Aalto University
School of Science
Degree programme in Engineering Physics and Mathematics

Identifying Determinants of District Heating Prices for Forecasting

Bachelor's thesis
19.12.2017

Jessica Norrbäck

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Jessica Norrbäck		
Title of thesis Identifying Determinants of District Heating Prices for Forecasting		
Degree programme Engineering Physics and Mathematics		
Major Mathematics and Systems Sciences		**Code of major** SCI3029
Supervisor Prof. Ahti Salo		
Thesis advisor(s) M.Sc. Vilma Virasjoki		
Date 19.12.2017	**Number of pages** 24+3	**Language** English

Abstract

In a cold country like Finland, where heating stands for approximately 67 % of the total energy consumption in buildings, the price of heating has significant impacts on the economy. Because district heating is the most common heating form, standing for nearly 50 % of all heating, the price of district heating is of great interest for many actors. This thesis searches for factors correlating with the district heating prices with the aim of making a forecasting model based on these factors, and comparing different methods for forecasting the prices.

The price of district heating varies between different municipalities depending on features such as the size of the district heating system, investments made, the age of the production facility, the structure of the municipality and the fuels used. Additionally, the district heating price varies in different types of buildings. This thesis studies the average district heating price in apartment buildings for every six months.

When studying data from 2007-2014, we found that fuel prices, the electricity price and the consumer price index are highly correlated with the district heating price. However, the prices of natural gas and woodchips are the only variables that can be included in a multiple linear regression model because of multicollinearity. The ex-post forecast of the multiple linear regression model for the period 2015-2017 is compared to forecasts produced by an ARIMA model and the naïve method. The ARIMA model is determined by the district heating price time series itself and the forecast of the naïve method is obtained by simply assuming that all future values are equal to the last observation.

From these three methods, the naïve method gives the most accurate ex-post forecast and will most likely give quite an accurate short-term forecast. None of these methods are, however, likely to give an accurate long-term forecast. For an accurate long-term forecast more factors need to be taken into account. More components of the production costs could be considered, and the pressure of the competitive situation, i.e. alternative heating forms and technological development driven by the need for sustainable energy production at the heating sector could also be considered. Additionally the impact of global warming could be considered, as the district heating price is dependent of the greatest heating demand during the coldest period of the year.

Keywords district heating, forecasting, multiple linear regression model, ARIMA model, naïve method

Aalto-universitetet
Högskolan för
teknikvetenskaper

Aalto-universitetet, PB 11000, 00076 AALTO
www.aalto.fi

Sammandrag av kandidatarbetet

Författare Jessica Norrbäck

Titel Att identifiera faktorer som påverkar fjärrvärmepriset för prognostisering

Examensprogram Teknisk fysik och matematik

Huvudämne Matematik och systemvetenskaper **Huvudämnets kod** SCI3029

Ansvarig lärare Prof. Ahti Salo

Handledare DI Vilma Virasjoki

Datum 19.12.2017 **Sidantal** 24+3 **Språk** Engelska

Sammandrag

I ett kallt land som Finland, där uppvärmning står för 67 % av byggnaders totala energikonsumtion, har priset på uppvärmning en stor inverkan på landets ekonomi. Då fjärrvärme står för nästan hälften av all uppvärmning är fjärrvärmepriset av stort intresse för många aktörer. I detta arbete studerar vi olika faktorer som påverkar fjärrvärmepriset med målet att göra en prognos på fjärrvärmepriset på basen av dessa faktorer. Dessutom jämför vi olika metoder för att prognostisera utvecklingen av fjärrvärmepriset.

Fjärrvärmepriset varierar i olika kommuner beroende på faktorer så som storleken på fjärrvärmesystemet, investeringar som gjorts, åldern på produktionsanläggningen, kommunens uppbyggnad samt bränslen som använts. Dessutom varierar priset i olika typer av byggnader. I detta arbete används medelpriset på fjärrvärme i höghus för varje halvt år.

Genom att studera data från 2007-2014 finner vi att bränslepriser, elpriset och konsumentprisindexet korrelerar starkt med fjärrvärmepriset. På grund av multikollinearitet kan dock endast priserna på naturgas och träflis inkluderas i en multipel linjär regressionsmodell. Prognosen av den multipla linjära regressionsmodellen för perioden 2015-2017 jämförs med resultatet av en ARIMA-modell och den naiva metoden. I ARIMA-modellen bestäms prognosen av tidigare värden på fjärrvärmepriset och i naiva metoden antas alla kommande värden vara lika som den sista observationen.

Av dessa metoder gav den naiva metoden den noggrannaste prognosen för perioden 2015-2017 och ger troligtvis en relativt noggrann prognos på kort sikt. Ingen av dessa metoder lär dock ge trovärdiga prognoser på lång sikt. För att få en bra prognos på lång sikt borde fler faktorer tas i beaktande. Man kunde till exempel analysera fler komponenter av produktionskostnaderna och ta i beaktande trycket konkurrensen på värmemarknaden sätter på fjärrvärmepriset. Dessutom kunde man beakta inverkan av den globala uppvärmningen på fjärrvärmepriset, eftersom en del av priset bestäms enligt den högsta konsumtionen under den kallaste perioden på året.

Nyckelord fjärrvärme, prognostisering, multipel linjär regressionsmodell, ARIMA-modell, naiv metod

Contents

Abbreviations

a ..... Year

ADF ..... Augmented Dickey-Fuller

ARIMA ..... Autoregressive integrated moving average

CHP ..... Combined heat and power

CPI ..... Consumer price index

DH ..... District heating

ESS ..... Explained sum of squares

k ..... Number of variables in a model

MSE ..... Mean squared error

MSR ..... Mean squared regression

MWh ..... Megawatt hour

n ..... Number of observations

RSS ..... Residual sum of squares

SE ..... Standard error

TSS ..... Total sum of squares

1 Introduction

The Finnish heating market is unregulated and competitive, in the sense that there is no specific legislation concerning the pricing of district heating and that customers are free to choose whichever heating form they want to use. However, there are some restrictions concerning the pricing of district heating and in some areas buildings may by law be obliged to connect to the district heating network (Finlex: Land use and building act 57a§, 1999). Particularly in a cold country like Finland, where 67 % of the energy consumed in residential buildings is consumed by heating sources (Statistics Finland: Asumisen energiakulutus, 2010-2015), it is extremely important to choose the heating method cost-effectively.

