# Mat-2.4177:
# Seminar on Case Studies in Operations Research
# Forecasting the Consumption of District Heating
# Final report

Lasse Lindqvist (Project manager)
Juulia Happonen
Joona Karjalainen
Mika Juuti

Client: Fortum

May 17, 2014

# Contents

# 1 Preface

This is the final report of our project for the course Seminar on Case Studies in Operations Research in Aalto University in 2014. Fortum was the client of the project. The main objective of the project was to develop a model to predict heat demand in district heating. A basic introduction to district heating is presented in section two. Applicable modeling techniques will be introduced and the results of a literature review on earlier research will be presented. Finally, the modelling approaches used will be discussed with the results and recommendations.

# 2 District heating

District heating is the main heating form in Finland, having a market share of 46 % of the total heat production. In the largest cities, it covers over 90 % of the provided heat [17]. In Europe, the market share district heat has a market share of some 10 % [16]. District heating is used for residential, commercial and industrial purposes. It is mainly used for space heating and hot-water consumption. Its applications also include space cooling, which can be used for cooling of data centers, for instance.

District heating is based on central production of the heat that is distributed to the consumers through a network of insulated pipes. Distribution of the heat is operated in a network covering all the customers. Due to the large size of the network, there is a lag of a few hours in the transfer from the plant to the client. The district-heating water can be hot, chilled or in the form of steam. In Finland and Europe, mainly hot water is in use [2] [10]. The temperature of the water in the pipes has some dependency on outside temperature [10]. Commonly the water temperature stays between 65 and 115 °C.

The district heating water is first heated at the heat production plant and directed to the distribution network. After that, the water moves to the heat distribution center of the building where it heats the consumer water. District-heating water itself is not consumed in the building. Instead, after cooling down it is directed back to the plant to be reheated. Some water losses can occur through leaks in the network, and losses in the network need to be replaced with fresh water.

Building a district-heating network requires big investments in both infrastructure and production technology. Heat losses occur, which is why district-heating is most suitable in areas with high density in both population and buildings and is mainly used in big cities [2]. Losses and investment costs are the bigger the longer distances the water needs to be transferred.

Main share of district heat in Finland is based on combined heat and power (CHP) production, however some heat is generated in separate heat plants as well [10]. CHP is a form of production where electricity and heat are generated simultaneously, thus getting better efficiency than if electricity and heat were generated separately. By the second law of thermodynamics, only some of the supplied energy can be converted to electricity. The remaining energy can then be utilized for district heating. In other Nordic countries, incineration of municipal solid waste is often used as the source of energy for district heating [14].

Compared to the methods of heating buildings individually, the capability to produce heat centrally enables achieving the economies of scale. For example, the fuel costs are decreased. In addition, many different energy sources can be utilized in the district heat production. For instance, heat from waste incineration and industrial processes are sometimes used in district heating. Downsides, on the other hand, are high investment costs and inefficiency in sparsely populated districts [11].

# 3 SARIMA and dynamic regression models

The Box-Jenkins methodology is a common approach to time series modeling and forecast. Following the notation in [4], the objective is to model the time series of interest $X_t$ by defining a differenced series $Y_t$

$$Y_t = (1-B)^d(1-B^s)^D X_t \tag{1}$$

where B denotes a lag operator, $s$ is the length of the season and $d$ and $D$ are the orders of differences. The differenced series is then modeled as

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t \tag{2}$$

where $\phi, \Phi, \theta$ and $\Theta$ are polynomials of orders $p$, $q$, $P$ and $Q$ (respectively) and $Z_t$ is a white noise process. A model of this type is called a seasonal autoregressive moving average model and it is denoted by ARIMA$(p,d,q)[s](P,D,Q)$.

The modeling approach consists of three stages : model identification, parameter estimation and model checking [21]. The model identification stage consists of selecting the orders $p$, $q$, $P$, $Q$, $d$, $D$ and the length of the season $s$ by e.g. visual inspection of the estimated autocorrelation and partial autocorrelation functions. The coefficients in the polynomials are estimated using a chosen method, such as maximum likelihood. The model is validated by using autocorrelation and partial autocorrelation plots of the residuals and statistical tests. Ideally, the residuals should be uncorrelated.

The Box-Jenkins methodology can produce several feasible models. These models can be compared using information criteria and error measures. In practice,it is possible to fit several models with different $p$, $P$, $q$ and $Q$, and compute the information criteria and the error measures and do the diagnostic tests afterwards.

An exogenous variable $T$ can be introduced in the time series model to explain the variation of $X_T$. The R software package provides routines to estimate models of the form

$$X_t = \beta T_t + n_t \tag{3}$$

where the error term $n_t$ satisfies the equation

$$\phi(B)\Phi(B)(1-B)^d(1-B^s)^D n_t = \theta(B)\Theta(B)Z_t. \tag{4}$$

This is a dynamic regression model, or a regression model with ARIMA errors. The model building can be done in the same way as with ARIMA models, and the same diagnostic checks and statistical tests can be used. With an exogenous variable, it is also appropriate to test the cross-correlations of the residuals and $T$, which should ideally be close to zero [19].

# 4 Static time series models

In contrast to ARIMA and other dynamic regression models, in a static time series model the present values are not dependent on past values. The four main assumptions of classical linear regression models are linearity, independence, homoscedasticity, and normality. Firstly, the mean dependent variable is supposed to depend linearly on the independent variable or variables. The assumption of independence here means that the error terms do not depend on each other. Homoscedasticity, on the other hand, refers to all error variables having a constant variance. Lastly, the error terms are assumed to follow the normal distribution. [29]

## 4.1 Simple linear regression

Let us follow the notation in [29]. The most straightforward regression model is the simple linear regression model, where causal behavior of the dependent variable is explained by one independent variable. The model is typically written in the form

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{5}$$

where $y$ is the dependent variable, $\beta_0$ is the $y$ intercept, $\beta_1$ is the slope of the linear regression line, $x$ is the independent variable, and $\varepsilon$ is the random error. In literature, the dependent variable is sometimes also referred to as the response variable. Also, the independent variable can be called the explanatory or the predictor variable.

We can state the simple linear regression model with respect to $n$ pairs of data in the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, 2, \ldots, n. \tag{6}$$

For all $\varepsilon_i$, we assume that $\mathbb{E}(\varepsilon_i) = 0$, variance is constant $\mathrm{Var}(\varepsilon_i) = \sigma^2$. In addition, it is assumed that all $\varepsilon_i$'s are independent.

A commonly used method to find good estimates for coefficients $\beta_0$ and $\beta_1$ is the principle of least squares estimation. For the simple linear regression model, the coefficients are estimated such that the the sum of the squared distance from the actual, observed $y_i$ and the predicted response $y_i$ is minimized. The aim in this is to determine estimates for the parameters by choosing the regression line that fits closest to all data points.

## 4.2 Multiple linear regression

Often the behavior of the response variable cannot be explained by one predictor only. In a multiple linear regression model, the causal changes in the dependent variable are explained by several independent variables. The predictor variables can contain other predictors and their higher-order terms. The multiple linear regression model can be stated in the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \tag{7}$$

or in the matrix-format

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8}$$

If there are only few possible explanatory variables, one can form all possible combinations of regressors and evaluate them with a help of model selection criteria. However, as the number of the regressors increases, the method of stepwise selection can prove to be more efficient. There are three different stepwise procedures: backward elimination, forward addition and stepwise search.

