Heating Demand Forecasting with Multiple Regression: Model Setup and Case Study

Karlis Baltputnis
Institute of Power Engineering
Riga Technical University
Riga, Latvia
karlis.baltputnis@rtu.lv

Roman Petrichenko
Institute of Power Engineering
Riga Technical University
Riga, Latvia
romans.petricenko@rtu.lv

Dmitry Sobolevsky
Institute of Power Engineering
Riga Technical University
Riga, Latvia
zdmi333@inbox.lv

***Abstract*—Accurate demand forecasting in district heating networks is an essential and imperative task in the everyday operation of both, the network itself and the heating energy suppliers. Multiple regression is one of the possible approaches to solving the forecasting problem with sufficient accuracy and little computational effort. This paper presents a polynomial regression model and offers several additions for its further improvement. It is found that grouping the model residuals by hour-of-day allows notably reducing the forecast error. The value of other modifications and the optimum size of the training set can vary over time, thus an automatic model parameter selection before each new forecast is advised.**

Keywords—district heating, forecasting, regression, automation, cogeneration

I. Introduction

Combined heat and power (CHP) plants are an important source of heating energy in district heating (DH) networks around the world. These plants are characterized by high efficiency due to the electricity produced alongside heat; this allows them to have lesser fuel consumption and smaller carbon footprint compared to when the two types of energy are produced separately [1].

The primary task of CHP plants connected to DH networks is supplying the heating energy, whereas electricity is often treated as a byproduct. However, for worthwhile participation in electricity markets, more certainty is necessary regarding the heating demand. Although there are measures which allow more flexibility in the production of electrical energy by somewhat untying it from the heat demand, i.e., heat storage tanks, peak water boilers, improved cycling operation [1], [2], proper scheduling and operational control of CHP plants nevertheless heavily relies on heating demand forecasts.

The forecasts necessary for CHP plant operation can be categorized in two groups depending on the prediction horizon: operational (subhourly to several hours-ahead) for near real time adjustments of the production output and day-ahead for unit scheduling and preparation of bids to a wholesale market [3].

A great variety of methods for DH heating demand forecasting can be found in recent literature, for instance, feed-forward neural networks [3]–[8], support vector machines [4], [6]–[11], random trees regression [4], [7], [12], ridge regression [9], [13], random forest [9] deep learning [13], extreme learning machines [5], [12], genetic programming [5], [6], [8] and even linear regression [4], [7], [13]–[15]. The methods vary in complexity and therefore also presumably in their time of execution, unfortunately, few authors provide comparable data on computational time.

However, several studies suggest that the simpler regression models can provide similar [4] or even better [13], [15] forecasting accuracy than machine learning approaches. This study aims to expand the literature on heating demand forecasting in DH networks with regression models by employing a very straightforward and effective polynomial approach and exploring the benefits of improving it with three types of modifications – decoupling hot water (HW) consumption from space heating demand, taking into account the residuals of the fitted regression model and filtering the input and output series. Furthermore, as currently the size of the overall historical dataset to be used for forecasting is seldom tested in the literature, this paper provides insights into identifying a reasonable look-back horizon for forecasting heating demand with regression methods.

The remainder of the paper is structured as follows. Section II provides description of the forecasting techniques employed and the dataset used. Section III lays out the case studies and results. Finally, the last section concludes the paper.

II. Methodology

A. The Underlying Regression Model

In general, regression allows us to approximate a mathematical relationship between two or more variables if their values are known in a number of points. Eq. (1) illustrates a multiple regression model (a polynomial), where the right-side terms can be both independent variables and functions of independent variables.

$$y_i = a_0 + \sum_{n=1}^{k} a_n \cdot x_i^n + \varepsilon_i, \tag{1}$$

where y_i – dependent variable at point i, x_i – independent variable at point i, n – power of each term, k – power of the last term (i.e., order of the polynomial), ε_i – error term at point i, a_0 – the intercept term, a_n – coefficient for the corresponding function of the independent variable .