As district heating is the most common heating form in Finland standing for nearly 50 % of all heating (Energiamaailma: Kaukolämpö, 2017), the price of district heating has significant impacts on the Finnish economy. This motivates this thesis, which aims to find factors correlating with the price of district heating to be able to forecast the future development of the district heating prices. Furthermore, it is of interest to estimate how the price of district heating develops in the future, since technology development makes alternative heating forms increasingly popular, especially in small residential buildings (Statistics Finland: Rakennus- ja asuntotuotanto, 2017). Therefore, estimating the future development of district heating prices is of great interest for many actors.

This thesis studies how different fuels correlate with the price of district heating and whether other factors, such the increasing popularity of other heating forms and changes in consumer price indices affect the district heating prices. Based on this analysis and extending into time-series models, this thesis compares three different methods and models for making ex-post forecasts of the district heating price. These models may also provide decent forecasts a few years ahead.

Section 2 presents background information on the district heating market in Finland. Section 3 presents the theory of the forecasting methods and motivates the choices. Section 4 describes the processes of constructing the models, and, finally, Sections 5 and 6 present the results, including a comparison of these methods.

2 Background

2.1 District Heating in Finland

District heating (DH) is the most common heating form in Finland. Nearly 50 % of all buildings are heated with DH (Energiamaailma: Kaukolämpö, 2017). In relation to the size of the population, Finland also has the highest DH production in the Nordic countries (Finnish Energy: Kaukolämmön tuotanto, 2017). As a rule of thumb, DH is more economic the more densely built the area is and the larger the houses are (Energiamaailma: Kaukolämpö, 2017), because this decreases network losses. Thus, DH is a commonly used heating form in urban areas and apartment buildings. Over 85 % of all apartment buildings in Finland are heated with DH (Statistics Finland: Asumisen energiakulutus, 2016).

DH can either be produced in combined heat and power (CHP) plants or heat-only boilers. In Finland 70 % of DH is produced in cogeneration with electricity in CHP plants with high efficiency ratings. Heat produced in the combustion plants is then transferred to customers through hot water cycling in the DH network.

Fuels used for DH production vary depending on the area. The share of each fuel used in DH production in 2015 can be seen from Figure 2.1. The most commonly used fuels are wood fuels, coal, peat and natural gas which together stand for 84 % of the total DH production.

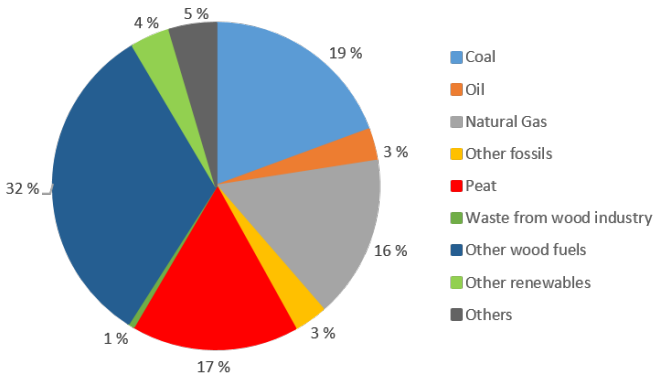

Figure 2.1: Fuels used for DH production in 2015 in Finland. Data source: Statistics Finland, Sähkön ja lämmön tuotanto (Electricity and heat production)

2.2 Price of District Heating

There is no legislation concerning the pricing of DH in Finland, because the market is unregulated. However, competition authorities supervise the pricing and operation of local DH companies. The supervision is based on the antitrust legislation, which states that the misuse of a dominant market position is forbidden. This means that the price level should be reasonable, the pricing should correspond to the costs, and customers have to be treated equally. Additionally, energy taxation regulates the pricing to some degree (Nuorikivi, 2009).

The DH price varies in different municipalities, depending on the size of the DH system, investments made, the age of the production facility, the structure of the municipality and the fuels used. From a customer perspective, a district heating bill consists of a fixed charge (€/a), an energy payment (€/MWh) and value-added tax (24 %). The fixed charge depends on the size of the contracted waterflow and is calculated based on the greatest heating effect during the coldest time. The energy payment is charged according to the consumed energy (Elenia: Kaukolämmön hinnat ja ehdot, 2017).

The DH prices used in this thesis are taken from the material bank of Finnish Energy (Finnish Energy: Kaukolämmön hintatilasto, 2017). Because 70 % of the DH used to warm up residential buildings is consumed by apartment buildings, the price used in this thesis is an average price for DH in apartment buildings. Until January 2010 the standard apartment building was defined in the data of Finnish Energy to have an annual energy consumption of 450 MWh and from then on the standard building was changed to a greater apartment building with an annual energy consumption of 600 MWh.

The price development of district heating is presented in Figure 2.2. It can be seen that the DH price has increased over the years, although the trend has leveled off during the past few years, 2014-2017.

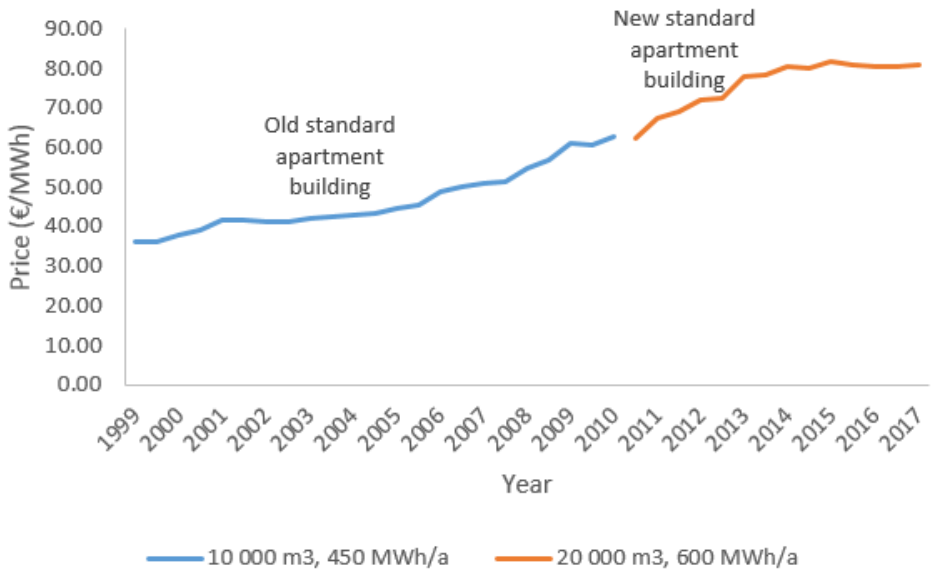

Figure 2.2: Development of district heating prices in a standard apartment building 1999-2017. Annual energy consumption of standard apartment building 1999-2010: 450 MWh, and from 2011 onwards: 600 MWh. Data source: Statistics Finland, Sähkön ja lämmön tuotanto (Electricty and heat production)