Backward elimination starts with a model that has all the possible regressors as explanatory variables. Each variable is tested separately with an F-test through a comparison of the initial model and the model where that particular variable has been removed. The least significant variable that has the highest $p$-value in F-test is removed from the model. This is then repeated until the $p$-value of the least significant predictor of the elimination round is lower than the threshold value and the final model is found. It is also computationally the fastest procedure [29].

Forward addition, on the other hand, starts with the model that has only the intercept term. The procedure tests all simple linear models and compares them to the original null model. The most significant explanatory variable with the smallest $p$-value is added to the model if the $p$ is lower than a specified threshold value. The algorithm continues with identification of the next most significant predictor to be added and is repeated until a regressor does not qualify to be added, i.e. has a too high $p$-value.

The stepwise search is somewhat similar to the forward addition procedure but after each addition it checks the model for predictors with the $p$-value higher than the threshold value and removes them. The stepwise search is meant to capture the best features from the other two stepwise procedures and is believed to perform best of the three. [29]

# 5 Earlier research

There is a lot of research on forecasting the consumption of electricity. Electricity is sometimes used for heating but serves mainly other purposes. District heating has seen less treatment in the literature.

## 5.1 Models for predicting heat demand in district heating

Static linear models and SARMAX models have been used for heat demand predictions in literature [6] [7] [12] [18]. Some differences between studies arise from the method of transforming the outside temperature (in degrees) into an explanatory variable in the model. As the relationship between the temperature and the heat demand is not linear, polynomials [6] and piecewise linear transformations [7] [12] have been used. The social component is generally included in the models either as a SARMA process (e.g. [6]) or by adding indicator variables in model e.g. for each hour of the week (as in [18]).

Most studies use more than one error measure to evaluate the results. However, time series analysis of the prediction errors and model comparison with information criteria is often omitted. Most studies note that the social component and the outside temperature are the most important explanatory variables, and that accurate weather forecasts are crucial for the success of the prediction [6] [7].

Dotzauer (2002) [7] presents a simple model for predicting loads in district-heating systems. The heat demand is modeled as a sum of a temperature-dependent component and a social component. The outside temperature is transformed using a piecewise linear function with equal numbers of data points corresponding to each segment. The social component is modeled using one indicator variable for each of the 168 hours of the week, and the coefficients are estimated from the data. Dotzauer's model is static in the sense that neither the temperature-dependent component nor the social component makes explicit use of previous inputs and outputs. Data on wind and global radiation could be included in the model as a correction for the temperature-dependent part instead of including them as separate variables. A simple model is justified by lack of measured data and uncertainties in the forecasts, and it is noted that good temperature forecasts are crucial for making relevant heat-demand predictions. Due to the similarities between prediction of heat demand and electrical power, the author suggests that same types of models could be used for both. It is suggested that one should focus on improving the weather forecasts rather than developing more advanced load prediction algorithms in order to improve the predictions.

Grosswindhager, Voigt and Kozek (2011) [12] use a SARIMA model for short-term forecasting the system heat load in the district heating network in Tannheim, Austria. The outside temperature is transformed using a piecewise linear function and subtracted from the heat load, and a SARIMA(2,1,1)(0,1,1)48 model with half-hourly data is used to model the remaining time series. Model selection is done using the Bayesian Information Criterion, and the autocorrelations of the residuals are tested using the Ljung-Box test. It is mentioned that calendar effects (Christmas, Easter) should be treated with dummy variables, but they are not included in their model.

Nielsen and Madsen (2006) [22] present a grey-box model for heat consumption in district heating systems. The prior information about the physical processes, such as ventilation and heat transfer through windows is used to reduce the number of models to consider and to prevent overfitting. The missing values in the heat consumption data are not replaced. However, missing and outlying data about climate variables are replaced with methods such as time series decomposition and smoothing. The model is validated with positive results by examining autocorrelation, partial autocorrelation and inverse autocorrelation functions of the residuals as well as cross-correlations. Likelihood ratio tests are used to test the statistical significances of the explanatory variables and the authors report that wind speed, solar radiation and temperature are significant, whereas the interaction between the wind and temperature is not. The difference between holy and half-holy days (including saturdays) is not statistically significant.

An introduction to heat-load modeling in large systems is presented by Heller (2002) [13]. It is noted that there are large variations in the significance estimations of different components in literature. The ambient temperature is by far the most important variable, whereas wind and humidity seem to be less significant.

Chramcov (2010) [6] applies the Box-Jenkins methodology to forecast heat demand in district heating in Most-Komořany and Litoměřice, Czechia. The presented model is a sum of a social component and a temperature-dependent component. The social component is modeled using SARIMA with daily and

weekly periods, whereas only the current value of the temperature is included in the model, not the previous values. A piecewise linear function and a third-degree polynomial are presented for transforming the outside temperature into a meaningful explanatory variable. However, no preference is given to one or the other. The performance of the model is evaluated using the MAPE and RMSE error measures, and the author reports that the accuracy of predictions decreases on the weekend days. Again, it is concluded that the accuracy of the temperature forecasts is crucial.

Lahdelma et al. (2012) [18] use a static time series linear regression model to predict the heat demand in district heating. The current temperature and a social component are used as explanatory variables. The temperature is used in the model without any transformation. The social component is modeled by using one parameter for each hour of the period, where the length of the period was either 168 hours or 72 hours. The inclusion of mid-week holidays is found to have very little effect on the accuracy of the predictions. The predictions are more accurate for large clients and sums of smaller clients compared to predictions of the consumption by individual small clients. The authors predict that the performance of different models will improve as more accurate information about consumption and the climate will be available.

## 5.2 Other approaches to time series prediction

Other approaches have been used for time series prediction, including artificial neural networks (ANN) and Gaussian processes. Compared to Box-Jenkins and static linear regression, these approaches seem to lack the wide availability in software packages and the easily applicable recipes for model building. Much research has been done in applying different models for energy and electricity consumption, whereas studies on district heating are more rare. At the time of writing, no studies about applying Gaussian processes to heat demand prediction are avalable. Tools for using neural networks are included in the Neural Network Toolbox for MATLAB and in the neuralnet library for R. Vanhatalo et al.(2013) [24] have implemented tools for using Gaussian Processes in the GPstuff toolbox for MATLAB.

Artificial neural networks are used by Bakker et al. (2008) [3] to predict the heat demand of individual households based on historical heat demand and weather influences as input. They conclude that neural network techniques can be used for such predictions and consider the results promising. However, their prediction horizon is only 24 hours. They suggest that including wind speed and other factors could improve the predictions.

Another application of ANN is presented by Yalcintas and Akkurt (2005) [27] for predicting a building's energy use. They conclude that the method is suitable for making energy predictions. However, they note that their approach cannot be considered a generalized method for energy predictions and they list several issues related to applications, such as the difficulty of determining the most suitable ANN architecture.

Wang and Meng (2012) [25] combine ARIMA and ANN models to predict energy consumption in Hebei province in China. They note that ARIMA cannot deal with nonlinear relationships, while neural networks alone are not able to handle linear and nonlinear patterns equally well. The RMSE, MAE and MAPE error measures are used to show that their model performs better than ARIMA

and ANN models when used separately.

Yan and Malkawi (2013) [28] use a Gaussian process model for predicting the building cooling and heating consumption. In their application the Gaussian process model produces better predictions than a neural network, although they note that further comparisons between the methods would be required. The authors note that the results of Gaussian process modeling express the uncertainty of predictions, whereas the uncertainty could not be quantified explicitly with neural networks. Leith et al. (2004) use Gaussian process models to predict electrical loads in Ireland with promising results. MAPE and MSPE are used as error measures for evaluating the models. The authors express some concerns related to difficulties in computation and the use of the models in long-term prediction.

