The 21st International Conference on Thermal Science and Engineering of Serbia
SimTerm 2024
Niš, Serbia, Oct 22-25 2024
doi: 10.5937/SimTerm24257S

Predicting Heat Demand of Residential Buildings with Lag and Time Variables

Mirko M. Stojiljković[a], Marko G. Ignjatović[b], Goran D. Vučković[c], Vladan S. Jovanović[d]

[a]*University of Niš, Faculty of Mechanical Engineering in Niš, Niš, RS, mirko.stojiljkovic@masfak.ni.ac.rs*
[b]*University of Niš, Faculty of Mechanical Engineering in Niš, Niš, RS, marko.ignjatovic@masfak.ni.ac.rs*
[c]*University of Niš, Faculty of Mechanical Engineering in Niš, Niš, RS, goran.vuckovic@masfak.ni.ac.rs*
[d]*University of Niš, Faculty of Mechanical Engineering in Niš, Niš, RS, vladan.jovanovic@masfak.ni.ac.rs*

Abstract: Building heat demand depends on various parameters: the properties of the materials, geometry, climate conditions, occupancy patterns, usage habits, etc. Short-term forecasts of the heat demand can be based on different subsets of these parameters, depending on the application and available data. Recently, black-box models that apply machine learning methods became widely used to predict the heat demand. This paper is a part of a broader research effort and investigates the time-series prediction properties of a simple model based on the Random Forest regression that uses only lag and time-related variables as inputs. The time resolution is one hour. The lag variables are the heat demand one hour before, 24 hours before, and 25 hours before the time of prediction. The time-related variables are the hour of day, day of week, and month. The model has the coefficient of determination of over 0.99, root-mean-square error 38.2 kWh, and mean absolute error 19.1 kWh. The most important predictor is the 24-hours lag. Large errors occur mainly early in the morning and late in the evening, when the heat demand changes have high values and the setpoint temperatures are being modified.

Keywords: building simulation, multi-story apartment building, random forest, time-series forecasting.

1. Introduction

Modern and smart design and operation of energy-efficient buildings, that have complex energy supply systems, require accurate forecasts of the building energy performance. Traditionally, this is achieved with building simulations, which are built around knowledge-based models. Recently, the models based on knowledge and physical laws are often replaced or supplemented with data-driven or surrogate models, which are rely on machine learning algorithms.

Data-driven models are often desirable because they can be very precise and do not require full detailed information on the variables that the building heating demand depends on the properties of the materials, geometry, climate conditions, occupancy patterns, usage habits, etc. Gathering such information if often time- or resource-consuming. These models can perform sufficiently well with different subsets of the afore-mentioned parameters, depending on the application and available data. In addition, they can also be faster and more accurate [1].

However, sufficiently precise surrogate models require input data of appropriate quantity and quality. The inputs are usually obtained either with real measured data [2] or synthetic data obtained with building simulations [3].

Surrogate models can be applied in the design phase [4], in the case of building retrofit analysis [5], or for the operation of building energy systems [2]. They can be used to predict either annual, monthly, daily, hourly, or sub-hourly energy consumption data. A particularly important category of machine learning models applied to the heating and cooling demand are the models that predict time series data.

Li et al. [3] applied several supervised machine learning models to create building meta-models that are supposed to replace computationally expensive building simulation models and predict time series. They used hourly data obtained with EnergyPlus [6] simulations of an office building. The input variables were heating and cooling demand lags, physical properties of the building elements, climate data, hour, and day type. The most important features for the prediction of the heating energy consumption appeared to be the day type and the first lag of the outdoor dry-bulb temperature, while for the cooling energy consumption, the outdoor dry-bulb temperature and solar heat gain coefficient were the most significant.

Liang et al. [7] applied the K-Nearest Neighbors method for surrogate modeling buildings in the pre-design stage, as well as predicting the heating and cooling load of retail, hotel, and office buildings. They used input variables related to the building geometry, indoor temperature, occupancy, lightning, and equipment.