2.3 Recent Research

The scientific literature on forecasting DH prices is not extensive. Presumably, this is more interesting for companies and customers in the heating sector than it is in the scientific community. Companies may have done research on the subject, but not published them in order to have an advantage over their competitors. However, there is a report made by Pöyry for Finnish Energy, where the position of DH in the future is discussed (Pöyry, 2011). The report also presents different scenarios for the development of DH prices in 2020 and 2030. The forecasts are made by estimating the changes in production costs, which comprise changes in fuel costs, emission allowances, taxes, investment costs, staff costs and other costs related to maintanance. Some other studies related to the subject focus on explaining DH pricing (Heikkilä, 2015) and comparing costs of DH to other heating methods (Heiskanen, 2013). None of these studies are based on mathematical models for forecasting the future development of DH prices, which is the main research question and contribution of this thesis.

3 Methods

3.1 Multiple Linear Regression Model

A simple linear regression model relates the given observed values of the independent variable X to corresponding values of the dependent variable Y (Amemiya, 1994). It is presumed that for each observation X, the observations on Y will vary in a random fashion. Hence, a random error component, ϵ_i is added to the model, which can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{3.1}$$

where Y_i is a random variable for the ith observation and X_i is nonstochastic and known. The term β_0 is constant and the regression coefficient β_1 measures the change in Y caused by a unit change in X (Pindyck & Rubinfeld, 1981).

The simple linear regression model can be extended to a multiple linear regression model in which the dependent variable Y is a linear function of a series of independent variables $X_1, X_2, ...X_n$ and random error term ϵ_i. The multiple regression model is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ni} + \epsilon_i. \tag{3.2}$$

Here X_{ji}, $j \in [1, n]$, represents the ith observation of the independent variable X_j. The coefficient β_j, $j \in [0, n]$, measures the change in Y caused by a unit change in the variable X_j as in the case of the simple linear regression model, but here under the assumption that all other independent variables are constant (Pindyck & Rubinfeld, 1981).

The assumptions of the multiple regression model according to Pindyck and Rubinfeld are as follows:

i The model follows the specification given by Eq. (3.2).

ii The independent variables X are nonstochastic and there are no exact linear relationships between two or more of them.

iii (a) The expected value of the error term is zero for all observations, i.e.,

$$E(\epsilon_i) = 0, \forall i \in [1, n]. \tag{3.3}$$

(b) The errors corresponding to different observations are uncorrelated:

$$\text{corr}(\epsilon_i, \epsilon_j) = 0, \forall i, j \in [1, n], i \neq j. \tag{3.4}$$

(c) The error variable follows a normal distribution:

$$\epsilon \sim N(0, \sigma). \tag{3.5}$$

Having several potential independent variables, it is important to scrupulously choose which ones to use. Redundant predictors should be removed, since it is desirable to explain the data in the most simple way (Armstrong, 2002). Unnecessary predictors only add noise to the estimation of other quantities and introduce additional degrees of freedom. The term degrees of freedom refers to the number of unconstrained observations (Pindyck & Rubinfeld, 1981). When testing hypotheses on the model statistically, confidence intervals depend on the degrees of freedom, which means that unnecessary variables affect the confidence intervals and may lead to different statistical results.

3.1.1 Parameter Estimation by Least Squares Method

The β parameter estimation is typically done by using the least squares method, in which the parameter estimates minimize the residual sum of squares (RSS) (Amemiya, 1994):

$$RSS = \sum \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}' \hat{\boldsymbol{\epsilon}} \tag{3.6}$$

where

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} \tag{3.7}$$

and

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{3.8}$$

Here, $\hat{\boldsymbol{\epsilon}}$ represents the $N \times 1$ vector of regression residuals, and $\hat{\mathbf{Y}}$ represents the $N \times 1$ vector of fitted values for $\mathbf{Y}$. Equations (3.7) and (3.8) can be

inserted into Eq. (3.6), to obtain

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y'Y} - \hat{\boldsymbol{\beta}}'\mathbf{X'Y} - \mathbf{Y'X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X'X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y'Y} - \mathbf{2}\hat{\boldsymbol{\beta}}'\mathbf{X'Y} + \hat{\boldsymbol{\beta}}'\mathbf{X'X}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.9}$$

The terms $\hat{\boldsymbol{\beta}}'\mathbf{X'Y}and\mathbf{Y'X}\hat{\boldsymbol{\beta}}$ are both scalars and equal to each other, hence the last step. The least-square estimators are determined by minimizing RSS

$$\frac{\partial \text{RSS}}{\hat{\boldsymbol{\beta}}} = -2\mathbf{X'Y} + 2\mathbf{X'X}\hat{\boldsymbol{\beta}} = 0$$

$$\implies \hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}(\mathbf{X'Y}). \tag{3.10}$$

The second-order condition requires that the matrix $\mathbf{X'X}$ is positive definite. In case $\mathbf{X}$ has a full rank this requirement is fulfilled. The assumption that $\mathbf{X}$ has rank k implies that $\mathbf{X'X}$ is nonsingular and has an inverse (Amemiya, 1985).

3.1.2 R^2 and adjusted R^2 ($\bar{R}^2$)

R^2 is often used as an informal measure for goodness-of-fit to the multiple regression model. It is defined as the explained sum of squares (ESS) divided by the total sum of squares (TSS)

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum \hat{\epsilon}_i^2}{\sum(Y_i - \bar{Y})^2}, \tag{3.11}$$

where Y_i is the original data value, $\hat{Y}$ is the value from the model and $\bar{Y}$ is the mean. R^2 measures the portion of variation in Y which is explained by the independent variables of the multiple regression model. The main concerns related to the use of R^2 are that all the statistical results follow from the initial assumption that the model is correct and that the addition of more independent variables cannot lower R^2; they usually raise it. A solution is to study the variances instead of variations so that the goodness-of-fit on the number of independent variables in the model is eliminated. The adjusted

R^2, or $\bar{R}^2$ aims to elimainate the weakness of R^2 (Amemiya, 1985). It is defined as

$$\bar{R}^2 = 1 - \frac{\text{Var}(\hat{\epsilon})}{\text{Var}(Y)}. \tag{3.12}$$

The sample variances of $\hat{\epsilon}$ and Y are calculated as

$$\text{Var}(\hat{\epsilon}) = s^2 = \frac{\sum \hat{\epsilon}_j^2}{n-k} \tag{3.13}$$

and

$$\text{Var}(Y) = \frac{\sum (Y_i - \bar{Y})^2}{n-1}, \tag{3.14}$$

where n is the number of observations and k is the number of independent variables. Thus, the relationship between R^2 and $\bar{R}^2$ is

$$\bar{R}^2 = R^2 \frac{n-1}{n-k}. \tag{3.15}$$

Even though RSS may decrease as more explanatory variables are added to the model, the residual variance does not necessarily do that. This makes $\bar{R}^2$ a more desirable and comparable goodness-of-fit measure, although it still does not solve all difficulties. The decision to add a new variable should still largely be based on a priori theoretical considerations.

