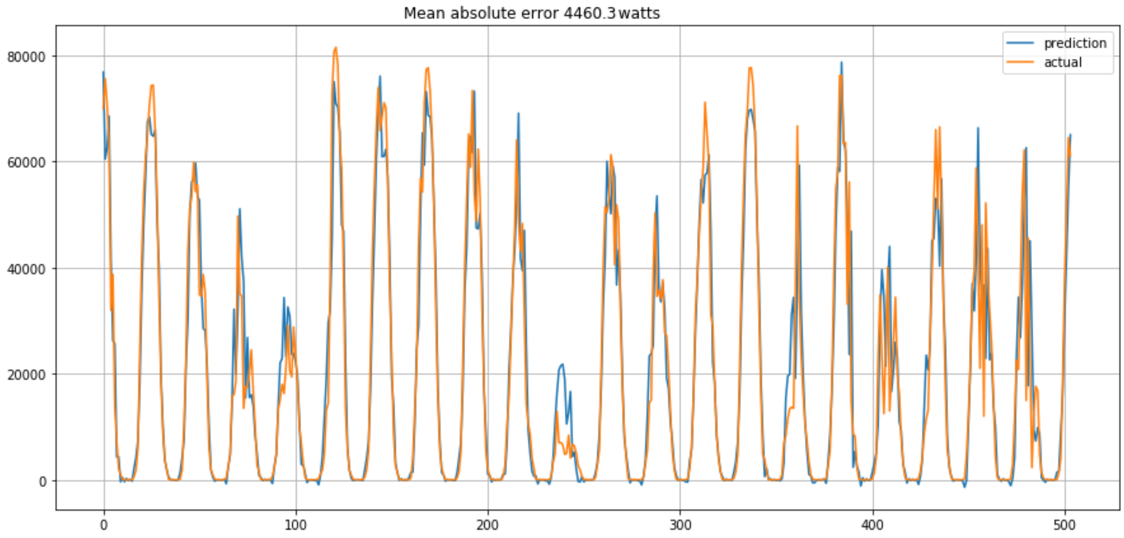

Figure 31. Plot of the electricity production forecasting based on the SARIMAX model in summer time by the Carport panel.

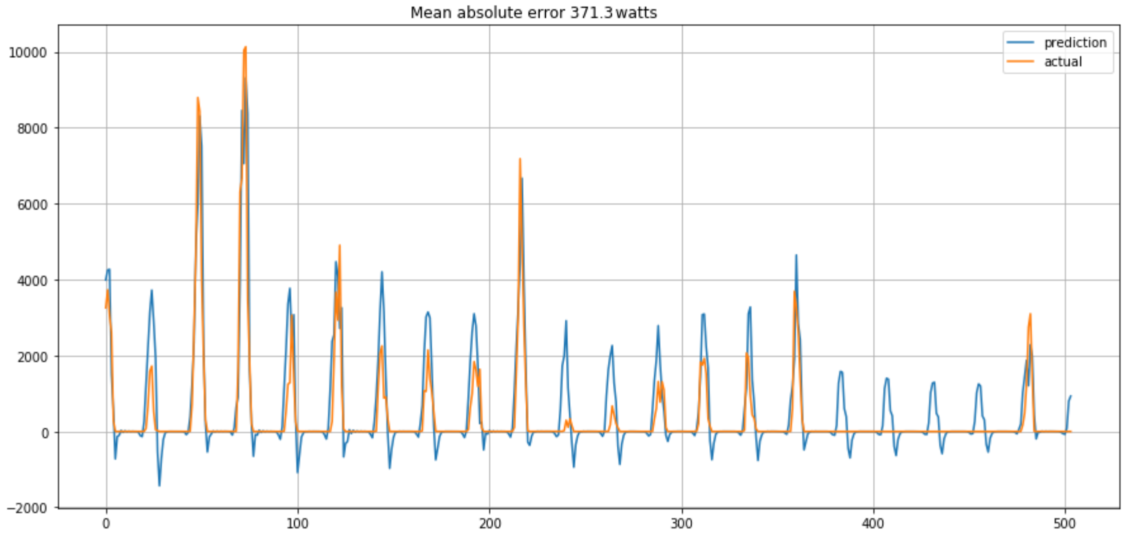

Figure 32. Plot of the electricity production forecasting based on the SARIMAX model in winter time by the carport panel.

6 DISCUSSION

In this work, several prediction methods have been discussed based on scientific articles and case studies. Linear Regression, SARIMAX, and SARIMA models were selected during the literature review and built by performing the parameters adjustment and feature selection.

According to the forecasting results, both of these approaches deserve attention. For the electricity consumption of some LUT buildings, Linear Regression-based model with the weather features shows a bit better result than SARIMA-based model. For some buildings it shows a bit worse result. Anyway this deviation is too tiny to deserve attention (mean value between the accuracy result is equal to 0.7 kWh or 0.02 for R^2 measurement).