Zhang et al. [8] presented an approach for day-ahead prediction of the cooling load of an office building by reorganizing historical data based on daily time and weather similarity and then applying machine learning models such as Linear Regression, Random Forest, Support Vector Regression, and Gaussian Process Regression.

Chou and Ngo [9] proposed a hybrid model for day-ahead forecasts of time-series data from smart grids, based on weekly sliding window. They combined a seasonal autoregressive integrated moving average model (SARIMA) with support vector regression (SVR) and used metaheuristic optimization firefly algorithm (FA) to tune the hyper-parameters. The combination of SVR and FA was applied to overcome the limitations of the linear nature of SARIMA. They analyzed the results obtained with the following inputs: the energy consumption lags, day type (weekday or weekend), hour of the day, and outdoor temperature. The conclusion is that the scenario with the first three features yields the best results. Ikeda and Nagai [10] presented a different hybrid of machine learning and metahuristic optimization. They optimized operation of a complex building energy system by applying deep neural networks on cooling towers and connected components, and metaheuristics to the rest of the system.

In addition to energy demand, machine learning models have been used to predict the electricity production of a photovoltaic system [11, 12], solar irradiation [13], temperature [14], etc.

This paper examines the predictive performance of a time-series model that forecasts hourly heating demand of a group of multi-story residential buildings. The model uses Random Forest method [15] for regression. The forecast horizon is one-step ahead. This part of a broader research is focused on the model that uses only time-related and heating-demand-lag variables, and attempts to analyze the performance in the cases when the weather-related data is not collected, as well as the data on building geometry, physical properties, occupancy and other relevant patterns, etc.

The buildings have the total floor area of 27000 m². Their behavior is simulated with EnergyPlus to obtain synthetic data for model training and testing. The buildings are divided into 524 thermal zones. The time step of the simulation outputs is one hour.

Prediction of the heating demand of these buildings is of interest for the district heating plant that provides the heat to them.

2. Methodology

As mentioned above, this paper presents the model that predicts the heating demand of a group of residential buildings. The demand is forecast in a time-series manner, one hour ahead. The prediction model is based on Random Forest (RF) regression method. RF is chosen because of its wide range of applicability, relatively fast and computationally non-intensive training procedure, and small number of hyperparameters.

The analyzed period is eight years (from 2014 to 2021). The first six years (from 2014 to 2019) are used to train the model and the last two years (2020 and 2021) remained for testing.

The performance indicators of the regression model are:

- Coefficient of determination (R^2)
- Root mean square error (RMSE)
- Mean absolute error (MAE)

The coefficient of determination, defined in Eq. (1), is dimensionless and shows the ratio of the variance of the predicted variable (in this case the heating demand) explained by the independent variables in the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where n is the number of samples, y_i is a true value of the predicted variable for the sample point i, which is obtained with the building simulation in this case, $\hat{y}_i$ is the predicted value, their difference $y_i - \hat{y}_i$ is the error of prediction, and $\bar{y}$ is the arithmetic mean of all true values, as shown in Eq. (2):

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

RMSE, defined in Eq. (3), corresponds to the quadratic error and has the same unit as the predicted variable, i.e. kWh in this case:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

MAE, defined in Eq. (4), is also expressed in kWh and shows the absolute error:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

The model contains only two kinds of variables: (1) time and (2) lag variables. Selection of the lag variables that should be included in the model is performed based on the values of the Pearson correlation coefficient. Figure 1 illustrates these values for the lag variables up to the 100th shift.