3.1.3 F-Test of Overall Significance

Hypothesis testing can be helpful in taking decisions about the addition of new variables in the model, because they test the statistical significance of the addition. The initial step in testing hypotheses is stating the null hypothesis, H_0, and the alternative hypothesis, H_1. The null hypothesis states that something is not significant and that there is no relationship between some factors. If it is significantly unlikely that the data occured with the null hypothesis being true, the null hypothesis is rejected and the alternative hypothesis is accepted (Amemiya, 1994).

The test for overall significance of the regression model is carried out using variances. It tests whether a significant linear relationship exists between

the dependent variable and at least one of the independent variables. The hypotheses for the F-test of overall significance are as follows

$H_0 : \beta_1 = \beta_2 = ... = \beta_n = 0$
$H_1 : \beta_j \neq 0$, for at least one j.

The test statistic F_0 is calculated by using the mean squared regression (MSR) and the mean squared error (MSE) (Moy, Chen & Kao, 2015)

$$F_0 = \frac{MSR}{MSE}. \tag{3.16}$$

The term MSR is obtained by dividing the the explained sum of squares ESS by the degrees of freedom for the model

$$MSM = \frac{ESS}{k-1}, \tag{3.17}$$

where k is the number of independent variables in the model. The term ESS is obtained by

$$ESS = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2. \tag{3.18}$$

The term MSE is calculated by dividing the RSS with the degrees of freedom for the error

$$MSE = \frac{RSS}{n-k}, \tag{3.19}$$

where n is the number of observations. The null hypothesis cannot be rejected, if the calculated F_0 statstic is within the confidence interval

$$F_0 < f_{\alpha,k,n-(k+1)}, \tag{3.20}$$

where α is the confidence level, k is the number of indepenent variables and n is the number of observations. The confidence interval can be calculated using the F-table or statistical software. If the null hypothesis can be rejected, at least one of the coefficients β_j is significant. The F-test does not, however, tell which coefficient it is.

3.1.4 T-test of Individual Coefficients

T-tests are carried out to test the significance of individual coefficients in the multiple linear regression model. The hypotheses for an individual coefficient, β_j, in the t-test are

$H_0 : \beta_j = 0$
$H_1 : \beta_j \neq 0.$

The test statistic, T_0, based on the t-distribution is

$$T_0 = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}, \tag{3.21}$$

where $SE(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$, the modeled value of β. The standard errors for the β_i coefficients are obtained by taking square roots from the diagonal of the variance-covariance matrix. The variance-covariance matrix, D, contains the variances and covariances of the independent variables X_j (Kennedy, 2003)

$$D = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Var}(X_n) \end{bmatrix}. \tag{3.22}$$

The standard error of $\hat{\beta}_j$ is thus obtained as follows

$$SE(\hat{\beta}_j) = \sqrt{D_{jj}}. \tag{3.23}$$

The null hypothesis cannot be rejected if the test statistic lies within the acceptance region for the null hypothesis

$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}, \tag{3.24}$$

where α is the confidence interval and n is the number of observations. The confidence interval is calculated using the t-distribution. If the null hypothesis is rejected, it can be concluded that β_j is significant at the chosen confidence level.

3.1.5 Partial F-test of Subsets of Regression Coefficients

The partial F-test is a more general form of the t-test, as it checks the significance of including one or several regression coefficients to the linear regression model. As the addition of more variables increases the regression sum of squares, RSS, the test studies the increase of RSS, called extra sum of squares. In the test the vector $\boldsymbol{\beta}$ is split into two vectors, $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$. The coefficient vector $\boldsymbol{\theta_1}$ contains the first $(n+1-r)\beta$ regression coefficients and the vector $\boldsymbol{\theta_2}$ contains the rest of the β regression coefficients. The hypotheses test the significance of adding the regression coefficients in $\boldsymbol{\theta_2}$ to the model with the regression coefficients from $\boldsymbol{\theta_1}$.

$H_0 : \boldsymbol{\theta_2} = 0$
$H_1 : \boldsymbol{\theta_2} \neq 0$

The test statistic, F_0, follows a F-distribution and is calculated as

$$F_0 = \frac{RSS(\boldsymbol{\theta_2} \mid \boldsymbol{\theta_1})/r}{MSE}, \tag{3.25}$$

where $RSS(\boldsymbol{\theta_2} \mid \boldsymbol{\theta_1})$ is the increase in the regression sum of squares that the inclusion of the regression coefficients in $\boldsymbol{\theta_2}$ contribute with. The parameter r is the number of elements in vector θ_2. The null hypothesis cannot be rejected, if

$$F_0 < f_{\alpha,r,n-(k+1)}. \tag{3.26}$$

The confidence interval is calculated using the F-distribution. If the null hypothesis is rejected, at least one of the coefficients in $\boldsymbol{\theta_2}$ contributes significantly to the regression model.

3.1.6 Motivation for the Use of Multiple Linear Regression Model

A multiple linear regression model is a good choice, when the data is linearly correlated with other explanatory variables. As mentioned in Section 2.2, the DH price is influenced by many factors that are known. Therefore, the DH price can be forecasted based on other variables for which data is available, assuming that there is a linear correlation between the DH price and the variables and that these explanatory variables do not exhibit multicollinearity. However, forecasts cannot be extrapolated into the future, unless the future values of the explanatory variables are known. Therefore, the independent

variables must be lagged by one period or more. As the DH price is set at a specific time for a fixed time, it is most likely that the independent variables in fact are lagged. Hence, the multiple linear regression model is likely to be suitable for developing a short-term forecast of the DH price.