# 6 Data

The data we have in our use contains usage data from 2013. The usage is recorded hourly giving us 8760 data points for every building. Temperature data is also hourly data collected from one measurement station. The essential information can be considered to be:

1. The quantity to be forecasted (MWh).
2. Characteristics of each building
   - Heated volume ($m^3$)
   - Year of construction
   - Type of building
3. Temperature from Sepänkylä measurement station in Espoo.

## 6.1 Omitted data and comments

The original data defines the building volume, floor area and heated floor area for each building. The heated volume gives a more appropriate estimate of the building's heat consumption than the floor area, and the floor area will thus be omitted. The year of construction surely affects the energy efficiency of the house and thus also the heat consumption in some way.

The original data also contains information such as the year of restauration (if any). The year of reconstruction is not reported in all cases. Additionally the term causes some ambiguity; in some cases even repainting the building is considered restauration. Thus, the reconstruction year will be omitted.

The temperature measurements contain some blank measurements. In order to use the data for modeling, these blanks must be filled. Due to the lack of better knowledge, these values are to be interpolated. When using the model for real forecasting, forecasts for temperature are also used. Usually forecasts for every hour should be available.

The same goes for consumption data. This data does not contain blanks, but it does contain some erronous values. These values too must be replaced with interpolated values. Automating consumption correction could be a hard problem, but an important one. Detecting and correcting errors has to be

implemented. We will describe in the report how erroneous values can affect the resulting model and give unacceptable results.

### 6.2 Correcting errors

Temperature data contains some blank cells. To use the data for modeling blanks must be filled. Without better knowledge, these values are to be interpolated using simple linear interpolation. When using the model for real forecasting, forecasts for temperature are also used. Usually forecasts for every hour should be available.

A simple automatic correction for errors in the consumption data can be constructed the following way. If the maximum possible consumption per hour is known, values larger than this can be corrected to equal the average of the preceding and successing values. If the successing value is also too high, the value can be changed to equal the preceding value only. This simple procedure should be powerful enough since errors are relatively rare.

A more powerful way to correct errors would be to use forecasted values. For every value a forecast would be made and this way errors would be corrected from beginning to end. This, however, would be time-consuming in processing time even if the forecasting was fully automated. It is not clear if the improvement in results would be noticable enough to justify this method.

Literature concerning error correction is broad. One simple way to detect erroneous values is Cook's distance. Cook's distance measures the effect of deleting the chosen data point.[26]

A whole field of statistics called data editing is centered around finding and correcting errors.[5]

## 7 Building the model

### 7.1 Outliers

Most data sets have outliers. Examining the individual buildings' consumptions reveals that many of the measurements are not updated regularly, meaning that the value does not change often enough (e.g. only once a month). This brings problems in understanding the relation between temperature and heat usage and between time and heat usage. Figure 1 shows the variability of the data. Examining the data reveals that a minimum requirement that the data varies every third measurement is appropriate. We can see that tens of measurements practically are constant throughout the one year long time series.

### 7.2 Groups

The grouping of building types is based on two fundamental differences between the buildings: the temporal rhytm of buildings and the relationship between heating volume and mean consumption. The correlation between the average consumption for each building in the groups and the building's volume is shown in Table 1.

There is a linear relationship between the heating volume and the mean consumption of the buildings. On closer inspection, it becomes apparent that some of the outlying data points follow a different relation between the heating

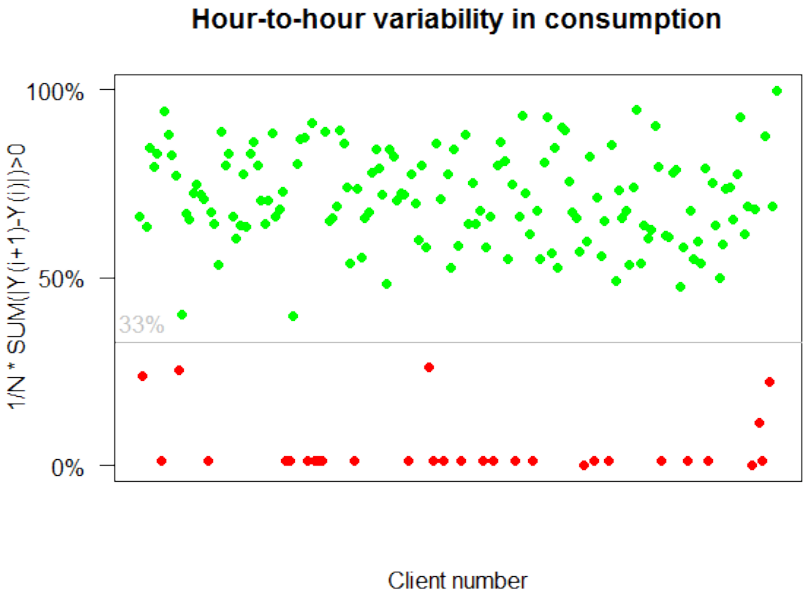


Figure 1: The variability of the data. The figure shows how often the heat usage data changes values in average. The data measurements that vary more than 33% in average are utilized in the model building.

volume and the mean consumption. Swimming halls have a very different relationship between the heat consumption and building volume compared to other sport halls, as can be seen in Figure 2. The buildings were divided into five groups: residential, daycare centres/work, commercial centres, swimming halls and other type of sport buildings. The groups are shown in Figure 3 in a more narrow interval. The coloring is the same as in Figure 2. Clearly, most of the residential data regards small residential buildings.

### 7.3 Lags of temperature

We can consider the temperature as a control variable to a discrete dynamic system. The step response of this system tells the lag, i.e. how long it takes before the temperature outside affects the temperature inside and thus the heat consumption of a building. There are some different ways to estimate the step response of a linear system. In our case, it is not possible to change the control variable freely (the control variable is temperature). One way is to use the correlation analysis algorithm CRA. This method gives a quick insight of the time constants and time delays of the system. This model assumes that the input $u(t)$ is uncorrelated with the error in the model $e(t)$

$$y(t) = \sum_{k=0}^{\infty} g_k u(t-k) + e(t) \tag{9}$$

This means that the correlation analysis does not produce credible results

Table 1: The correlations between volume and average consumption for buildings in different groups.

| Group | $h$ |
|---|---|
| All values | 0.63 |
| Residential | 0.90 |
| Daily | 0.77 |
| General sport | 0.70 |
| Swimming | 0.89 |
| Centre | 0.66 |

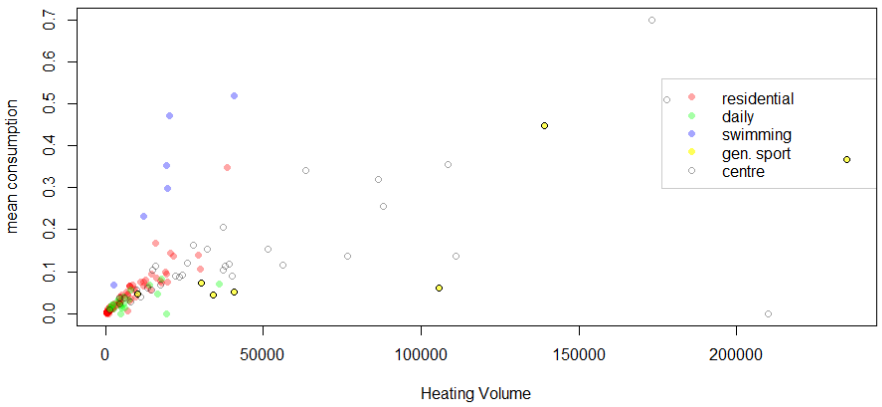


Figure 2: The groups in the data.

in the case of output feedback. In our case this would mean that the use of central heating affects the future temperature in the measuring station, which is implausible. The correlation analysis is done with the algorithm CRA as described in [19]. The input signal is whitened first with an AR-filter.

