Comparison of Black Box Models for Load Profile Generation of District Heating Networks

G. Steindl
Research Cluster of Smart Energy Systems
Forschung Burgenland GmbH, Eisenstadt, Austria
e–mail: gernot.steindl@forschung-burgenland.at

Ch. Pfeiffer
Center of Methodological Competence
Forschung Burgenland GmbH, Eisenstadt, Austria
e–mail: christian.pfeiffer@forschung-burgenland.at

ABSTRACT

Black box modeling is a fast and efficient way of creating models for generating the heat demand of a district heating networks. A sufficient amount of high quality data has to be collected to form the basis for a valid model that can serve as training and test stand for the models. The model parameters and their influence on the heat demand are investigated and a model structure is derived. With this structure, five data mining algorithms, namely Multiple Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), k-Nearest Neighbor (k-NN) and Artificial Neural Networks (ANN) are utilized for creating the load models for a small district heating network located in southeast of Austria.
Except for LR, all algorithms showed a good performance. They are well suited for that kind of task. K-NN has the best regression score metric with an average MAPE of 13.49 %.

KEYWORDS

District heating network, black box model, machine learning, heat load profile, simulation, data mining

INTRODUCTION

The high fluctuation of renewable energy sources like wind and photovoltaics cause problems in the electric power grid. Storage systems can help overcome some of these problems by decoupling production and consumption. District heating networks can act as affordable thermal storage systems for a smart grid. To improve control strategies of power-to-heat in combination with photovoltaics and wind energy production, simulation-based studies have to be carried out. For these simulations not only the production of wind and PV has to be simulated, also the heat consumption in the district heating network has to be modelled.

Therefore, different kinds of models are used in building energy estimation, which can be classified in *Forward* approaches and *Data-driven* approaches [1]. Forward approaches use mathematical equations to describe the physical behavior. These types of models can become very complex and not very suitable as a load model for district heating networks.

The data driven approach can be further separated into a gray-box and a black-box approach. Gray-Box models use knowledge about some physical principles in the model domain, for example the thermal characteristics of a building, like in [2]. A similar approach for district heating networks is used in [3], where the theoretical relations of heat transfer is taken into account to create a model structure. This leads to a more complicated mathematical model for which the parameters have to be identified.

Black box models do not use any physical background. They are purely data driven and usually use some kind of supervised machine learning algorithm. Quality and the amount of input data is crucial for such models. If there is enough previously measured data of a district heating network, black box modelling is a fast and efficient way of creating such load models without considering the physical background. Data-driven approaches in general have the advantage that they allow to extract models from a large volume of data, which further down the line allows adapting and updating the model for new sets of data [4].

Another important factor in designing black box models is choosing the right input parameters. Research shows that the weather as well as the social behavior has the greatest influence on the heat load in district heating networks [5]. The outdoor temperature as well as the humidity correlates the most with the heat demand in a district heating network [6]. So these parameters are chosen as inputs for the investigated black box models. The social behavior is taken into account implicitly by using the day of the week and the current month as an additional input. This will reflect just a small part of the social behavior but keep the model complexity as low as possible. To capture the dynamics of the buildings which are connected to the heat net also past values of outdoor temperature and humidity are used.

The mathematical principles of the presented black box models are almost similar to the techniques used in heat demand forecasting. These forecast models were used for planning and optimizing heat generation of district heating networks [7,8]. In the presented work these techniques are used in a different context. The models are used to generate realistic heat load profiles for a specific period in time, based on well-known weather data for a specific region. The generated heat load profiles can be used in different simulation environments.

This work should give an overview over different black box models and test their performance. The validation of the algorithms is done with measurements from a small district heating network located in southeast of Austria.

METHODOLOGY

The important steps, which have to be taken in supervised Machine Learning, are data identification and collection as well as pre-processing [9].

The data collection was done with a digital heat meter, which was installed to measure a whole branch of the district heating network. This heat meter was connected via M-Bus to a PC, which

stored the values for further analysis. The data included the current heat demand with a timestamp. Because of the communication system design, the data was received in an unevenly manner and some data was missing. Therefore, the data had to be cleaned and transformed into an evenly spaced time series.