3.2 ARIMA models

To be able to construct a time series model, the time series has to be stationary. For a time series to be classified as stationary, the mean, variance, autocorrelation and other statistical properties ought to be constant over time. However, nonstationary series can be transformed into stationary by, e.g,. differencing (Pindyck & Rubinfeld, 1981).

The series y_t is homogenous nonstationary of order d, if

$$w_t = \Delta^d y_t = y_t - y_{t-d} \tag{3.27}$$

fills the criterions of a stationary series. After the series y_t has been differenced to produce the stationary series w_t, time series models can be applied on w_t.

Homogenous nonstationary time series can be modeled as ARIMA(p, d, q) processes, where p comes from the autoregressive part of the model, d comes from the order of differencing and q comes from the moving average part of the model. In autoregressive models the current observation, y_t, is generated by a linear combination of previous observations of the variable going back p periods. Thus, an autorregressive model of order p, denoted AR(p), is defined as

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p} + \delta + \epsilon_t, \tag{3.28}$$

where δ is a constant term and ϵ_t is white noise. An autoregressive model is similar to a multiple regression model, but it uses its own lagged values as explanatory variables (Hyndman & Athanasopoulos, 2012).

In moving average models past forecast errors going back q periods are used to generate each observation y_t. A moving average process of order q, denoted MA(q) is defined as:

$$y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - ... - \theta_q \epsilon_{t-q} \tag{3.29}$$

where μ is the mean, ϵ_t is a random disturbance. It should be noted that the mean $E(y_t) = \mu$ is independent of time.

Additionally, each random disturbance ϵ_t is assumed to be generated by the same white noise process (Pindyck & Rubinfeld, 1981), which means that

$$E(\epsilon_t) = 0 \tag{3.30}$$

$$E(\epsilon_t^2) = \sigma_\epsilon^2 \tag{3.31}$$

$$E(\epsilon_t \epsilon_{t-k}) = 0, k \neq 0. \tag{3.32}$$

An ARIMA (AutoRegressive Integrated Moving Average) model is obtained by combining differencing with autoregression and a moving average model (Hyndman & Athanasopoulos, 2012). The backward shift operator is useful when working with time series, as it denotes lags through

$$By_t = y_{t-1}. \tag{3.33}$$

We can write the equation for the ARIMA(p, d, q) process using the backward shift operator

$$\Phi(B)\Delta^d y_t = \delta + \theta(B)\epsilon_t, \tag{3.34}$$

where

$$\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p \tag{3.35}$$

and

$$\theta(B) = \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q. \tag{3.36}$$

The term $\Phi(B)$ is called the autoregressive operator and $\theta(B)$ is called the moving average operator (Pindyck & Rubinfeld, 1981).

The practical problem in modeling an ARIMA process is choosing the most appropriate values for p, d and q. Since the specification of the ARIMA model is not the main focus of this thesis, we will specify the ARIMA model using a built-in function in R.

3.2.1 Motivation for the Use of ARIMA model

We know that the DH price is dependent of external factors, but we do not know exactly how the DH price is determined. By using a time series model we can, instead, forecast the DH price by its own lagged values. It is interesting to see how the results differ when the forecast is made on external versus internal factors. Studies show that when the expected changes are small, the information about relationships are of little value. Thus, in short-term forecasts extrapolation methods often perform as well as econometric methods (Armstrong, 2002). This motivates the use of a time series model as a means of comparison for the regression model. We consider the ARIMA model to be a suitable time series model, because any homogenous nonstationary process can be modeled as an ARIMA(p, d, q) model (Pindyck & Rubinfeld, 1981).

3.3 Naïve Method

The naïve method is an extremely simple forecasting method, that provides a good benchmark for other forecasting models. It should be noted, though, that the method is only applicable for time series data. The principle of the naïve method, is that all future values are set equal to the last observation. This can be notated as:

$$y_{t+h|t} = y_t \tag{3.37}$$

where t is the time of the last observation and h is the amount of periods we want to forecast ahead. As simple as it is, the method still works quite well for economic and financial time series, which are hard to predict due to their unregular patterns (Hyndman & Athanasopoulos, 2012).

Naïve methods often give adequate short-term forecasts, when data has been stable for a long time (Armstrong, 2002). The increase in the price of district heating has stagnated in 2014 and thereafter stayed on quite a constant level (Fig 2.2). Therefore, the naïve method is a suitable benchmark to compare the forecasts from the multiple regression model and ARIMA-model with.

3.4 Research and Modeling Software

The data in this thesis is analyzed using Excel and R. The correlations are determined using Excel's $CORREL()$-function, which calculates the corre-

lation between two data sets x and y by the formula

$$\text{Correl}(x, y) = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})\Sigma(y - \bar{y})}}. \tag{3.38}$$

The forecasting models are constructed using R. The multiple linear regression model is made with the $lm()$-function, using the least squares method (R Documentation, ND). The ARIMA model is constructed using the $auto.arima()$-function, which uses a variation of the Hyndman and Khandakar algorithm to obtain an ARIMA model (Hyndman & Athanasopoulos, 2012) taking into account the given arguments. For more information, see Hyndman & Athanaopoulos, Chapter 8.7.

4 Construction of Models

4.1 Potential Explanatory Variables of Multiple Linear Regression Model

The main factors which influence the DH price in different municipalities are the size of the DH plant, investments made, the age of the production facility, the structure of the municipality and the fuels used. When focusing on the average price of DH, the only factors that can reasonably be taken into account from the list above are the fuel prices. Additionally, the consumer price index (CPI) can be considered as a potential explanatory variable, because changes in the CPI most likely influence the DH price.

Even if the DH price corresponds to the production costs in accordance with the antitrust legislation, the prices of alternative heating forms need to be taken into account according to the principles of competitive pricing (Berends, 2004) when setting the DH price. Figure 4.1 shows the energy consumed by different energy sources in apartment buildings in 2015. After DH, electricity is the second most popular heating form. The other sources make up a significantly smaller share and are thus not considered as potential explanatory variables.