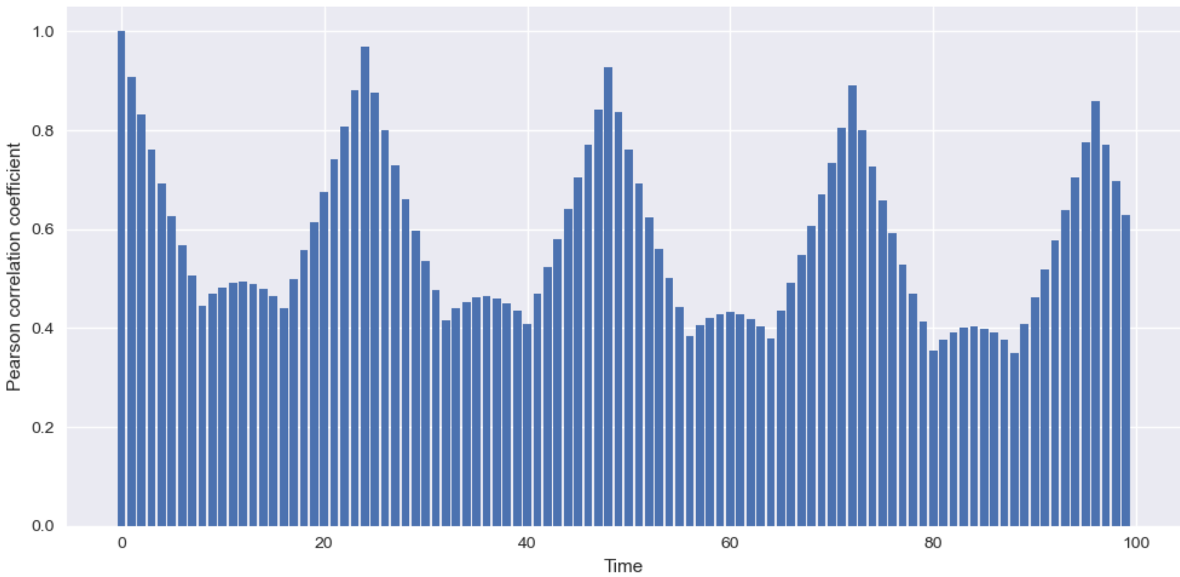

Figure 1. Pearson correlation coefficients for the heating demand and its lags

In addition to the expected drop after the first lag, the correlation coefficient increases again as the time approaches 24 h and again for 48 h, 72 h, and so on. This raises an assumption that the 24th lag gives more information than most of the earlier lags. For that reason, this lag is included in the list of input variables. Since the first lag is naturally involved in the prediction, it seemed as an interesting idea to include the 25th lag as well, so that some relations between the values of the heating demand that occurred 24 and 25 hours in the past can be learned and applied to the demand value from the past hour.

Another interesting detail is noticed for the autocorrelation between the heating demand and its first lag. The Pearson correlation coefficient is very close to one for all the hours of the day except for the morning and evening hours. These are the hours when the changes in the heating demand are usually very large. Figure 2 shows this pattern.