In heat load forecasting, the dependent variable is, of course, the heating demand itself, whereas various different factors can serve as the independent variables or predictors. All of the reviewed studies agree on outdoor temperature as the most important predictor in heating demand forecasting. However, some additional parameters have been employed as well. For instance, papers [4]–[6], [8], [13] also consider time-lagged heating demand values. Time factors like hour-of-day, day-of-week and day-of-year are also sometimes used for forecasting [4], [7], [13]. If the forecasting

algorithm is intended to be applied for a smaller supply area (i.e., one substation as opposed to the whole DH network), the physical parameters of the DH substation can be used as well [4]. Study [9] stands out in that it considers dew point as a predictor variable. Finally, solar irradiation [11] and wind speed [11], [14] is employed as well, however, the impact of wind on the forecasts can vary a lot across different buildings and, on a larger scale (i.e. the whole DH network), can even out [14].

However, the inclusion of multiple input variables in predictive models can negatively affect their interpretability and predictive power. Additionally, it can reduce their generalization capability [16]. Consequently, in this paper, we focus on outdoor temperature as the most influential predictor [9], [11], [16].

Thus, we can formulate the function for heating demand forecasting. If we assume a third order polynomial relationship, the model can be expressed as in (2).

$$\hat{Q}_t = a_0 + a_1 \cdot \hat{T}_t + a_2 \cdot \hat{T}_t^2 + a_3 \cdot \hat{T}_t^3, \quad (2)$$

where $\hat{Q}_t$ – the forecasted heating demand (output) at hour t, $\hat{T}_t$ – the temperature forecast (input) at hour t, a_0, a_1, a_2, a_3 – polynomial coefficients (model parameters).

The model parameters are obtained by solving a least squares problem where the sum of the model residuals is minimized. The solution can be expressed in matrix formulation as:

$$\begin{bmatrix} a_3 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} = (V' \times V)^{-1} \times (V' \times Y), \quad (3)$$

where Y – a vector of dependent variable values (in our case, heating demand), V – the Vandermonde matrix for the independent variable (outdoor temperature).

B. Modifications

Ref. [14] identified HW as an important social component in the heating demand curve. In this paper, we will test if a polynomial regression model can provide higher accuracy for our testing datasets if it is supplemented by an additional component for HW handling. While the recorded heating demand data does not discriminate between space heating and HW, the energy spent on water heating has to be identified implicitly. For this, we assume that most of the consumption during summer is specifically for HW and thus we can obtain the social component by averaging the recorded points over the corresponding time period. Afterwards, the approximate HW hourly profile can be subtracted from the model training dataset and added back to the forecast as a temperature-independent component.

Another addition to the polynomial regression model tested in this paper is in handling the residuals of the fit. It is done by assigning information on hour-of-day to the error term ε_i from (1) for each element i. The residuals are then grouped by the respective hours of the day and, thus, an average error profile for a full day is obtained. This profile is subtracted from the forecast in an expectation to decrease the inaccuracy.

$$\hat{Q}_t = a_0 + a_1 \cdot \hat{T}_t + a_2 \cdot \hat{T}_t^2 + a_3 \cdot \hat{T}_t^3 - \overline{\varepsilon_t}, \quad (4)$$

where $\overline{\varepsilon_t}$ is the average error of the model in the training dataset for each particular hour of the day t (1..24, since we aim to use the forecasting model for day-ahead scheduling of CHP plants).

A third modification to be tested is applying a smoothening filter by calculating the weighted double-sided moving average of different lengths. This can be applied to either the model training data (historical heating demand, dubbed input hereinafter), the forecasted demand series (output), both or neither. The formulae for smoothening the output is provided in (5), but the same principle would be used for the input.

$$\begin{aligned} \hat{y}_t^{\text{fl.}} &= \hat{y}_t \text{ for } t \in \{1,24\}; \\ \hat{y}_t^{\text{fl.}} &= \frac{\hat{y}_{t-1} + \hat{y}_t + \hat{y}_{t+1}}{3} \text{ for } t \in \{2,23\}; \\ \hat{y}_t^{\text{fl.}} &= \frac{0.5 \cdot \left(\hat{y}_{t-2} + \hat{y}_{t+2}\right) + 0.7 \cdot \left(\hat{y}_{t-1} + \hat{y}_{t+1}\right) + \hat{y}_t}{3.4} \text{ for } t \in [3,22] \end{aligned} \quad (5)$$

Finally, the size of the training dataset is also a model feature to be determined. We test 24 different setups from one week (7 days) to 168 days (roughly 6 months).