In order to understand the essential time lags for the temperature variable in a SARIMAX model, correlation analysis is used. The temperature will be regarded as the control signal in this model. Correlation analysis relies on the fact that the control signal is a random variable with a zero mean. In this case the model error and the filtered control signal will have an expected value of zero and the correlation between the variables can be seen in the cross-correlation function. The temperature should clearly not be random; cold weather at hour $t$ clearly implies cold weather at hour $t + 1$. It is necessary to whiten the temperature with a linear filter.

Forecasting the temperature is difficult and it is clear that a linear process model cannot capture the full dynamics of the process. However, we can create a linear process model that is sufficiently good that the residuals are white for a couple of hours forward in time. This will produce a linear filter that can be used to filter the input signal (temperature) and output signal (heat usage). The residuals from this model can be tested against the whitened output in correlation analysis.

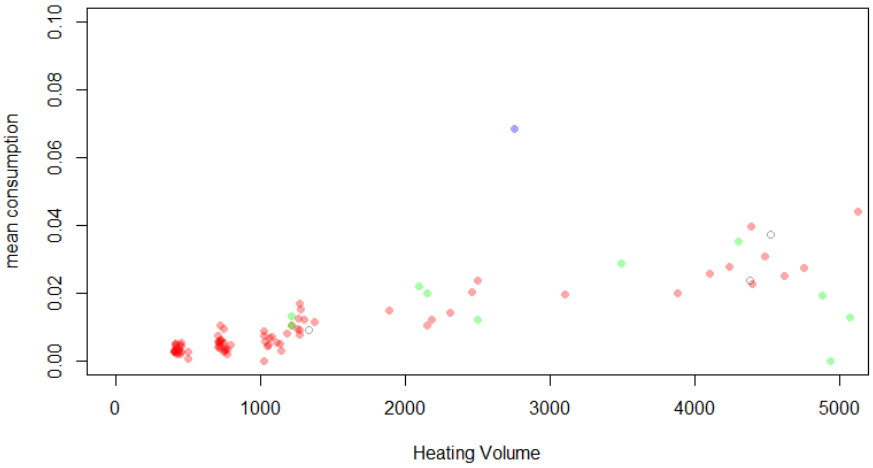


Figure 3: The group separation.

The temperature time series is whitened with a AR(4,1,0)-filter. The cross-correlation analysis did not produce credible results that suggested that other lags than 0 could be useful in estimating the heat usage. Other lags can be credibly motivated.

# 8 Validating and verifying models

Crossvalidation with real weekly data is used to test the model validity. The data is divided into two parts: a wintertime model and a summertime model. The winter week starts with a the date and hour 2013-04-04 09:00 EEST and ends with 2013-04-11 08:00 EEST. The summer week is validated between 2013-10-20 03:00 EEST and 2013-10-27 02:00 EEST in the yearly daylight saving time model.

## 8.1 Measures of error

Measuring the error in forecast can be done in many different ways. Cumulative Forecast Error (CFE) is the most straightforward way. This is the simple cumulative sum of all errors. Mean error (ME) is the same but divided by the number of data points. This is a little more representative way of measuring. The problem with these is that negative and positive errors can average out and the measure can be close to zero even if the forecast is not good. Mean Percentage Error (MPE) is similar to ME but presents the error in percentages.

Mean Squared Error (MSE) is the sum of squared errors. Root Mean Squared Error (RMSE) is the square root of (MSE). RMSE is calculated as

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - x_t)^2}{n}}, \tag{10}$$

where $y_t$ is the real value and $x_t$ the predicted value for time $t$.

Mean Absolute Deviation (MAD) is the sum of the absolute values of errors. Mean Absolute Percent Error (MAPE) is similar to MAD but presents the error in percentages. MAPE is calculated as

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} |\frac{y_t - x_t}{y_t}|. \tag{11}$$

Mean Absolute Scaled Error (MASE) is a somewhat new measure of error. Proposed in 2006 by Australian statistician Rob J. Hyndman MASE measures scaled absolute errors. The scaling is done by a sum of naive forecasting. The formula for MASE is

$$MASE = \frac{1}{n} \sum_{t=1}^{n} \frac{y_t - x_t}{\frac{1}{n-1}\sum_{i=2}^{n} |y_t - y_{t-1}|} = \frac{\sum_{t=1}^{n} |y_t - x_t|}{\frac{n}{n-1}\sum_{i=2}^{n} |y_t - y_{t-1}|}. \tag{12}$$

Another measure of error is the residuals of the forecast. Autocorrelation with lag of one (ACF1) can be used as a statistic in studying error.

Academians have traditionally had a strong preference for RMSE and it is frequently used in forecasting. MAPE is perhaps the most widely used unit-free statistic. [1] MASE is a newer method with a somewhat different approach having a simple interpretation.[15] With all different statistics having their strengths and weaknesses, we will use these three to study the success of the forecast. They are by no means the only three possible. The choice is somewhat arbitrary.

Maximum Absolute Deviation (MAD) is yet another error measure that tells how much the forecast has been wrong in the worst case in the forecasted period. This measure might be especially interesting to get a hint of the accuracy of the model. The unit for this measure is the same as for the forecasted value, in this case MW. The positive error measure is

$$MAD_p = \max_t x_t - y_t. \tag{13}$$

# 9 Modeling with ARMA

## 9.1 Building the model

To use a simple SARIMAX model we can do the choice of order automatically by the means of information criteria. Using the temperature and its third power as the external variable and Akaike Information Criterion with a correction for finite sample sizes (AICc) as the information criteria we get the following.

From the beginning of normal time (as opposed to summer time) until the end of year 2013 we have 1579 hours. We leave 168 hours (a week) of these out to test our prediction. Estimating an ARIMA(1,1,1)[24](6,1,7) model for the 1411 hours gives us the following.

Figure 4 shows the forecast with model ARIMA(1,1,1)[24](6,1,7). We can calculate the prediction error for this in many ways. Using MAPE we get MAPE = 8.229122. This is somewhat big. Calculating MASE gives MASE = 2.05344. This means that the error is twice as big on average compared to using always the previous hour as our guess. This does not mean we could use the previous

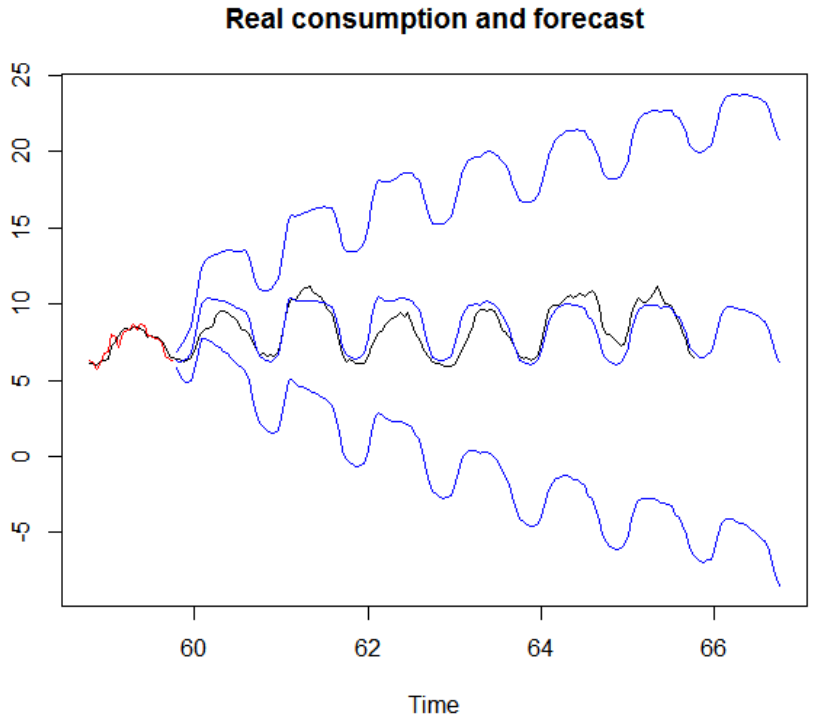


Figure 4: Forecast with ARIMA(1,1,1)[24](6,1,7). In red is the real consumption, with black the future real consumption. Blue lines are the forecast and 95 % confidence levels.

hour for the whole week. We can only use it for the next hour since we don't know the real values into future.