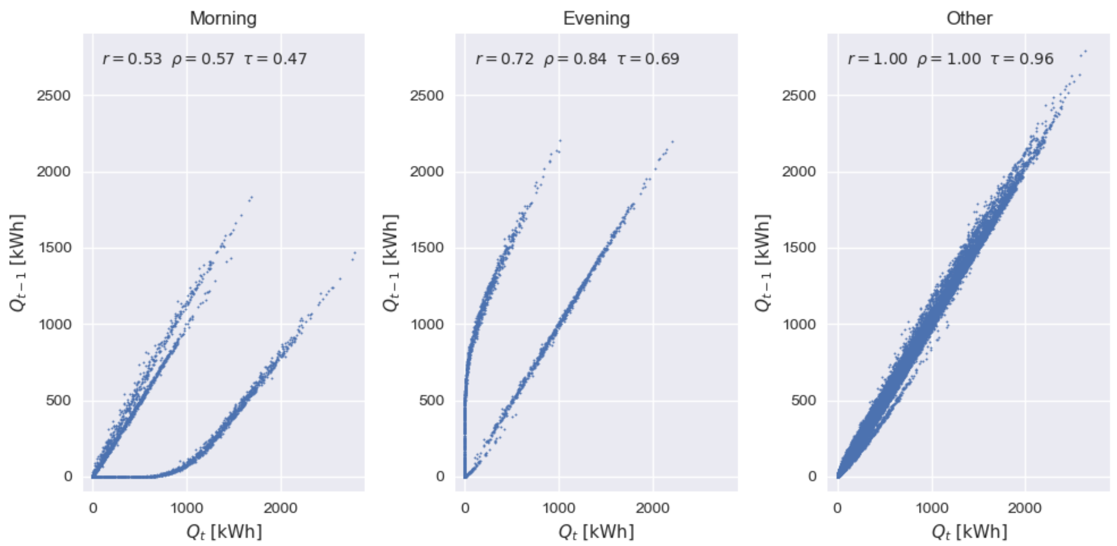

Figure 2. Autocorrelation according to the time of day

Finally, eight features are selected:

- Time variables: hour of day, day of week, day of month, month, and year.
- Lag variables: the heating demand one, 24, and 25 hours before the time of prediction.

The hyperparameters of RF model are defined based on experience and include:

- Number of decision trees: 200,
- Quality of the split: Friedman Mean Squared Error
- Minimal sample size needed to split an internal node: 20
- Number of considered features during the search for the best split: 6

RF model is implemented in Python programming language using Scikit-learn library [16].

3. Results and discussion

The performance of the model is evaluated using the indicators defined in Eqs. (1), (3), and (4). Table 1 illustrates the obtained values of these indicators for training and test data. The values of the coefficient of determination are very high. The values of RMSE and MAE are satisfying, bearing in mind that the maximal simulated value of the heating demand is 2179 kWh.

Table 1. Predictive performance of the model

Indicator	Training data	Test data
Coefficient of determination (R^2)	> 0.99	> 0.99
Root mean square error (RMSE)	30.41 kWh	38.24 kWh
Mean absolute error (MAE)	13.44 kWh	19.08 kWh

Figure 3. shows high correlation between the simulated and corresponding predicted values of the heating demand, which also indicates good overall performance of the model.

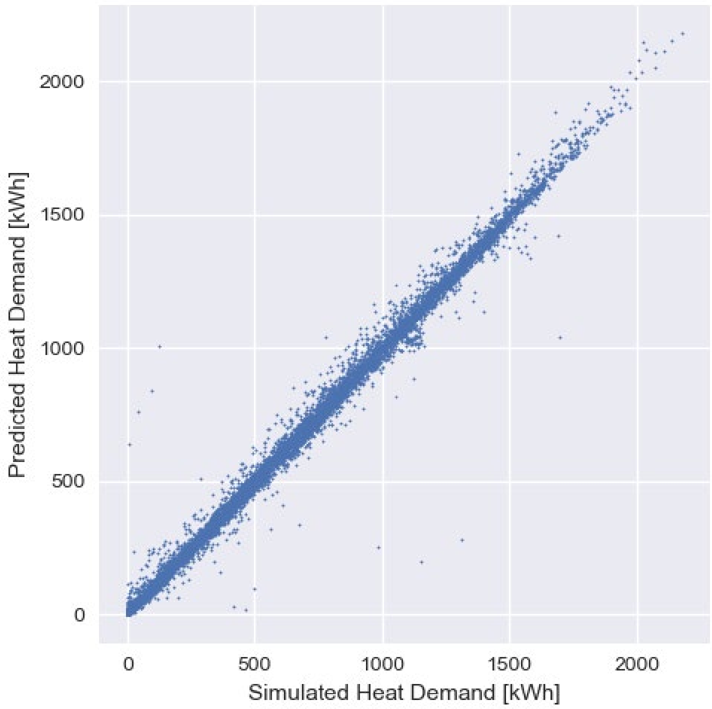

Figure 3. Correlation between simulated and predicted values

However, the maximal absolute error on the test set is 1028.65 kWh, which is too high. Moreover, there are some samples with very high errors. There are 185 sample points in the test set with errors above 100 kWh. The histogram on the left-hand side shown that the most such errors, i.e. 38 of them, have a value between 1000 and 1250 kWh. Most of these data points, 151 of them, are related to the morning hours (between 05:00 and 07:00) and evening hours (between 21:00 and 22:00), which coincide with the significant demand rises and drops that follow the setpoint temperature change.