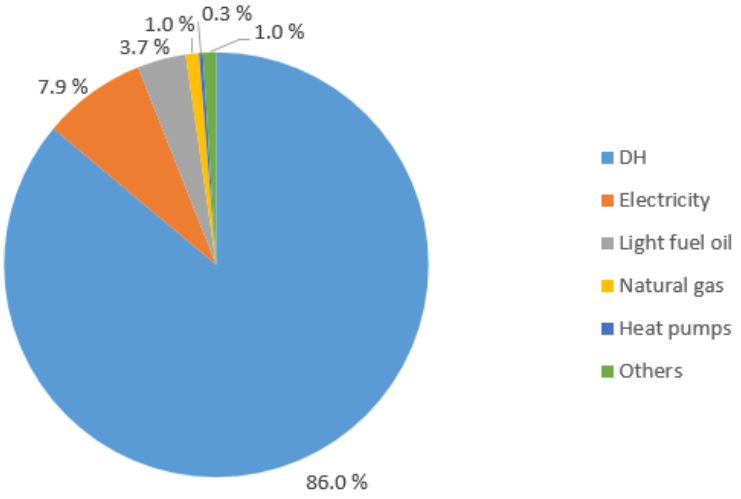

Figure 4.1: Energy consumed by different energy sources in apartment buildings 2015. Data source: Statistics Finland, Asumisen energiakulutus energialähteittäin (Energy consumption of housing by energy source)

There may, however, be some problems related to adding the CPI and electricity price as explanatory variables to the model. According to the assumptions of the multiple linear regression model in Section 3.1, there can be no exact linear relationship between two or more of the independent variables. The existance of highly correlated explanatory variables (i.e. multicollinearity) may cause problems. Table 1 shows the correlations between the potential explanatory variables discussed above (CPI and electricity price) and the other potential variables in the model, the fuel prices. Changes in the consumer price index are most likely reflected in the fuel prices as well, as indicated by the high correlation between the CPI and fuel prices (Table 1). Adding the elctricity price to the model may be problematic, since 70 % of the district heat is produced in cogeneration with electricty. Thus, there is also a risk of multicollinearity as indicated by Table 1. Therefore, neither one is added to the model to avoid multicollinearity.

A multiple regression model is thus created with the four most popular fuels as potential explanatory variables: woodchips, coal, peat and natural gas. In section 3.1.6 we assumed that these as independent variables are lagged. By studying the correlations between average DH prices and independent variables at different time lags, it can be noted that the independent variables are indeed lagged (Table 2). For the domestic fuels, woodchips and peat, data is only available for every quarter. The correlation between woodchips

and DH prices is highest when the woodchip prices are lagged 16 months. Respectively, the correlation between peat and DH prices is highest when the peat prices are lagged 13 months. Taken into account that both fuels are domestic and that the lags are quite small at one month too, it seems like a coincidence that the lags are highest at 13 and 16 months. We will therefore use a one months lag for domestic fuels, as information should be disseminated quickly within the country. For the imported fuels, coal and natural gas, it can be assumed that the lag is slightly higher than for the domestic fuels. The correlation between coal and DH prices is highest when both are lagged 12 months, which seems reasonable. Thus, we will choose a 12 months lag for the imported fuels.

Table 1: Correlations without lag representing the potential for multicollinearity.

	DH price	Woodchips	Peat	Coal	Natural gas
CPI	0.977	0.878	0.958	0.891	0.947
Electricity price	0.908	0.898	0.784	0.788	0.783

4.2 Choice of Explanatory Variables for Multiple Linear Regression Model

The multiple regression model should be kept as simple as possible, because the inclusion of insignificant independent variables does not improve the performance of the model. Table 3 shows a short summary of different models that we test. Woodchips and natural gas prices have the highest correlation with the DH prices. Thus, one of these could be taken as the first explanatory variable to the model. As Table 3 shows, any second variable cannot be added to the model so that both variables are statistically significant, meaning that the t-statistics from the t-test on individual regression coefficients do not exceed $t_{0.025,11} = 2.201$. Neither do the F-statistics of the partial F-test exceed $f_{0.05,1,10} = 4.9646$. Thus, Table 3 indicates, that model_{1b}, a simple model with only the natural gas price as an independent variable is the best, because all regression coefficients in it are statistically significant and it has a higher adjusted R^2 than the other model with all regression coefficients statistically significant (model_{1a}).

Table 2: Correlation with DH price at different lags. The lags of the bolded values are the ones chosen to the model.

Lag (months)	Woodchips	Peat	Coal	Natural gas
0			0.806	0.945
1	**0.939**	**0.896**	0.744	0.898
2			0.729	0.907
3			0.773	0.923
4	0.908	0.847	0.837	0.934
5			0.800	0.941
6			0.806	0.945
7	0.896	0.883	0.801	0.937
8			0.777	0.925
9			0.829	0.946
10	0.867	0.851	0.829	0.951
11			0.851	0.919
12			**0.860**	**0.957**
13	0.898	0.928	0.818	0.938
14			0.791	0.938
15			0.832	0.943
16	0.985	0.919	0.829	0.942
17			0.844	0.937
18			0.853	0.929

However, natural gas only accounts for 16 % of the DH production, wherefore forecasts of the DH price based solely on the natural gas prices may be weak. For this reason, a second independent variable is added to the model. The adjusted R^2 is highest for the model with woodchips and natural gas as explanatory variables (model_{2c}). What makes model_{2c} even more attractive, is that is has both a domestic and a imported fuel, which means that factors influencing the fuel prices in different places are taken into account in this model. Therefore, model_{2c} is chosen to be the final multiple regression model. The summary of model_{2c} is presented in Table 4.

Table 3: Alternative models. The bolded independent variables are statistically significant.

Name	Independent variables	Adjusted R-squared
model_{1a}	**Woodchips**	0.870
model_{1b}	**Natural gas**	0.909
model_{2a}	**Woodchips**, Peat	0.895
model_{2b}	**Woodchips**, Coal	0.882
model_{2c}	**Natural gas**, Woodchips	0.926
model_{2d}	**Natural gas**, Peat	0.925
model_{2e}	**Natural gas**, Coal	0.915

Table 4: Summary of model_{2c} (std = standard error, dof = degrees of freedom)

	Estimate	Std. Error	t value
Intercept	39.0868	2.5310	15.443
Woodchips	0.8727	0.4633	1.884
Natural gas	0.5276	0.1729	3.052
Residual standard error:	2.24 on 10 dof		
Adjusted R^2:	0.9259		

4.3 Construction of ARIMA(p, d, q) model

The parameters p, d and q of the ARIMA model are chosen using built-in functions in R. When simply using the *auto.arima*()-function, the model obtained is a ARIMA(1,1,0) model with a drift. However, it gives a very incorrect forecast. If drift is not allowed, we obtain an ARIMA(0,1,0) model, which is called a random walk model (Fan & Yao, 2017) and is given by

$$y_t = y_{t-1} + \epsilon_t. \tag{4.1}$$

This is basically the same as the naïve method, but with a random error term. It is, however, quite an uninteresting model, because we already are using the naïve method. We will therefore try to construct some other ARIMA model.