For the electricity production SARIMA-based model shows better result than the Linear-Regression based model with weather features. SARIMA improves the accuracy for every production panel set. Mean value of the improvement is equal to 0.29 kWh or 0.06 for R^2 measurement. Presumably it is because of the seasonal differences.

During this Thesis, a couple of problems have been faced. Encountered problems are related to the differences in seasonality between the electricity production and consumption data (this problem has been solved with using of different models) and missing a large part of the Local Weather Station dataset (this problem has been solved by the adjusting of the target variable data length).

Forecasting results can be considered quite satisfying but could be better. It is worth to try other models suitable for time-series or maybe do more experiments with features, for instance, the random forest method. Another possible way to improve is to build a model using extended data, for instance, gather more relevant features or get a dataset with information for much more than couple of years.

Another presumable step is to implement the real-time forecasting system based on models described in this Thesis. For instance, it can be developed as the web-application.

7 CONCLUSION

Review of the methods has been done using case-studies articles and review papers. Input datasets have been analyzed and preprocessed with all necessary steps. Linear Regression, SARIMA, and SARIMAX models have been built. Forecasting models are built for the prediction and consumption purposes.

Results related to the production forecasting can be estimated with the R_2, because unfortunately the structure of the data spoils the MAPE estimation result. Plots analysis show that they are more accurate then the MAPE error can tell. It is caused by the wide spread of production results during different seasons.

It has been discovered based on the experiment that the linear regression model and SARIMA model show approximately the same results for the prediction of the electricity consumption (mean R^2 value for each building is equal to 0.95). It has been also discovered that SARIMA model shows better results for the electricity production that the Linear Regression. For the model based on SARIMA, mean $R2$ value is equal to 0.86.

The reason of differences between a suitable model can be a seasonality of the production data and a difference amongst the data structure. Local Weather Station provided such a small amount of data, otherwise, predictions would be more accurate.

Accuracy of the forecasting is quite high. It can be compared with the results achieved in investigated case studies and is considered as good by the representative of the Energy Systems department. So objectives set before the beginning of the work are impleme nted but of course, there is much to grow.

REFERENCES

[1] Y. M. Hassan, A. M. Mat Daut, H. Abdullah, A. H. Rahman, M. P. Abdullah, and F. Hussin. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: a review. *Renewable and Sustainable Energy Reviews*, 70:1108–1118, 2017.

[2] K. Methaprayoon, W. Lee S. Rasmiddattaa, J. Lia, and R. Ross. Multistage artificial neural network short-term load forecasting engine with front-end weather forecast. *Energies*, 43:1410–1416, 2007.

[3] W. Zeyu and R. S. Srinivasan. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75:796–808, 2017.

[4] Production figures - Green Campus. https://www.lut.fi/web/en/green-campus/green-campus-in-numbers/production-figures, 2018. [Online; accessed February and 15, 2018].

[5] C. Deb, F Zhang, J. Yang, S. E. Leea, and K. W. Shaha. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, pages 902–924, 2017.

[6] T. Czernichow, A. Piras, P. Caire Y. Jaccard B. Dorizzi, and A. Germond. Short term electrical load forecasting with artificial neural networks. *Engineering Intelligent Systems*, 2:85–99, 1996.

[7] C.-H. Wua, G.-H. Tzeng, and R.-H. Lin. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications*, 36:4725–4735, 2009.

[8] H. Khosravani, M. Castilla, M. Berenguel, A. Ruano, and P. Ferreira. A comparison of energy consumption prediction models based on neural networks of a bioclimatic building. *Energies*, 9:1–24, 2016.

[9] K. Kavaklioglu. Modeling and prediction of Turkey's electricity consumption using support vector regression. *Applied Energy*, 88:368–375, 2011.

[10] M. Olsson and L. Soder. Modeling real-time balancing power market prices using combined SARIMA and Markov processes. *IEEE Transactions On Power Systems*, pages 443–450, 2008.

[11] P. Chujau, N. Kerdprasop, and K. Kerdprasop. Time series analysis of household electric consumption with ARIMA and ARMA models. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 1, 2013.

[12] G. R. Newsham and B.J. Birt. Building-level occupancy data to improve ARIMA-based electricity use forecasts. *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 1:13–18, 2010.

[13] K.Y. Chen and C. H. Wang. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in taiwan. *Expert Systems with Applications*, pages 254–264, 2007.

[14] E. Egrioglu, C. H. Aladag, U. Yolcu, M. A. Basaran, and V. R. Uslu. A new hybrid approach based on SARIMA and partial high order bivariate fuzzy time series forecasting model. *Expert Systems with Applications*, pages 7424–7434, 2008.

[15] M. Bouzerdoum, A. Mellit, and A. Massi Pavan. A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, pages 226–235, 2013.

[16] Z.Mohammed and P.Bodgeri. Forecasting electricity consumption in New Zealand using economic and demographic variables. *Energy*, 30:1833–1843, 2003.

[17] V. Bianco, O. Manca, and S. Nardini. Electricity consumption forecasting in Italy using linear regression models. *Energy*, 34:1413–1421, 2008.