Figure 5. illustrates feature importance calculated by RF model. The most important feature is the heat demand that occurred 24 h before the prediction time, followed by 1 h and 25 h lags, and the hour of the day. This corresponds to the correlation coefficients from Figure 1.

4. Conclusions

This paper presents a case of time-series, one hour ahead prediction of the heating demand of a group of residential buildings. The hourly demand data is obtained from detailed building simulations, using EnergyPlus software. The machine learning model is based on Random Forest regressor with 200 decision trees and has only time and lag variables as the inputs.

It appears that the 24^{th} lag variable is highly correlated to the predicted heating demand, and it is the variable with the highest feature importance. In general, the heat demand lag variables have much higher importance compared to the time variables.

The performance indicators, the coefficient of determination, root mean squared error, and mean absolute error, suggest good overall prediction performance, but there is a number of samples that are poorly predicted. These samples are mostly related to the early morning and late evening hours and correspond to high demand changes and setpoint temperature shifts.

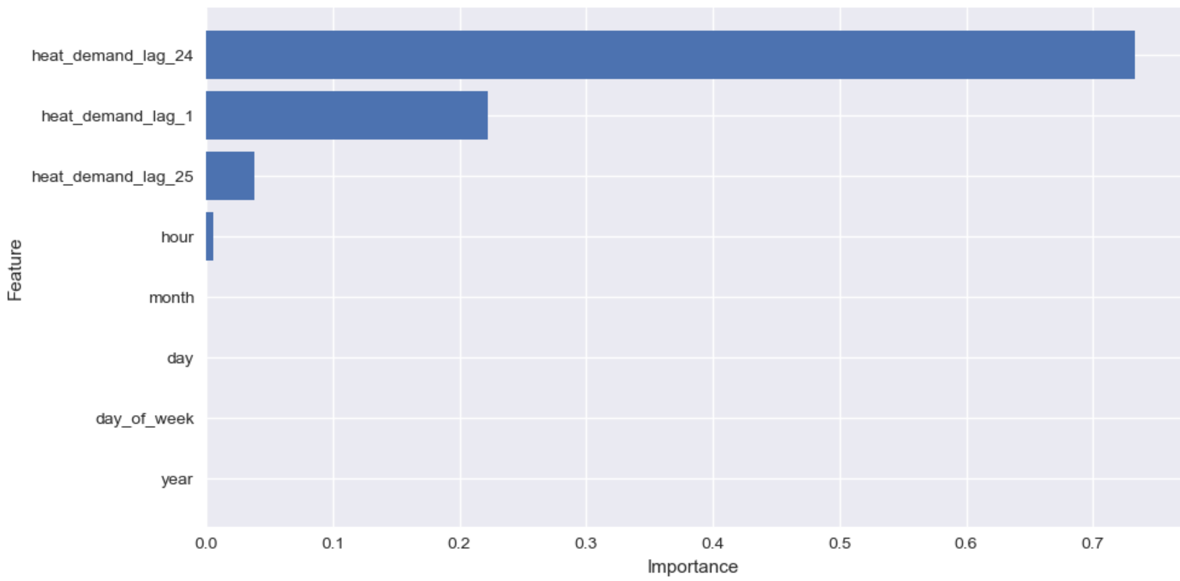

Figure 5. Feature importance

Acknowledgments

This research was supported by the Science Fund of the Republic of Serbia, Grant No. 23-SSF-PRISMA-206, Explainable AI-assisted operations in district heating systems - XAI4HEAT.