Figure 4 also shows the confidence intervals. It should be noted that these are not sensible in the end because the consumption cannot be negative. The levels are just normal approximations.

## 9.2 Uselfulness of temperature as external variable

The usefulness of temperature as an independent variable can be questioned since the cycle of temperature is mainly the same as the 24 hour day cycle. Even though previously discussed literature showed the importance of temperature, it is wise to check this with our data.

Figure 5 shows the difference of forecasts with and without temperature with the same data as before. The differences might seem small.

Table 2 shows the forecast error measures for the two different forecasts. We can see that the forecast with temperature is consistently better at least with this data. The table also shows the error measures for forecast with only linear component of the temperature. It can be seen that it is very much the same as the forecast with third power though a little better. It is therefore not clear

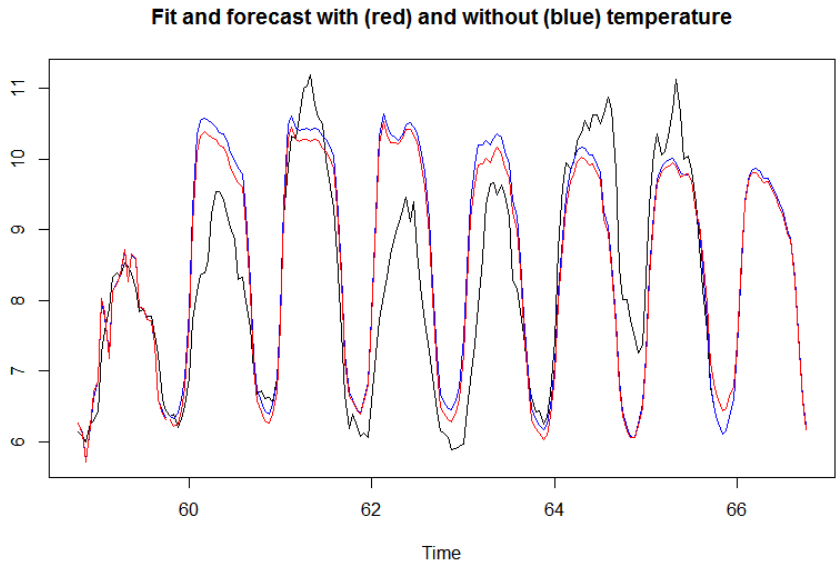


Figure 5: Forecast with ARIMA(1,1,1)[24](6,1,7). In red we have the forecast with temperature and its third power as an independent variable and in blue without it.

Table 2: Measures of error for different models.

| Measure / Model | No temperature | Temperature | Only linear temperature |
|---|---|---|---|
| RMSE | 1.002158 | 0.9312565 | 0.93962 |
| MAPE | 8.760943 | 8.229122 | 8.295957 |
| MASE | 2.173188 | 2.05344 | 2.069228 |

whether the third power should be included in the model. It is important to compare the forecast error measures and not the model because the temperature would always win the model competition due to more variables.

## 9.3 Computational considerations

Fitting an ARMA model is heavy in computational sense. The choice of the final model, its parameters $p$, $P$, $q$, $Q$, $d$ and $D$ can be automated in R with a function called `auto.arima`. With this function the choice can be made stepwise but even so the process can take up to an hour with an average computer. With this in mind if this timeframe is too much the process could be changed or a more powerful computer must be used.

# 10 Modeling with static linear regression

## 10.1 Building the model

Static linear regression is an alternative to ARMA model with the advantage of computational simplicity among other things. Using static linear regression for modeling and forecasting is slightly different from time series models. The difference boils down to whether future heat consumption depends on previous values or not. When using static linear regression previous values are not used.

When using linear regression we have the consumption as the dependent variable. Independent variables include temperature and an indicator of time. Temperature can be handled similarly to the ARMA model.

Linear regression can be written in vector form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{14}$$

where $\mathbf{y}$ is the vector for consumptions, $\mathbf{X}$ is the matrix for the independent variables, $\beta$ is the vector for the constants and $\epsilon$ the error term.

How to use time in the model is a trickier question. There are three prominent approaches. First is to label every hour of day with a label of their own and have 24 indicator variables. The second approach uses one indicator for each day of week and thus has 168 variables. The third approach is something between these two. It uses 72 variables, 24 for weekdays and 24 for saturdays and sundays. When doing calculations one of these variables must be dropped out to avoid linear dependency.

In the case of 24 indicator variables and a linear temperature dependency, the estimate for a specific hour would be Equation (15):

$$\hat{y}_t = \beta_\tau \hat{x}_\tau + \sum_i^{24} \mathbf{1}_{(t,24)} \beta_i x_i, \tag{15}$$

where $x_i$ is the mean value of this time slot, $\hat{x}_\tau$ is the temperature estimate for the estimated time and $\beta$ are ordinary least squares estimates for the coefficients. $\mathbf{1}_{(t,24)}$ is the indicator function

$$\mathbf{1}_{(t,24)} = \begin{cases} 1 & t \bmod 24 = i \\ 0 & t \bmod 24 \neq i \end{cases} \tag{16}$$

that equals 1 if $t$ is divisible by 24 and 0 otherwise. The errors are assumed to be Gaussian.

## 10.2 Temperature

A static linear regression model was estimated for the sum of the heat consumptions in 178 buildings (hourly data in 2013). Using the temperature as an independent variable without transformation results in $R^2 = 0.8429$. Using a polynomial of third degree gives $R^2 = 0.8687$ and of fifth degree gives $R^2 = 0.8721$. Already with this analysis we can see that using more than third order is not wise and leads to overfits. This is true regardless that we get p-values lower than 2e-16 for all degrees up to fifth when fitting a polynomial of fifth order.

## 10.3 Structure of the week

Periods of 24, 72 and 168 hours were used to model the seasonality in the data, as was done in [18]. With more variables fitting the model gives a larger $R^2$. Therefore we must use something else to study which method is the best.

Using 24 or 168 variables is the easiest choice. When using them we don't need to explicitly know which hour is which. With 72 variables we need this information because the weekend has a special meaning.

Table 3: Error measures when using time periods in forecasting the heat consumption for the last week of 2013.

| Measure / Model | 24 | 72 | 168 |
|---|---|---|---|
| RMSE | 0.0038004 | 0.0039487 | 0.0041116 |
| MAPE | 6.5521 | 6.6046 | 6.9141 |
| MASE | 0.80030 | 0.80651 | 0.84002 |

Table 3 shows the error measures when using the fall time series to forecast the heat consumption for the last week of the year. The last week of the year is however a very special week because it includes the Christmas week. Results for another forecast is shown in table 4. In this case the forecast is for the last week in the spring before clocks are moved.