[18] H. Tanaka, S. Uejuma, and K. Asai. Linear regression analysis with fuzzy model. *IEEE Transactions on Systems and Man and and Cybernetics*, 12:903–907, 1982.

[19] K. B. Song, Y. S. Baek, D. H. Hong, and G. Jang. Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE Transactions On Power Systems*, 20:96–107, 2005.

[20] F. F. Nobre, A. B. S. Monteiro, P. R. Telles, and G. D. Williamson. Fuzzy modeling to forecast an electric load time series. *Statistics in Medicine*, 55:395–404, 2015.

[21] Prophet: forecasting at scale. https://research.fb.com/prophet-forecasting-at-scale/, 2018. [Online; accessed April, 17, 2018].

[22] G. Dudek. Short-term load forecasting using random forests. *Intelligent Systems'2014: Proceedings of the 7th IEEE International Conference Intelligent Systems*, 2:821–828, 2015.

[23] Seasonal ARIMA with python. http://www.seanabu.com/2016/03/22/time-series-seasonal-ARIMA-model-in-python/, 2016. [Online; accessed May, 06, 2018].

[24] D. Kwiatkowski, P. C. B. Philips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54:159–178, 1992.

[25] H. S. Hippert, C.E. Pedreira, and R. C. Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on Power Systems*, 16:44–55, 2001.

[26] U. Kumar. ARIMA forecasting of ambient air pollutants (o3 and no and no2 and co). *Stochastic Environmental Research and Risk Assessment*, pages 751–760, 2010.

[27] F. F. Nobre, A. B. S. Monteiro, P. R. Telles, and G. D. Williamson. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine*, 20:3051–3069, 2001.

[28] Regression with lagged explanatory variables. http://personal.strath.ac.uk/gary.koop/OheadsChapter8.pdf, 2016. [Online; accessed May, 06, 2018].

[29] P. Zhang, X. Wu, and X. Wang S. Bi. Short-term load forecasting based on big data technologies. *Journal of Power and Energy Systems*, 1:59–67, 2015.

[30] J. L. Deng. Grey modeling in energy-gathering space. *Journal of Grey System*, 19:301–308, 2007.

[31] K. Methaprayoon, W. Lee, S. Rasmiddattaa J. Lia, and R. Ross. Multistage artificial neural network short-term load forecasting engine with front-end weather forecast. *IEEE Industrial Applications*, 43:1410–1416, 2007.

[32] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[33] V. Barnett and T. Lewis. *Outliers in statistical data*. John Wiley And Sons Ltd., 3rd edition, 1974.

[34] M. King. *Statistics for Process Control Engineers: A Practical Approach*. John Wiley And Sons Ltd., 1st edition, 2017.

[35] Anomaly detection with local outlier factor. http://scikit-learn.org/stable/autoexamples/neighbors/plotlof.html, 2018. [Online; accessed March and 05, 2018].

[36] Find outliers in data - MATLAB isoutlier. https://ch.mathworks.com/help/matlab/ref/isoutlier.html, 2018. [Online; accessed March and 5, 2018].

[37] F. W. Scholz and M. A. Stephens. K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, pages 918–924, 2012.

[38] Detect and replace outliers in data - MATLAB filloutliers. https://ch.mathworks.com/help/matlab/ref/filloutliers.html, 2018. [Online; accessed March, 06, 2018].

[39] Weather station at LUT campus. https://www.lut.fi/web/en/green-campus/green-energy-and-technology/weather-station, 2018. [Online; accessed April, 18, 2018].

[40] D. Lee and R. Baldick. Short-term wind power ensemble prediction based on gaussian processes and neural networks. *IEEE Transactions On Smart Grid*, 5:501–510, 2014.

[41] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, H. Abdullah, and R. Saidur. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33:102–109, 2014.

[42] T. Senjyu, K. Takara, K. Uezato, and T. Funabashi. One-hour-ahead load forecasting using neural network. *IEEE Transactions on Power Systems*, 17:113–118, 2002.

[43] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence*, 2:1137–1143, 1995.

[44] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 2007.

[45] How to backtest machine learning models for time series forecasting. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/, 2016. [Online; accessed May, 05, 2018].

[46] Cross-validation for time series. https://robjhyndman.com/hyndsight/tscv/, 2016. [Online; accessed May, 05, 2018].

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

[48] B. Yildiz, J. I. Bilbao, and A. B. Sproul. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, 73:1104–1122, 2017.

[49] S. Hyojoo and K. Changwan. Short-term forecasting of electricity demand for the residential sector using weather and social variables. *Resources, Conservation and Recycling*, 123:200–207, 2017.

[50] H. Wan. Load forecasting via deep neural networks. *Procedia Computer Science*, 122:308–314, 2017.

[51] A. Kadir and M. N. El-Gohary. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, 81:1192–1205, 2016.

[52] A. A. Weiss and A. P. Andersen. Estimating time series models using the relevant forecast evaluation criterion. *Journal of the Royal Statistical Society*, 147:484–487, 1984.

[53] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:697–688, 2006.