The data identification process deals with the problem of finding the right model input parameter. If too many input parameters are chosen, the model tends to over-fit the data. To provide generalization, as few as possible input parameters are used, but the parameters selected actually have an impact on the heat demand.

To investigate and prove the input parameters' impact on the heat demand, scatter plots, cross correlations, and box plots are used.

As the weather has much influence on the heat demand, weather data (temperature and humidity) were obtained from ZAMG [10] with a time resolution of 1 hour.

With the identified input parameter a model structure was derived. With this structure, five data mining algorithms, namely Multiple Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), k-Nearest Neighbor (k-NN) and Artificial Neural Networks (ANN) are utilized for creating the load models.

LR is a traditional approach to specify causal relationships between one or more independent attributes with some dependent attribute. Parameters for the independent variables are estimated by the method of least squares [11].

RF is an ensemble method that combines the prediction of many decision trees; it can be applied for classification, regression purposes as well as for ranking candidate attributes [12,13].

Support Vector Machines map some given data set from an input space into a high dimensional feature space using a kernel function. This approach is also suited for regression problems [14,15].

ANN are inspired by the way human brain processes information. In the multilayer perceptron type of ANN, the artificial neurons are organized in several layers of adaptive weights. In a feed-forward architecture, the outputs of one layer are used as the inputs of the following layer using a non-linear activation function without feedback loops [16]. The application of ANN is common in forecasting issues [17].

K-NN is an instance-based learning method in which each new instance is compared with existing ones and k closest existing instances are used to assign the class to the new ones. Euclidean distance function is often applied to compute difference between instances. Because of its competitive properties, it is widely considered in time-series applications and pattern-recognition problems [18,19].

To compare their performance, the regression score metrics root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), as well as the coefficient of determination (R^2) are calculated, given by

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right)^{\frac{1}{2}} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (2)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

where $\hat{y}_i$ is the predicted load, y_i is the observed load, $\bar{y}$ is the mean of the observed loads and n is the number of samples forecasted. Also, the year's duration curve of the district heating network is used as a special metric to analyze the model performance at peak and off-peak situations.

RESULTS

Input data analysis and pre-processing

The district heating network is a relatively small one with a star topology and length of about 1,860 m. The network is designed for a supply temperature of 90 °C and a return temperature of 60 °C. The whole net is fed at a single point by a combined heat and power unit (CHP) with a nominal thermal power of about 400 kW. A biomass boiler with a nominal power of 880 kW is used for peak load production. The heat is used for heating and tap water production in the connected buildings.

The daily average heat demand of the network measured is shown in fig. 1. As expected the heat demand during summer is very low, because most of the connected buildings are offices. In these buildings the tap water consumption is typically low during the whole year. The maximum heat demand during the measurement was about 790 kW but the average consumption is just 220 kW in the one-year-period that was observed.

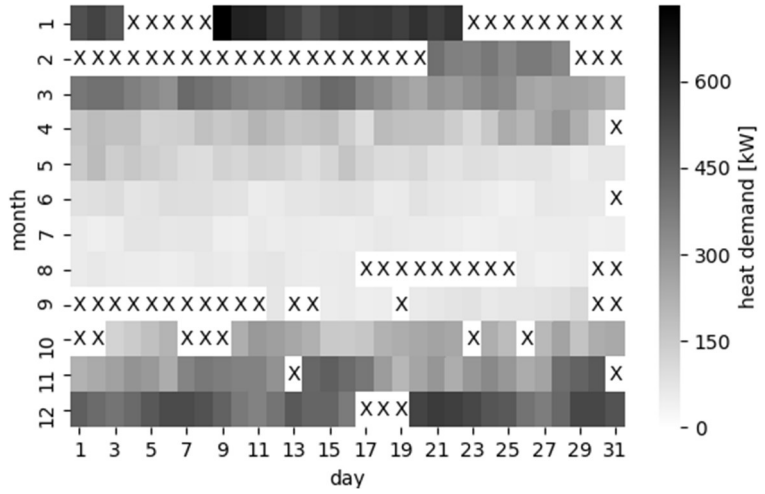

Figure 1. Heat demand of the district heating network in a one-year-period

The heat load was measured for a whole year in about 5 to 15 minute intervals. Because of measurement system design and delays during the data communication, the data was present as an unevenly spaced time series. Therefore, the data has to be pre-processed. It was transformed into an equally spaced series with a sampling interval of 15 minutes by linear interpolation. The evenly spaced time series is used for further analysis and as training data for the models. As the black box model should be able to produce the heat demand of the network for every hour, the hourly mean values were calculated and used as input data for the model.