One possibility is to determine the order of differencing using an augmented Dickey-Fuller (ADF) test, which has a null hypothesis that a unit root is present in the time series data and an alternative hypothesis that the data is stationary. According to the ADF test the DH data should be differenced twice to become stationary. When using the *auto.arima*()-function again, but requiring the order of differencing to be 2, an ARIMA(1,2,0) is obtained. This model is chosen as the final ARIMA model, since it gives a more interesting result than the random walk model.

5 Results

5.1 Comparison of Models

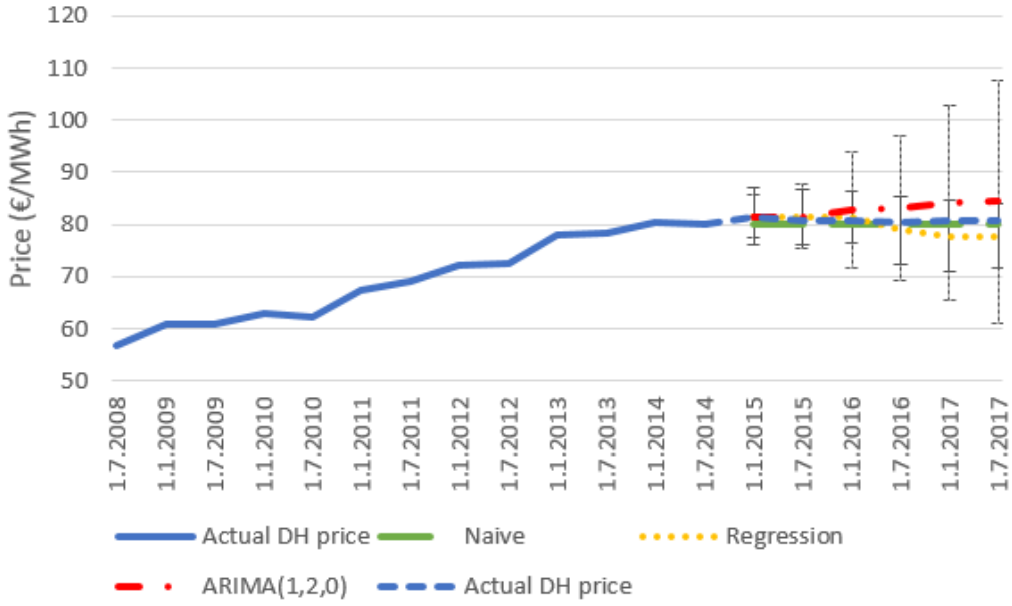

Figure 5.1: Fitted values of multiple regression model, ARIMA model and naive model and confidence intervals for regression and ARIMA models. Confidence interval for regression model is solid, confidence interval for ARIMA model is dashed.

Ex-post forecasts of the DH price using different models are presented in Figure 5.1. The term "ex-post" refers to forecasts made over a period, when the actual data is known, in this case 1.1.2015-1.7.2017. The forecast from

the multiple regression model is slightly above the actual DH price until 1.1.2016, and thereafter starts decreasing, contrary to the actual DH price. The forecast from the ARIMA model continues to grow in the same trend as it has done earlier, while the actual price has stayed on quite a constant level. The result obtained using the naïve method is a constant 80.06 €/MWh.

The 95 % confidence intervals for both models are obtained using R. The actual values of the DH price during the period from 1.1.2015 to 1.7.2017 are within the confidence intervals for both models. The confidence interval for the ARIMA model grows over time, while the confidence interval of the regression model is practically constant.

Table 5 presents the relative errors for the forecasted values from the different models. As the DH price has stayed on quite a constant level, the errors from the naïve model are relatively small, all under -1.85 %. The errors from the two other modes are not significant either. The error of the regression model is at its minimum in the first period and increasing up to -3.82% by the sixth period. The error of the ARIMA model is also smallest in the first period and increases to 4.41% by the sixth period. All these models can be seen as reasonably accurate, since their errors are relatively small.

For the period of 1.1.2015-1.7.2017, the naïve method gives the best forecast, since the errors are the smallest. This is due to the fact that the price of DH has not changed significantly over the examined period. However, it should be noted that the multiple linear regression model and ARIMA model are quite accurate the first two periods.

Table 5: Errors of ex-post forecasts for all models.

Date	Multiple regression	ARIMA	Naïve
1.1.2015	0.00%	0.02%	-1.85%
1.7.2015	0.81%	0.90%	-0.95%
1.1.2016	1.13%	2.65%	-0.73%
1.7.2016	-1.64%	3.37%	-0.30%
1.1.2017	-3.71%	4.05%	-0.90%
1.7.2017	-3.82%	4.41%	-0.99%

5.2 Models as Means of Forecasting

Because the DH price depends on external factors, it is most likely to be determined through some kind of a function with multiple variables. This motivates the use of a multiple regression model. But because we do not know the exact variables, coefficients and form of the function, there is plenty of uncertainty associated with the multiple linear regression model we developed.

The ex-post forecasts were made for the period 1.1.2015-1.7.2017 based on data from 1.7.2008-1.7.2014. Since the DH price has been nearly constant from 1.1.2014 onwards, the naïve method was quite accurate. The two other models were very accurate the first period, but their absolute errors increased over time.

For a short-term forecast ahead (ex-ante) the naïve method is likely to provide the most accurate forecast, since the DH price has been quite constant over the last eight periods. Thus, it will most likely be quite accurate at least one or two periods ahead. It is, however, very unlikely that the DH price will stay on the same level very long, wherefore the naïve method may not be the best model to produce a long-term forecast, i.e., more than a few periods ahead.

It is difficult to say how the regression model will perform in the future, because the lags for the domestic fuels are quite small and, thus, the future values of the independent variables are not known. If the lags were greater, it would be possible to extrapolate the model further into the future, without the need of forecasting independent variables. There is, however, a lot of uncertainty related to forecasting based on forecasts. Therefore, the multiple regression model is not very suitable for making a long-term forecast. A forecast could possibly be made one period ahead using the regression model, but considering that in our example, the absolute error of the regression model has increased over time, it would probably not be very accurate.