Table 4: Error measures when forecasting the consumption for one week in spring.

| Measure / Model | 24 | 72 | 168 |
|---|---|---|---|
| RMSE | 0.0068932 | 0.0051683 | 0.0072998 |
| MAPE | 11.941 | 8.6720 | 12.574 |
| MASE | 1.5944 | 1.1400 | 1.6831 |

Table 4 shows that in this normal week the forecast with 72 variables is clearly the best. Forecasts with 24 and 168 variables are equally bad compared to the one with 72. So far, this result would suggest that using 72 variables in linear regression is preferable.

## 10.4 The value of groups

The regression results during a week in autumn are shown in Table 5. This table shows the results of regressing the past usage data with the temperature and its third power with different lengths of past data examined. The data is regressed over a subset of the variables. In column "All data" all data points are used. In "no outliers" all data points that passed the initial requirements for varibility were used (i.e. not too many equal consecutive measurement values). The variability test was shown in Figure 1.

In "combined regression" the data points were separated into five groups and each group was regressed. The predictions and the realized values were summed for each day. The error measure was calculated for the summed values.

We can see that forecasting with the combined values generally produces smaller

errors than forecasting with all the data. This is because the groups use heating very differently according to the time of the day. Interestingly, the varying values regression fares more poorly than the "all data" regressors. However, the smallest maximum deviation are usually found by using the "no outliers" regressor. Regressing the data with the all values produces better error measures in average than with the "no outliers" values' regressor. This is because "all data" contains certain values that are largely constant throughout the measurement period. The data does not contain the proper, actual variability in the consumption.

The maximum absolute deviation (MAD) error is arguably the most important error measure. It tells how gravely the forecasting algorithm overestimates (+) or underestimates (−) the consumption at worst. Forecasting far too much means that an additional power plant needs to be started in vain. Forecasting too little means that some building might not get enough heating in the winter.

Table 5: The performance of the combined static linear regression models for the groups. The lower the error values are, the better. The lowest errors measured for each $h$ is written in bold characters.

| Error measure | $h$ | All data | No outliers | Combined regressors |
|---|---|---|---|---|
| MAPE | 24 | 7.938 | 8.798 | **7.724** |
| | 72 | 8.826 | 9.656 | **8.568** |
| | 168 | **7.415** | 8.591 | 7.515 |
| MASE | 24 | 1.923 | 2.148 | **1.844** |
| | 72 | 2.150 | 2.372 | **2.054** |
| | 168 | **1.797** | 2.108 | 1.80 |
| RMSE | 24 | 0.8637 | 0.8183 | **0.8135** |
| | 72 | 0.940 | 0.883 | **0.8827** |
| | 168 | 0.863 | 0.827 | **0.7931** |
| MAD | 24 | $+2.326, -1.195$ | $+\mathbf{2.137}, -\mathbf{1.102}$ | $+2.278 - 1.404$ |
| | 72 | $+2.402, -1.148$ | $+\mathbf{2.247}, -\mathbf{1.057}$ | $+2.382, -1.374$ |
| | 168 | $+2.639, -1.106$ | $+2.406, -\mathbf{0.9641}$ | $+\mathbf{2.20}, -1.198$ |

## 10.5 Temperature trend

We could use the temperature trend to divide the data into a summer/winter model. In Jaakko Luhtala's Master's thesis [20], the heat production at power plants were piecewise linearly dependent on the outside temperature. Figure (6) shows this switch. At 16 degrees the heat production basically saturated to a certain level. At this temperature, houses are not warmed anymore. However, people will still use warm water in the shower etc. Based on this, the summer time and winter time was divided into two parts: a winter model is used if the temperature trend is below 16 degrees and a summer model is used if the temperature trend is above 16 degrees. The temperature during the year, the temperature trend and the temperature switch is shown in Figure (7).

The validation period in the winter model is still 2013-04-04 09:00 - 2013-04-11 08:00. The validation period for summer was 2013-08-05 15:00 - 2013-

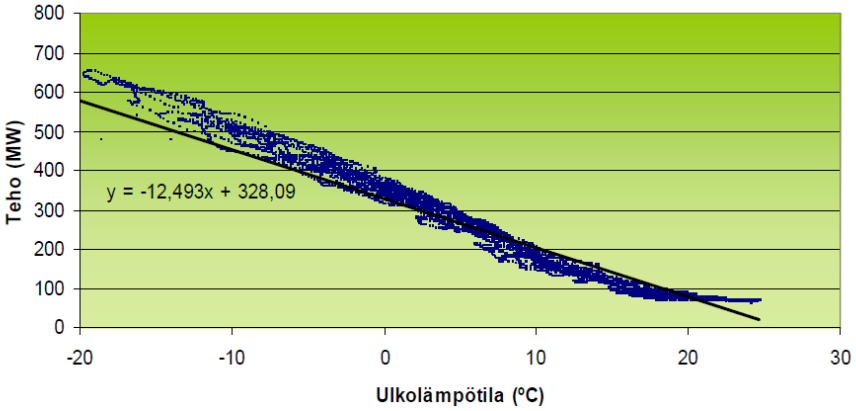


Figure 6: A scatter plot of the dispatch water temperature in from a power plant plotted against the outside temperature.

08-12 14:00. The results for forecasting on the winter and summer periods are shown in Table 6 and 7. We can see that the winter model performs much more favourably when the outliers are removed from the training and validation data. This is because the correct temperature-consumption relationship is captured better in the winter model.

All summer models have higher error measures than the winter models. The one exception is MAD, where the summer tests have smaller deviations. The reason for this is that the consumption is lower on summer. The maximum deviations will also remain lower.

The winter models capture the real behaviour better than the summer models because the effects that temperature has on the consumption are stronger.

When we compare the results to the model where summer time and winter time are determined using the daylight saving time, we see that the model with no outliers perfoms better (correct relationships between temperature and consumption). Comparing the results of the temperature trend models and daylight saving models reveals that all measurements are of the same size. The MAPE is in the interval $[7.4\%, 8.6\%]$ in all tests with 168 variables.

The root mean square error is particularly dependant on outliers. The winter model's increase in RMSE might be because of including particularly cold periods in the summer into the winter model.

The mean absolute scaled error reveals that the division of data according to the temperature is better. If MASE equals one, the forecast one hour ahead is as good as persistence, i.e. estimating that the value remains unchanged. Larger values indicate that the estimate is worse. Since MASE has decreased, the forecasting power one hour ahead has improved in both the summer and winter model compared to the winter model using the DST. In the winter models, 168 hour models produce the smallest errors.

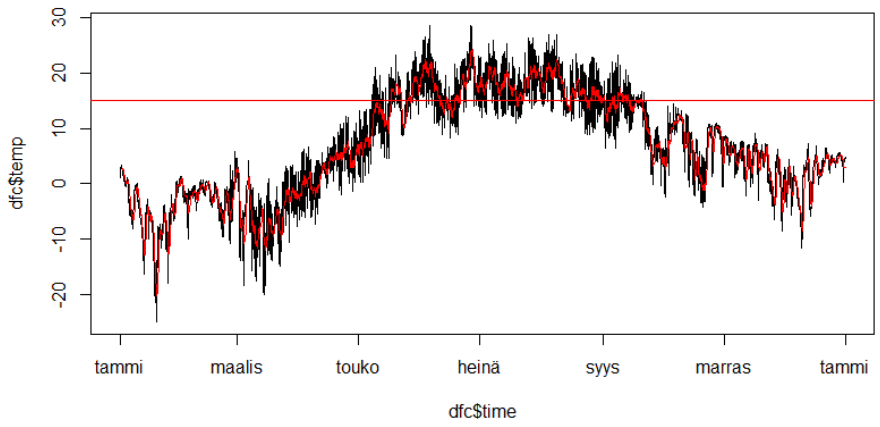


Figure 7: The temperature time series and the linear trend of the temperature. The trend is calculated linearly over 28 hours.