Due to failures of the data acquisition system, some data is lost. Short periods of missing data, up to one hour, were reconstructed by linear interpolation. Fig. 1 also shows the missing data, which are indicated with an X in the heat map plot. The reason for the data acquisition failure couldn't be found, but longer gaps are marked as missing data and were excluded from the training set.

Model structure

Weather data for the region in which the district heating network is located were used to train the heat load models. To analyze the impact of the chosen input parameters, cross correlations of the ambient temperature and heat demand of the district heating network were investigated. Fig. 2 shows the scatter plot and the correlation coefficients. There is a high negative correlation between the temperature and the heat demand of the network. The humidity scatter plot shows less correlation with the heat demand. The highest correlation coefficient is found at a lag of 3

hours. This can be interpreted as the dynamics of the buildings, which are connected to the heating network.

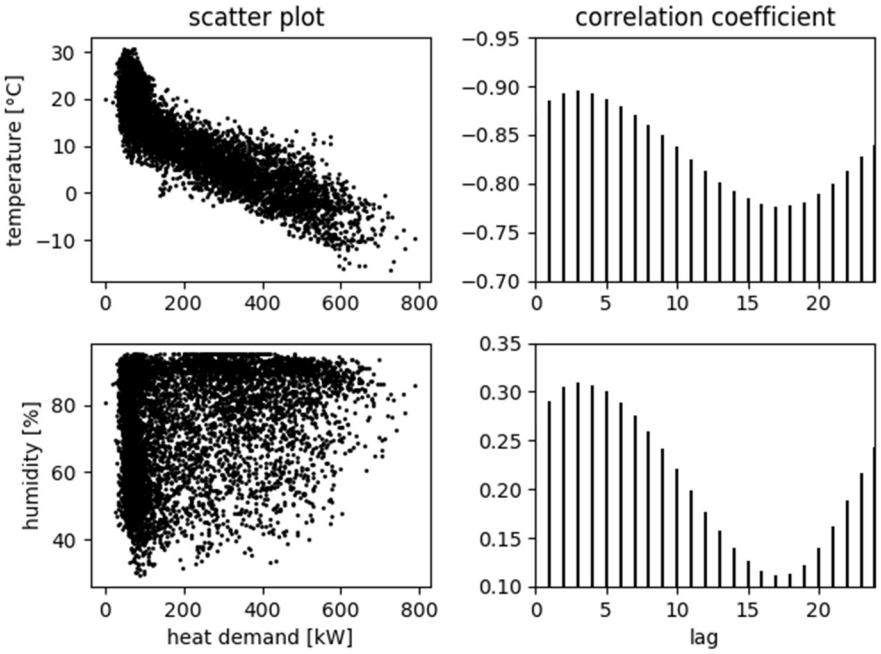

Figure 2. Scatter plots and cross-correlation coefficients of temperature and humidity

The day of the week is another input parameter, as well as the current month. As the variation of the heat demand over the months is obvious, because of the outdoor temperature change over the year, the heat demand variation over the week days is also investigated. Fig. 3 shows the box plots of the hourly heat demand values, separated for every single day. A slight variation across the week can be identified and confirmed by an analysis of variance ($F = 5.72$, $p < 0.01$). The reduced heat demand on weekends can be explained by the utilization of the supplied buildings. Only offices and industry are connected in the measured network.

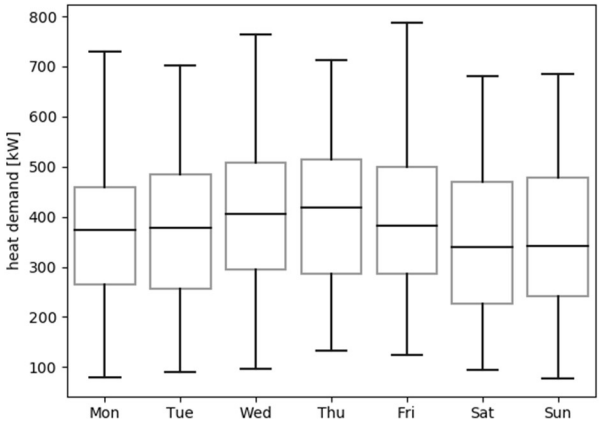

Figure 3. Comparison of heat demand for different days of the week from March until May