C. Setup of the Simulations

The performance of the forecasting model is evaluated using mean absolute percentage error (MAPE).

$$MAPE = \frac{1}{\mathrm{m}} \sum_{i=1}^{\mathrm{m}} \left| \frac{Q_i - \hat{Q}_i}{Q_i} \right|, \quad (6)$$

where Q_i – the actual heating demand at point i, m – total number of points in the forecast.

In order to simulate the intended application of the forecasting model (i.e., in day-ahead scheduling), the model is utilized in a rolling horizon manner – it moves iteratively though each day in the testing dataset and performs a 24-hour prediction; the MAPE for the day is calculated and saved; afterwards, the current day is added to the training dataset and a forecast for the next 24-hour period is performed. Once MAPEs for each of the days in the testing dataset are obtained, they are averaged out to find the mean error for the whole set. In order to test the effect of the features described in section II.B, the model runs are carried out a total of 384 times.

Finally, another approach to using the described multiple regression model features is tested, whereupon the model selects its features (HW exclusion on or off, model residual subtraction on or off, type of data filtering and size of the

training dataset) before each 24-hour period by exhaustively enumerating the possible model configurations on data from the previous day and selecting the best performer for the following day.

D. Data Set

For validation of the proposed multiple regression model and its modifications, we use historical data from Riga, Latvia, particularly, the largest DH network in the right bank of the city (annual consumption about 2.4 TWh). The dataset used in this study contains heating demand and outdoor temperature records from Jan. 1, 2015 to Oct. 31, 2016.

The forecasting simulation experiments will be run twice in this dataset. Case study 1 will forecast demand for days from Jan. 1, 2016 to Mar. 1, 2016 (91 days), whereas Case study 2 will perform forecasts from Oct. 15 to Oct. 31, 2016 (17 days). The former represents the middle of the heating season, while the latter – the beginning. It should be noted that only period when the heating season is assumed to be in full effect is included in the regression model (i.e., period from April to mid-October is excluded). The hourly forecasts are performed in a sliding horizon manner with 24-hour increments, but, for comparison purposes, only the final MAPE for each case study (and each model setup) will be presented.

In this study, we used the recorded temperatures as predictors instead of temperature forecasts. The reason for this was to isolate the effects from the regression model configuration, since the external temperature forecasts would introduce inaccuracies which do not depend from our model setup. An evaluation of the impact of temperature forecast imperfections is already offered in [3] and [12].

III. Results

A. Selection of Polynomial Order

Multiple regression with polynomials up to the 5th order was tested. In Case Study 1, the 2nd order polynomial proved to provide the best accuracy with a MAPE of 5.98 %, while the 3rd order was close behind with 6.07 %. In Case Study 2, both of these parameters again showed very similar results albeit with the 3rd order prevailing (at 4.64 % vs 4.68 %). The performance of each of the five models depending on the training set size is summarized in Fig. 1 (for both case studies combined). Evidently, higher order models tend to overfit if the training set is small, but the more the training set is increased, the more similar the performance of the various polynomials becomes.

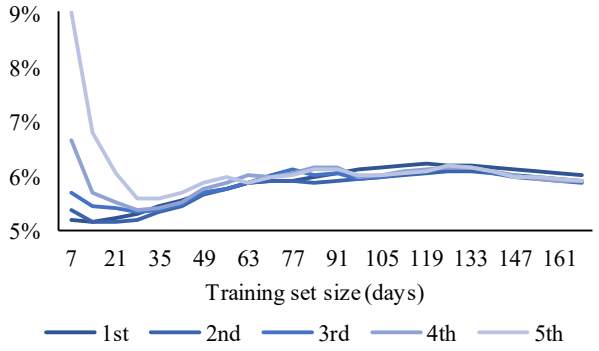

Fig. 1. MAPE per different polynomial orders and look-back horizon

In the remainder of the paper we will focus on the 3rd order model and not vary this parameter further as it is not the main subject of this study.

B. Effect of Modifications and Look-Back Horizon

Results from the various modified model runs for Case Study 1 are summarized in Table I. These are the MAPE values averaged over the different look-back horizons. Fig. 2 and Fig. 3 presents the disaggregated results with the impact of the training set size observable.