The absolute error in the models we compare is highest for the ARIMA model, wherefore it will perhaps not give the most accurate short-term forecast. The model is, however, quite easy to extrapolate into the future, but taken into account that the error of the model increased over time, a long-term forecast obtained by the ARIMA will neither be likely to perform very well. However, if the ARIMA model actually were used for forecasting, it could be updated regularly as new data became available, giving more accurate forecasts.

6 Conclusions

The aim of this thesis was to find factors that correlate with the DH price and to compare ex-post forecasts of different forecasting methods. We found that fuel prices, the electricity price and the consumer price index strongly correlated with the DH price, but an accurate forecasting model could not be made based on all these factors. We concluded that we can get quite accurate short-term forecasts by using the naïve method, but for an accurate long-term forecast of the DH price, there is a need for better methods.

There are many factors that need to be taken into account when making a long-term forecast for the DH price. One alternative is to make a more thorough analysis on the different components of the DH price, as done in Pöyry's report. The long-term forecasts made by Pöyry in 2010 (Pöyry, 2011) are to date not very accurate, but the prospect of estimating changes in production costs, emission allowances, taxes, and all other costs seems promising.

Another fact to be taken into account in forecasting the long-term development of the DH price is that the more the DH price rises, the less competitive DH will become. Already in some apartment buildings, the DH systems are switched to heat pumps, because heating is cheaper with heat pumps and the investment usually pays itself back in 15 years. This puts pressure on the DH price, because a further increase in the DH price may leed to a customer loss in areas, in which customers are free to choose whichever heating form they want. In order to keep DH a competitive heating form, producers of DH should by any means try to press down the price.

In addition, the climate change may impact the DH price, since part of the DH price is determined by the greatest heating effect during the coldest time. It is forecasted that by 2060, winters in Northern Europe will be 2-7°C warmer (Ilmatieteen laitos, ND). This could lead to lower DH prices when the maximum heating demand during the coldest period decreases. However, it is not known how the entire heating system will have changed by that time.

References

Amemiya. (1985), Advanced econometrics. Basil Blackwell Ltd., Oxford, UK.

Amemiya. (1994), Introduction to statistics and econometrics. Harvard University Press, Cambridge, USA.

Armstrong. (2002), Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers, New York, USA.

Berends. (2004), Price & profit: the essential guide to product & service pricing and profit forecasting. Berends & Associates, Ontario, USA.

Elenia: Kaukolämmön hinnat ja ehdot. Accessed: 30.8.2017. Available at: http://www.elenia.fi/lampo_ja_kaasu/hinnoittelu_kaukolampo

Energiamaailma: Kaukolämpö. Accessed: 30.8.2017. Available at: http://energiamaailma.fi/mista-virtaa/kaukolampo/

Finnish Energy: Kaukolämmön hintatilasto. Accessed: 30.8.2017. Available at: https://energia.fi/ajankohtaista_ja_materiaalipankki/materiaalipankki/kaukolammon_hintatilasto.html

Finnish Energy: Kaukolämmön tuotanto. Accessed: 30.8.2017. Available at: https://energia.fi/perustietoa_energia-alasta/energiantuotanto/kaukolammon_tuotanto

Fan & Yao. (2017), The elements of financial econometrics. Cambridge Univrsity Press, Cambridge, USA.

Finlex: Land use and building act (1999). Accessed: 30.8.2017. Available at: http://www.finlex.fi/fi/laki/ajantasa/1999/19990132#L7P57a

Heikkilä. (2011), Kaukolämmön hinnoittelurakenteet. University of applied sciences thesis. Accessed: 31.10.2017. Available at: https://www.theseus.fi/bitstream/handle/10024/32684/Insinoorityo%20TimoHeikkila%2015.4.2011.pdf?sequence=1

Heiskanen. (2013), Kaukolämpö- ja maalämpöjärjestelmän kustannusvertailu pientalon lämmitysjärjestelmänä. University of applied sciensces thesis. Accessed: 31.10.2017. Available at: http://www.theseus.fi/bitstream/handle/10024/66804/Heiskanen_Mirkka.pdf?sequence=1&isAllowed=y

Hyndman. (2012), Constants and ARIMA models in R. Accessed: 23.9.2017. Available at: https://robjhyndman.com/hyndsight/arimaconstants/

Hyndman & Athanasopoulos. (2012), Forecasting: principles and practice. Open-access online textbook. Accessed: 6.7.2017. Available at: https://www.otexts.org/fpp

Ilmatieteen laitos: Ilmastonmuutos. Accessed: 6.7.2017. Available at: http://ilmatieteenlaitos.fi/ilmastonmuutoskysymyksia

Kennedy. (2003), A guide to econometrics. The MIT Press, Massachusetts, USA.

Moy, Chen & Kao. Study guide for statistics for business and financial economics. Springer, Switzerland.

Nuorikivi. (2009), Kaukolämmön hinnoittelumallit. Accessed: 5.7.2017. Available at: http://188.117.57.25/sites/default/files/kaukolammon_hinnoittelumallit_2009.pdf

Pindyck & Rubinfeld. (1981), Econometric Models and Economic Forecasts. USA.

R Documentation, adf-test. Accessed: 4.11.2017. Available at: https://www.rdocumentation.org/packages/aTSA/versions/3.1.2/topics/adf.test

R Documentation, Fitting linear models. Accessed: 20.9.2017. Available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/lm.html

Statistics Finland: Asumisen energiakulutus. Accessed: 30.8.2017. Available at: http://www.stat.fi/til/asen/2015/asen_2015_2016-11-18_tau_001_fi.html

Statistics Finland: Asumisen energiakulutus energialähteittäin 2010-2015. Accessed: 13.9.2017. Available at: http://www.le.ac.uk/users/dsgp1/COURSES/ELOMET/LECTURE3.PDF

Statistics Finland: Asumisen energiakulutus vuosina 2010-2015. Accessed: 31.10.2017. Available at: http://www.stat.fi/til/asen/2015/asen_2015_2016-11-18_tau_001_fi.html

Statistics Finland: Rakennus- ja asuntotuotanto. Accessed: 30.8.2017. Available at: http://www.stat.fi/til/ras/2016/09/ras_2016_09_2016-11-25_kat_001_fi.html

Työ- ja elinkeinoministeriö, Energiateollisuus ry: Kaukolämmön asema Suomen energiajärjestelmässätulevaisuudessa. Accessed:

30.8.2017. Available at: http://www2.energia.fi/kaukolampo/klasemaloppuraportti52a14971.pdf