As the inputs and their influence on the heat demand are analyzed, the structure of the black box models is used, as shown in fig. 4. Temperature, humidity, the day of the week and the month are used as input parameters. As the correlation coefficient of the temperature and humidity shows a

lag of three hours, the input vector consists of the current and time delay values. This is indicated in fig. 4 with the delay operator of the Z-transformation z^n, where n is the time delay in hours.

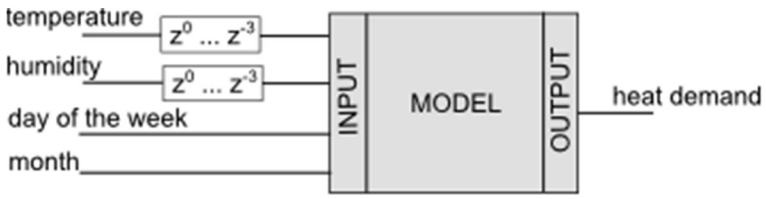

Figure 4. Model structure with input parameters

Model evaluation

With this black box model structure, different kinds of algorithms were evaluated. The RF was built with 500 trees. ANN was specified as a feed forward network with three hidden layers consisting of eight, five and three neurons, respectively. K-NN performed best when only two neighbors are considered.

Table 1 provides a comparison between the introduced data mining algorithms in terms of four statistical measures. The table reveals that the k-NN model outperforms the RF, SVR and ANN models in terms of each measure, whereas the LR model performs clearly worst. In winter and fall, comparatively high values of MAE are determined while values of MAPE indicate better properties in winter for predictive analyses.

Table 1. Comparison of regression score metrics

metric	method	test season				average
		Winter	Spring	Summer	Fall	
RMSE	LR	58.66	41.86	24.24	55.57	46.61
	RF	46.18	38.48	15.45	51.72	40.39
	SVR	53.45	37.65	12.29	53.60	42.33
	ANN	50.21	36.54	12.92	52.10	40.41
	k-NN	**46.12**	**37.32**	**10.95**	**45.39**	**37.77**
MAE	LR	46.83	30.94	18.88	43.44	34.64
	RF	36.44	27.08	9.57	38.23	27.86
	SVR	41.97	25.70	9.69	40.77	29.22
	ANN	39.51	26.59	9.51	39.24	27.91
	k-NN	**33.58**	**25.04**	**7.87**	**32.11**	**24.74**
MAPE (%)	LR	12.17	23.83	33.98	16.62	21.87
	RF	9.16	19.76	17.27	14.72	15.76
	SVR	10.65	18.39	17.78	14.87	15.77
	ANN	9.94	19.58	17.12	14.92	15.85
	k-NN	**8.64**	**17.77**	**13.99**	**11.49**	**13.49**
R^2	LR	0.77	0.79	0.71	0.83	0.93
	RF	**0.86**	0.82	0.83	0.85	0.94
	SVR	0.81	0.83	0.82	0.84	0.94
	ANN	0.83	**0.84**	0.83	0.85	0.94
	k-NN	**0.86**	0.83	**0.85**	**0.89**	**0.95**

Due to a reduced heat demand, RMSE and MAE measures in spring and summer are comparatively small. In terms of MAPE, measures compared with winter and fall are even

worse, indicating a poorer adequacy for prediction. In particular, the LR model behaves even more inaccurate compared to the other models in summer. Disregarding LR in summer, spring seems to be the most difficult season for predictions regarding heat demand.

Fig. 5 indicates the performance of RF and k-NN models in a one-week-time span in winter and spring. The graph shows that RF (dashed grey line) yields less accuracy than k-NN (dotted black line) in both seasons. In winter, RF is not able to reach the peaks and tends to underestimate the average heat demand. In comparison to the other models, the most precise model regarding the error measures appears to be k-NN as it catches the observed values in winter quite well. However, problems can be detected in cases of low heat demand in winter. On the one hand, viewing the accuracy on Sunday in the test week, heat demand is highly underestimated, on the other hand k-NN ignores low heat demand tails from Monday until Thursday. The test week in spring shows that RF behaves rather volatile without revealing a clear trend of under- or overestimating heat demand. In contrast, k-NN shows more accuracy in spring. However, some predictions still remain different from the observed values, confirming the reported error rates.