Table 6: The performance of the combined linear regression models for the winter model. All models use 168 hours in the week for the regression. The best values for each error measure are highlighted with bold.

| Error measure | All data | No outliers | Combined regressors |
|---|---|---|---|
| RMSE | 0.9725 | **0.8767** | 0.9043 |
| MAPE | 8.627 | **7.741** | 7.769 |
| MASE | 1.396 | **1.234** | 1.255 |
| MAD | $+2.202, -0.658$ | $+\mathbf{2.05}, -0.8013$ | $+2.09, -\mathbf{0.7624}$ |

## 10.6 Uncertainty of the temperature forecast / long-term forecast

The temperature forecasts become more uncertain the farther we look at the future. As an alternative model to forecasting future values, static linear regression using the volume of the house as an explanatory variable could be used.

A large part of a demand prediction error is based on the inaccuracy of the temperature and weather forecasts[8]. The uncertainty in temperature forecasts is caused through errors in both the weather prediction model formulation and the initial state and conditions of the model [9]. For instance, the weather and climate models can have up to $\mathcal{O}(10^7)$ degrees of freedom [23].

Among other reasons, there are economic grounds for predicting the uncertainty lying in the weather forecasts. With a careful assessment of the forecast uncertainty, the credibility of the weather forecast will increase, and the errors in the demand predictions based on temperature will decrease, yielding more accurate demand forecasts. This will in turn decrease the need to reschedule the heat generation, for example, and thereby make the planning of heat generation more efficient.

There is not much research on the prediction of the uncertainty of the

Table 7: The performance of the combined linear regression models for the summer model. All models use 168 hours in the week for the regression. The best values for each error measure are highlighted with bold.

| Error measure | All data | No outliers | Combined regressors |
|---|---|---|---|
| RMSE | **0.2852** | 0.2874 | 0.2898 |
| MAPE | **9.652** | 9.733 | 9.671 |
| MASE | 1.669 | 1.661 | **1.623** |
| MAD | $+0.3625, -\mathbf{1.337}$ | $+\mathbf{0.3553}, -1.34$ | $+0.3581, -1.35$ |

weather forecast in cases of district heating demand prediction. However, there are studies on how the uncertainty of the temperature forecast impacts electrical load forecasting [8]. Douglas et al. [1998] study the effect of uncertainty on electrical load forecasts using a recursive prediction algorithm based on Bayesian estimation. In their case, the difference in normalized errors in load forecasts can vary as much as 30 to 50 percent between load forecasting with realised temperature and load forecasting with temperature forecasts. On the other hand, their research shows signs that the impact of the errors in temperature forecasts can differ greatly with the annual seasons. In the summer, for instance, the electrical load seems much more temperature-sensitive than in the winter when the normalized errors were estimated to be of nearly the same size.

# 11 Case: consumption in the T3 area

We received consumption data from the T3 area that is Tapiola, Keilaniemi and Otaniemi. The data consists of 796 buildings.

Like before, the data has some problems. Errors in the consumptions had to be corrected as usual. In addition to that, there were errors in the metadata concerning the properties of the buildings. This was a nuisance in closer examination. Of all the buildings 583 were defined to be "good" meaning that their measurements were done hourly and had proper variability.

The data runs from March 2013 to February 2014. It was again divided to three parts according to the daylight saving time. This made the first part much shorter than the rest, only about 1000 hours. The second part was the summer and the third the following winter.

A prediction was made for the third part. Using only the good buildings and predicting for the last week of the third part using the earlier part of the third part gives the following statistics. The model used in this is the static linear regression model with 168 hours.

Table 8: Error measures for forecasts with static linear regression in Tapiola-Keilaniemi-Otaniemi area.

| Measure | Value |
|---|---|
| RMSE | 4.845 |
| MAPE | 8.792 |
| MASE | 3.005 |

We can see from Table 8 that the result is very much like before. Just under

10 % of the variation cannot be explained by this linear regression model.

The supplied data contained a unique problem. Although the area contained 583 good buildings, only 14 of these buildings were labeled according to the given groups. Further, 365 buildings (62%) contained no metadata information about the building type. This creates a huge uncertainty about how to assign group values. Because the varying values model compared favourably to the combined regressors model among the temperature trend sensing model, it is reasonable to use the simpler "no outliers" model to forecast the demand for the area. Using the combined regressors model creates additional uncertainty in determining the groups.

# 12 Conclusions

In this report, two different approaches for modeling the district heat demand are studied. Time-series approach is the more general model allowing the autoregression and moving averages. When these properties are left out we are left with static linear regression. In our model we compensate this by using the hour as an indicator variable.

The result of the modeling was similar in both approaches. Both models had a decent success in forecasting the heat consumption. When choosing between these two the choice then comes down to simplicity and other factor. The basic linear regression method is simple to use and fast to calculate whereas fitting the time series model automatically can take hours on a normal computer. Taking all things into consideration the recommended model is basic static linear regression.

Other things have to be taken care of when using the model for forecasting. In demand forecasts the temperature must be known in advance and temperature forecasts are used for this purpose. The variance in the temperature forecasts increases very quickly and usually most of the uncertainty in heat and electricity forecasts comes from this uncertainty in the temperature. The effect of uncertainty in the temperature forecasts is difficult to remove or diminish because it would require better weather forecast and climate models.

Filtering variables according to the data variability is advantageous in the winter data in Table (6). If the building type data is available, this can also be used in the forecast. The results were similar for either using the grouped filtered variables or the filtered variables data. Regressing the raw data during winter produces poorer results than using the filtered variables. In the summer data in Table (7) the filtering of the data had no effect on the performance. All models performed very similarly.

# References


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

[2] International District Heating Association, editor. *District Heating Handbook - A Design Guide*, volume 1. International District Heating Association, 4 edition, 1983.

[3] V. Bakker, A. Molderink, J. L. Hurink, and G. J. M. Smit. Domestic heat demand prediction using neural networks. In *Systems Engineering, 2008. ICSENG'08. 19th International Conference on*, pages 189–194. IEEE, 2008.

[4] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting, Second Edition*. Springer-Verlag New York, Inc., 2002.

[5] Statistics Canada. Data editing, 2013. [Online; accessed 10-March-2014]; http://www.statcan.gc.ca/edu/power-pouvoir/ch3/editing-edition/5214781-eng.htm.

[6] B. Chramcov. Heat demand forecasting for concrete district heating system. *International Journal of Mathematical Models and Methods in Applied Sciences*, 4(4):231–239, 2010.

[7] E. Dotzauer. Simple model for prediction of loads in district-heating systems. *Applied Energy*, 73(3):277–284, 2002.

[8] A. P. Douglas, A. M. Breipohl, F. N. Lee, and R. Adapa. The impacts of temperature forecast uncertainty on Bayesian load forecasting. *IEEE Transactions on Power Systems*, 13(4):1507–1513, 1998. Institute of Electrical and Electronics Engineers.

[9] M. Ehrendorfer. Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Zeitschrift*, 6:147–183, 1997. Gebrüder Borntraeger.

[10] Fortum. Kaukolämmön tuotanto — fortum.fi, 2014. [Online; accessed 24-February-2014]; http://www.fortum.com/countries/fi/yksityisasiakkaat/kaukolampo/tutustu-kaukolampoon/kaukolammon-tuotanto/pages/default.aspx.