Evidently, in Case Study 1, the impact of time series filtering is very small – in the range of 0.05 percentage points. The best result is achieved if only the output is filtered. The inclusion of a social component for HW handling has not improved the model performance. The explicit correction of hour-of-day specific model residuals, however, has more notably improved the forecasting performance, i.e., by 0.27 percentage points. In terms of training set size, the best results were achieved with a look-back horizon of 28 days (5.34 %). The results are similarly accurate for the range from 14 to 49 days, but with larger training sets the MAPE quickly increases.

TABLE I. Results of Case Study 1 (MAPEs)

Filtering		**Error correction**		**HW component**	
no filtering	5.92 %	***included***	5,78 %	***included***	5.92 %
filtered I	5.96 %	***not included***	6.05 %	***not included***	5.92 %
filtered O	5.86 %				
filtered I/O	5.91 %				

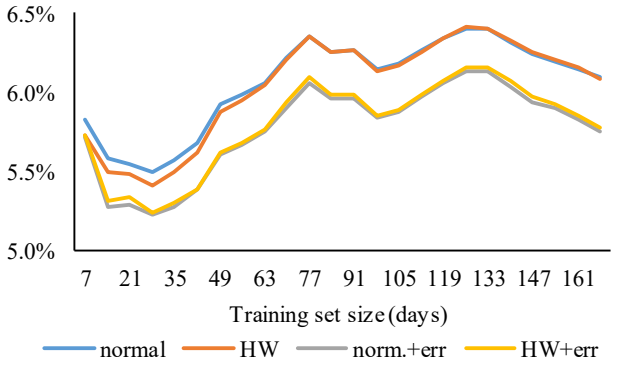

Fig. 2. MAPE per model modification and training set size (Case Study 1)

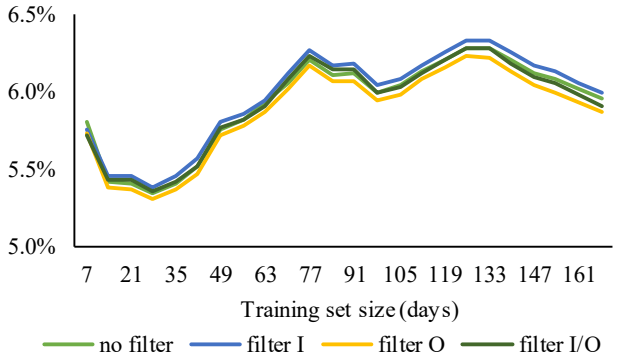

Fig. 3. MAPE per filtering type and training set size (Case Study 1)

TABLE II. RESULTS OF CASE STUDY 2 (MAPES)

Filtering		Error correction		HW component	
no filtering	4.40 %	***included***	4.18 %	***included***	4.36 %
filtered I	4.37 %	***not included***	4.59 %	***not included***	4.42 %
filtered O	4.38 %				
filtered I/O	4.40 %				

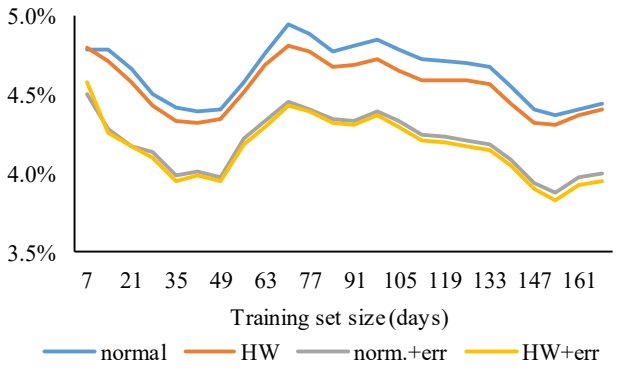

Fig. 4. MAPE per model modification and training set size (Case Study 1)

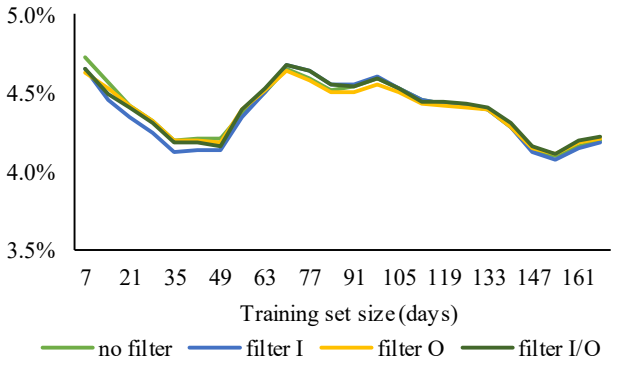

Fig. 5. MAPE per filtering type and training set size (Case Study 1)

The results of the Case Study 2 are similarly summarized in Table II, Fig. 4 and Fig. 5.