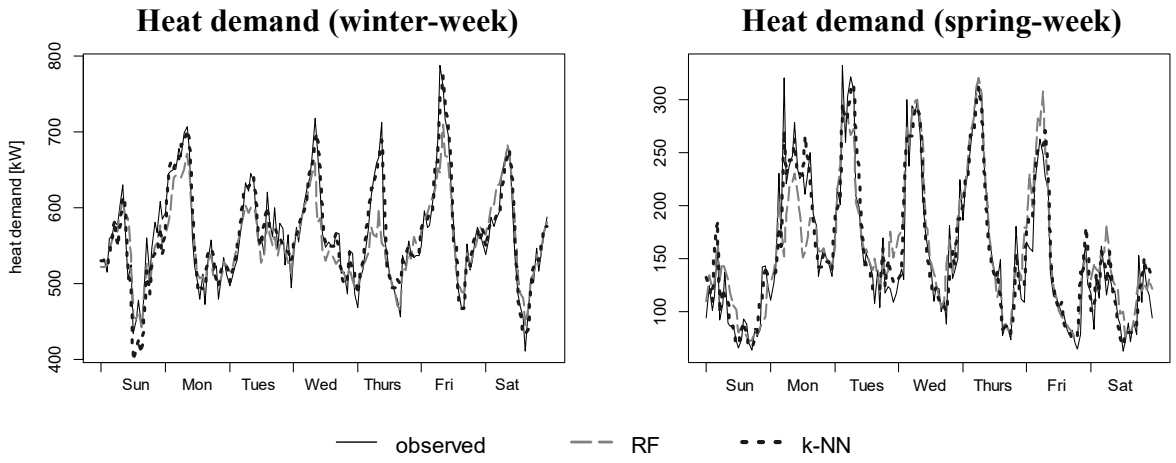

Figure 5. Observed and predicted heat demand in a one-week-time span in winter and spring

The duration curve of the considered heat net with predictions from the LR, RF and k-NN models is represented in fig. 6. Due to the fact that some data are missing, only about 6800 hours are implemented in the curve. Given a few peaks around 800 kW, about half of the time energy of 200 kW or less has been consumed. As shown in the bottom left figure, there is an abrupt decrease of heat demand within 100 hours. Only the k-NN model catches these peaks adequately. Especially LR, but also RF is not able to predict peak situations in an acceptable way. On the other hand, the bottom right figure shows the off-peak demand of the heat net. It is obvious that LR overestimates the off-peak demand for a long duration until its predicted load level decreases abruptly, even yielding some negative values. RF also overestimates the entire off-peak demand. Regarding k-NN, a marginal overestimation is identifiable with respect to the lowest load levels.