[11] J. Goop. District heating in the Nordic countries - modelling development of present systems to 2050. Master's thesis, Chalmers University of Technology, 2012.

[12] S. Grosswindhager, A. Voigt, and M. Kozek. Online short-term forecast of system heat load in district heating. In *Proceedings of the 31st International Symposium on ForecastingOnline, Prague, Czech Republic*, 2011.

[13] A. J. Heller. Heat-load modelling for large systems. *Applied Energy*, 72(1):371–387, 2002.

[14] K. Holmgren and A. Gebremedhin. Modelling a district heating system: Introduction of waste incineration, policy instruments and co-operation with an industry. *Energy Policy*, 32:1807–1817, 2004.

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

[16] Finnish Energy Industries. District heat on a global scale — energiateollisuus, 2014. [Online; accessed 28-February-2014]; http://energia.fi/en/energy-and-environment/district-heat-and-district-cooling/district-heat-global-scale.

[17] Finnish Energy Industries. Kaukolämmitys — energiateollisuus, 2014. [Online; accessed 4-March-2014].

[18] R. Lahdelma, K. Kontu, T. Fang, and R. Hotakainen. Etäluentatiedon käyttö kaukolämpöjärjestelmän hallinnassa, raportti energiateollisuudelle, 2012.

[19] L. Ljung and T. Glad. *Modeling of Dynamic Systems*. Prentice-Hall, 1994.

[20] J. Luhtala. Espoon kaukolämpöverkon toiminnan tehostaminen mittaustietojen avulla, 2008. Master's thesis.

[21] S. Makridakis and M. Hibon. ARMA models and the Box–Jenkins methodology. *Journal of Forecasting*, 16(3):147–163, 1997. Wiley Online Library.

[22] H. A. Nielsen and H. Madsen. Modelling the heat consumption in district heating systems using a grey-box approach. *Energy and Buildings*, 38(1):63–71, 2006.

[23] T. N. Palmer. Predicting uncertainty in forecasts of weather and climate, 1999. ECMWF - Meteorological Training Course Lecture Series; Online; accessed 20-April-2014; http://www.ecmwf.int/newsevents/training/rcourse_notes/pdf_files/Predicting_uncertainty.pdf.

[24] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with gaussian processes. *The Journal of Machine Learning Research*, 14(1):1175–1179, 2013. JMLR.org.

[25] X. Wang and M. Meng. A hybrid neural network and arima model for energy consumption forcasting. *Journal of Computers*, 7(5), 2012. Academy Publisher.

[26] Wikipedia. Cook's distance — wikipedia, the free encyclopedia, 2014. [Online; accessed 10-March-2014]; http://en.wikipedia.org/w/index.php?title=Cook%27s_distance.

[27] M. Yalcintas and S. Akkurt. Artificial neural networks applications in building energy predictions and a case study for tropical climates. *International journal of energy research*, 29(10):891–901, 2005. Wiley Online Library.

[28] B. Yan and A. M. Malkawi. A bayesian approach for predicting building cooling and heating consumption. In *Proceedings of BS2013: 13th Conference of IBPSA*, pages 3137–3144. 2013, International Building Performance Simulation Association.

[29] X. Yan and X. Gang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific, 2009.

# A Appendix: Self-assessment (in Finnish)

## Projektin kulku

Projekti alkoi hyvin. Tapaamiset Fortumin kanssa saatiin järjestettyä nopeasti, jonka ansiosta projekti lähti hyvin käyntiin.

Periaatteeksi otettiin, että jokainen osa projektista on hyvissä ajoin valmis, ja tämä onnistuikin hyvin. Projektin laajuus oli suurinpiirtein selvillä jo alussa. Ensimmäisen seminaaritapaamisen jälkeen projektia tarkennettiin vielä Fortumin edustajien kanssa.

Vapun alla ryhmään iski hieman hiljaisempi vaihe. Osittain tämä johtui työmäärän vähenemisestä; projekti oli lähes valmis. Osittain tämä myös johtui muiden kurssien työtehtävien kasautumisesta huhtikuun alkuun (koeviikkoon).

Loppuvaiheilla projektiin pohdittiin useita eri lisäyksiä, kuten sääennusteen epävarmuuden ja auringonpaisteen huomiointi. Nämä otettiin jossain määrin huomioon projektissa, mutta vaatisivat molemmat vielä lisäselvityksiä.

## Työnjako

Virallista työnjakoa ei tehty, mutta projektin kuluessa roolit tulivat selväksi. Lasse ja Mika keskittyivät enemmän koodipuoleen ja Joona ja Juulia enemmän kirjallisuuspuoleen. Kirjoitettu koodi ja kirjallisuustutkimus jaettiin, ja kaikki tutustuivat jossain määrin muiden aikaansaannoksiin. Näistä myös keskusteltiin tapaamisissa.

Projektin työmäärä oli virallisesti 115 tuntia rivijäsenellä ja 170 projektipäälliköllä. Nämä arviot pitävät jossain määrin paikkaansa, vaikkakin projektipäälliköllä ei kulunut 55 tuntia enemmän aikaa projektiin kuin rivijäsenellä.

Työmääriä tasattiin jonkin verran jakamalla vastuita opponoinnissa ja esitysten laatimisessa.

## Fortum

Dataa saatiin Fortumin puolelta mukavasti ja ajoissa, joten ennustaminen saatiin toteutettua hyvin.

Kun ennustemalli oli saatu toteutettua, ryhmä kysyi Fortumilta, kuinka he haluaisivat käyttää tätä hyväksi. Se, missä muodossa, jos missään, mallia Fortumilla tulevaisuudessa käytetään, ei ole vielä selvää. Mahdollisimman hyvä koodin ja projektin dokumentointi mahdollistaa mallin helpon jatkokäytön.

## Paranneltavaa

Ryhmä tapasi noin kerran kahdessa viikossa. Tapaamisissa keskityttiin tulevien tehtävien laatimiseen ryhmänjäsenille. Tapaamisissa kerrottiin lyhyesti viime tehtävistä, tosin jälkeenpäin ajatellen ehkäpä liian lyhytsanaisesti. Tietämättömyys muiden jäseneiden puuhista osittain jumitti ryhmänjäsenet samoihin tehtäviin. Näin ollen työtehtävät kasaantuivat eri jäsenille enemmän kuin toisille.

Työtehtävien tasoittamiseksi voisi tehdä enemmän työtä samanaikaisesti samalla paikalla, jolloin myös jäseneiden taidot siirtyvät paremmin eri jäsenien välillä kuin yksintyössä. Tämä tarkoittaa myös parempaa riskienhallintaa ryhmien jäsenten sairastumisten tai muun käyttämättömyyden varalta.

Tiheämpi kokoontumisväli ja pidemmät tapaamiset olisivat voineet myös edesauttaa tiiviimmän ryhmähengen muodostumisessa. Tapaamiset olisivat tällöin saattaneet olla avoimempia, jolloin esimerkiksi uusia tai vaihtoehtoisia näkökulmia olisi esitetty mahdollisesti vielä enemmän ja työtä olisi voitu kehittää vielä pidemmälle.

Projektityön rakennetta ja osia olisi voinut yrittää hahmotella tarkemmin jo projektin alussa. Näin työn suuruutta olisi pystynyt arvioimaan paremmin ja työvaiheiden valmistuessa seuraaviin siirtyminen olisi ollut selkeämpää. Monesti uusia tehtäviä keksittiin sitä mukaa, kun edelliset tulivat valmiiksi. Myös tämä olisi voinut auttaa työtehtävien paremmassa allokoinnissa.