The MAPE of the Case Study 2 is overall notably smaller. This signifies a season-specific reason for the inaccuracies. Similarly to the previous case, filtering does little to affect the results (range of only 0.03 percentage points) with input filtering providing the smallest error (4.37 %). In this case, however, HW component has slightly improved the results (by 0.06 percentage points). The residual component once again provides the most notable accuracy improvements (by 0.41 percentage points). Unlike in Case Study 1, here the best results are obtained by a 154 day look-back horizon (4.09 %), but there is also a range with low error estimates in the 28 to 49 days period.

C. *Automatic Feature Selection*

One of the main takeaways of the previous subsection is the difficulty to draw strong conclusions on the best forecasting model setup, since if applied to different portions of the dataset, the modified features offer varying advantages and disadvantages. Due to this uncertainty and the low computational effort the regression model requires (the 91 day testing dataset for Case Study 1 handled in less than a second), an automatic model setup is proposed and tested.

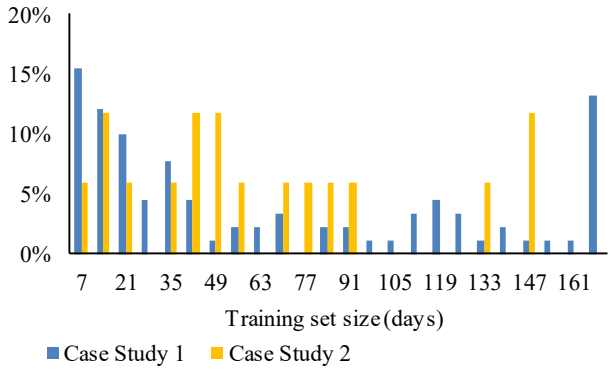

Fig. 6. Frequency of look-back horizon used in both case studies

If before each day-ahead forecast the model can self-select those parameters which would have provided the best forecast for the previous day, the overall MAPE for the testing dataset decreases more significantly – 5.19 % in Case Study 1 and 4.27 % in Case Study 2, a 0.73 and 0.12 percentage point improvement versus the average MAPE in the previous simulations respectively.

The automatic forecasting algorithm chose to employ the HW component for 30.77 % of days in Case Study 1 and 35.29 % of days in Case Study 2. The usage of the residual handling feature was more active – 72.53 % and 70.59 % respectively. Filtering wise, in both cases, I/O filtering was used most often (35.16 %, 35.29 %) while solely input filtering was the least used (13.19 %, 17.65 %).

Fig. 6 summarizes the frequency of training dataset size selected in both case studies. While generally this model feature has varied a lot, a tendency to cluster towards smaller look-back horizons can be observed.

IV. CONCLUSION

Multiple (polynomial) regression has proven to be an effective tool for heating demand forecasting. One of its main strengths is the negligible computational time it takes to perform forecasts without loosing much in terms of accuracy.

Furthermore, the forecasting model can be improved by certain modifications, the most promising of which has turned out to be subtraction of the model residuals averaged over hour-of-day. While other modifications (HW component and time series filtration) did not produce a consistently beneficial effect over the whole dataset, there were days when their inclusion aided in improving the accuracy. Thus, a model which automatically selects the features the forecasting program should consider before each daily forecast is advisable. Additionally, it should consider automatic selection of the training set size, since the optimum look-back horizon tends to vary during the heating season.

While the model tested in this paper already provides forecasts with adequate accuracy, further improvements are necessary. One promising venue for future work lies in further integration of the multiple regression model with an ANN-based forecasting approach. Another important research topic concerns forecasting the heat energy demand in the DH network specifically during the very beginning and end of the heating season, when space heating is gradually connected/disconnected by building managers.