References

[1] Soheil Fathi, Ravi Srinivasan, Andriel Fenner, Sahand Fathi, Machine learning applications in urban building energy performance forecasting: A systematic review, Renewable and Sustainable Energy Reviews 133 (2020) 110287, https://doi.org/10.1016/j.rser.2020.110287

[2] Haohan Sha, Majd Moujahed, Dahai Qi, Machine learning-based cooling load prediction and optimal control for mechanical ventilative cooling in high-rise buildings, Energy & Buildings 242 (2021) 110980, https://doi.org/10.1016/j.enbuild.2021.110980

[3] Guangchen Li, Wei Tian, Hu Zhang, Xing Fu, A novel method of creating machine learning-based time series meta-models for building energy analysis, Energy & Buildings 281 (2023) 112752, https://doi.org/10.1016/j.enbuild.2022.112752

[4] Flavian Emmanuel Sapnken, Mohammad M. Hamed, Božidar Soldo, Jean Gaston Tamba, Modeling energy-efficient building loads using machine-learning algorithms for the design phase, Energy & Buildings 283 (2023) 112807, https://doi.org/10.1016/j.enbuild.2023.112807

[5] Mirko M. Stojiljković, Goran D. Vučković, Marko G. Ignjatović, Classification of retrofit measures for residential buildings according to the global cost, Thermal Science 25 (2021), No. 4 Part A, pp: 2677-2689, https://doi.org/10.2298/TSCI200825306S

[6] EnergyPlus, Energy Simulation Software, Available at: https://energyplus.net/, [accessed 01-Sep-2024].

[7] Yumin Liang, Yiqun Pan, Xiaolei Yuan, Wenqi Jia, Zhizhong Huang, Surrogate modeling for long-term and high-resolution prediction of building thermal load with a metric-optimized KNN algorithm, Energy and Built Environment 4 (2023) 709–724, https://doi.org/10.1016/j.enbenv.2022.06.008

[8] Xu Zhang, Yongjun Sun, Dian-ce Gao, Wenke Zou, Jianping Fu, Xiaowen Ma, Similarity-based grouping method for evaluation and optimization of dataset structure in machine-learning based short-term building cooling load prediction without measurable occupancy information, Applied Energy 327 (2022) 120144, https://doi.org/10.1016/j.apenergy.2022.120144

[9] Jui-Sheng Chou, Ngoc-Tri Ngo, Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption pattern, Applied Energy 177 (2016) 751–770, http://dx.doi.org/10.1016/j.apenergy.2016.05.074

[10] Shintaro Ikeda, Tatsuo Nagai, A novel optimization method combining metaheuristics and machine learning for daily optimal operations in building energy and storage systems, Applied Energy 289 (2021) 116716, https://doi.org/10.1016/j.apenergy.2021.116716

[11] Gianni Di Giovanni , Marianna Rotilio, Letizia Giusti, Muhammad Ehtsham, Exploiting building information modeling and machine learning for optimizing rooftop photovoltaic systems, Energy & Buildings 313 (2024) 114250, https://doi.org/10.1016/j.enbuild.2024.114250

[12] Marina M. Milićević, Budimirka R. Marinović, Machine learning methods in forecasting solar photovoltaic energy production, Thermal Science 28 (2024), No. 1, Part B, pp: 479-488, https://doi.org/10.2298/TSCI230402150M
[13] Omer Dagistanli, Hasan Erbay, Hasim Ahmet Yurttakal, Hakan Kor, Solar irradiation forecast by deep learning architectures, Thermal Science (26) 2022, No. 4, Part A, pp: 2895-2906, https://doi.org/10.2298/TSCI2204895D
[14] Ihsan Tugal, Fatih Sevgin, Analysis and forecasting of temperature using time series forecasting methods a case study of Mus, Thermal Science (27) 2023, No 4, Part B, pp: 3081-3088, https://doi.org/10.2298/TSCI2304081T
[15] Leo Breiman, Random Forests, Machine Learning, 45 (2001), pp. 5-32, https://doi.org/10.1023/A:1010933404324
[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A.,Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011), pp. 2825-2830