ACKNOWLEDGMENT

This work has been supported by the European Regional Development Fund within the Activity 1.1.1.2 "Post-doctoral Research Aid" of the Specific Aid Objective 1.1.1 "To increase the research and innovative capacity of scientific institutions of Latvia and the ability to attract external financing, investing in human resources and infrastructure" of the Operational Programme "Growth and Employment" (No. 1.1.1.2/VIAA/1/16/021).

This work has been supported by Latvian Council of Science, project: Management and Operation of an Intelligent Power System (I-POWER) (No. lzp-2018/1-0066).

REFERENCES

[1] M. A. Sayegh *et al.*, "Trends of European research and development in district heating technologies," *Renew. Sustain. Energy Rev.*, vol. 68, pp. 1183–1192, Feb. 2017.

[2] O. Linkevics, P. Ivanova, and M. Balodis, "Electricity Market Liberalisation and Flexibility of Conventional Generation to Balance Intermittent Renewable Energy – Is It Possible to Stay Competitive?," *Latv. J. Phys. Tech. Sci.*, vol. 53, no. 6, pp. 47–56, Dec. 2016.

[3] K. Baltputnis, R. Petrichenko, and A. Sauhats, "ANN-based city heat demand forecast," in *2017 IEEE Manchester PowerTech*, 2017, pp. 1–6.

[4] S. Idowu, S. Saguna, C. Åhlund, and O. Schelén, "Applied machine learning: Forecasting heat load in district heating system," *Energy Build.*, vol. 133, pp. 478–488, Dec. 2016.

[5] S. Sajjadi *et al.*, "Extreme learning machine for prediction of heat load in district heating systems," *Energy Build.*, vol. 122, pp. 222–227, Jun. 2016.

[6] M. Protić *et al.*, "Forecasting of consumers heat load in district heating systems using the support vector machine with a discrete wavelet transform algorithm," *Energy*, vol. 87, pp. 343–351, 2015.

[7] D. Geysen, O. De Somer, C. Johansson, J. Brage, and D. Vanhoudt, "Operational thermal load forecasting in district heating networks using machine learning and expert advice," *Energy Build.*, vol. 162, pp. 144–153, Mar. 2018.

[8] E. T. Al-Shammari *et al.*, "Prediction of heat load in district heating systems by Support Vector Machine with Firefly searching algorithm," *Energy*, vol. 95, pp. 266–273, 2016.

[9] S. Bandyopadhyay, J. Hazra, and S. Kalyanaraman, "A machine learning based heating and cooling load forecasting approach for DHC networks," in *2018 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2018, pp. 1–5.

[10] M. Protić *et al.*, "Appraisal of soft computing methods for short term consumers' heat load prediction in district heating systems," *Energy*, vol. 82, pp. 697–704, 2015.

[11] M. Dahl, A. Brun, O. Kirsebom, and G. Andresen, "Improving Short-Term Heat Load Forecasts with Calendar and Holiday Data," *Energies*, vol. 11, no. 7, p. 1678, Jun. 2018.

[12] C. Johansson, M. Bergkvist, D. Geysen, O. De Somer, N. Lavesson, and D. Vanhoudt, "Operational Demand Forecasting In District Heating Systems Using Ensembles Of Online Machine Learning Algorithms," *Energy Procedia*, vol. 116, pp. 208–216, Jun. 2017.

[13] G. Suryanarayana, J. Lago, D. Geysen, P. Aleksiejuk, and C. Johansson, "Thermal load forecasting in district heating networks using deep learning and advanced feature selection methods," *Energy*, vol. 157, pp. 141–149, Aug. 2018.

[14] T. Fang and R. Lahdelma, "Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system," *Appl. Energy*, vol. 179, pp. 544–552, 2016.

[15] R. Petrichenko, K. Baltputnis, A. Sauhats, and D. Sobolevsky, "District heating demand short-term forecasting," in *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe)*, 2017, pp. 1–5.

[16] D. Petković, M. Protić, S. Shamshirband, S. Akib, M. Raos, and D. Marković, "Evaluation of the most influential parameters of heat load in district heating systems," *Energy Build.*, vol. 104, pp. 264–274, 2015.