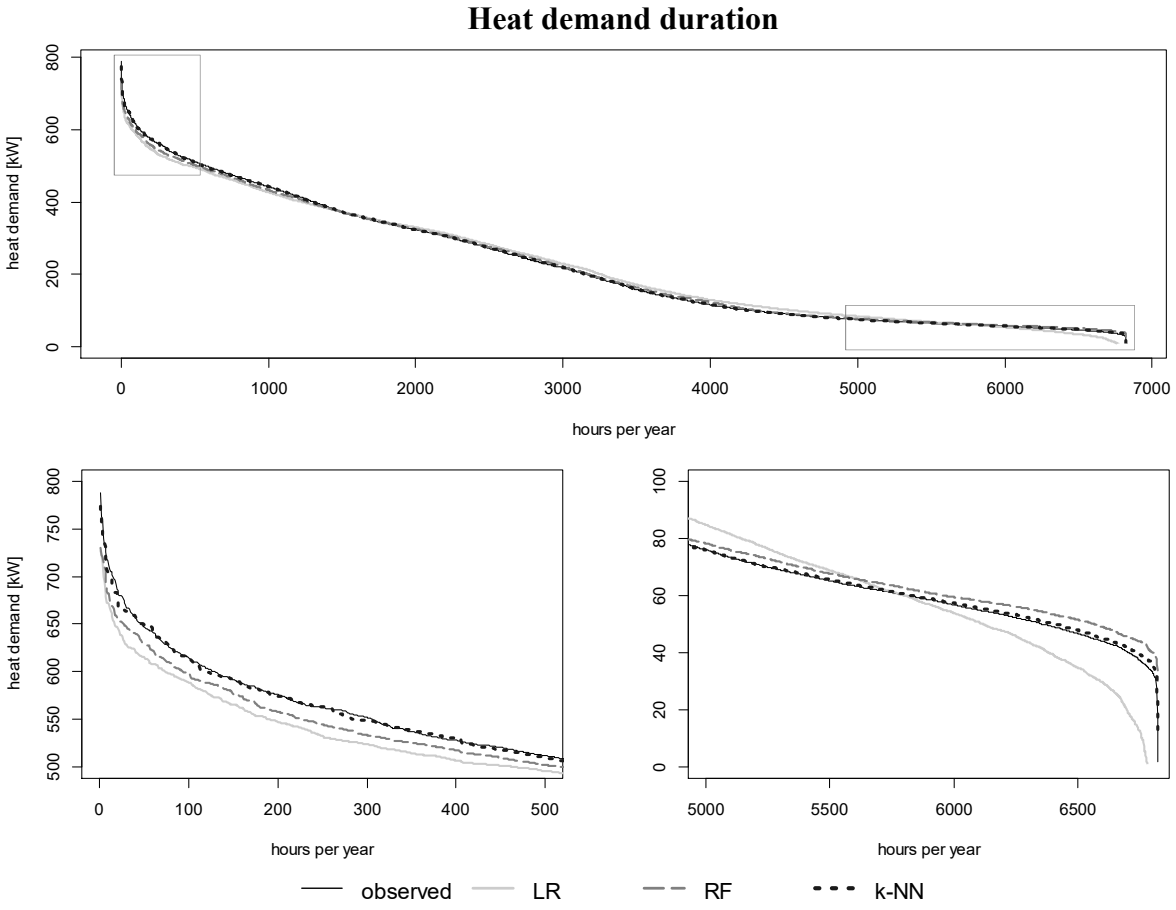

Figure 6. Heat demand duration curves

CONCLUSION

All investigated models, except the LR-model performed well for this type of application. They all have similar regression score metrics, although the k-NN has a slightly better performance with an average MAPE of 13.49 % and an overall R^2 of 0.95. But as shown, the model performance varies within the year, which has to be taken into account for further investigations.
The heat demand duration curve of a network is often used to characterize the whole district heat net. Our comparison showed that the models also perform well in reflecting this characteristic. In this curve it can be easily seen if a model performs well in peak as well as in off-peak situations. It can be seen that especially LR underestimates the heat demand at low or full load and only performs well in part load situations.
The heat net under investigation is a very small one; this has a negative impact on our models' performance, because in such small networks the stochastic influence is much higher than in large ones. Short and high heat load demands are very hard to model. In addition, the missing data leads to a reduced training set. This also has a negative influence on the model performance. For further improvement more data has to be gathered to enlarge the training data.
As shown, black box modeling and its data driven approach is a fast and sufficient way to generate models of district heating systems. However, as it is a data driven approach, it always requires enough data to perform well. The comparison shows the metrics of all models, except for the LR model, are almost equal. Although k-NN performed best, the difference is too little to

recommend only this algorithm for district heating network modeling. It is possible to increase the performance of the models by tuning their parameters, e.g. vary and optimize the amount of layers, neurons and activation functions for NN.

ACKNOWLEDGEMENT

The authors want to thank all supporters, the project Hybrid Grids Demo is supported by the Climate and Energy Fund within the funding scheme “Smart Cities Demo”.

NOMENCLATURE

ANN … Artificial Neural Network
CHP … Combined Heat and Power Unit
k-NN … k-Nearest Neighbor
MAE … Mean Average Error
MAPE … Mean Average Percentage Error
LR … Multiple Linear Regression
R^2 … Coefficient of Determination
RF … Random Forest
RMSE … Root Mean Squared Error
SVR … Support Vector Regression

REFERENCES

1. Fumo, N., A review on the basics of building energy estimation, *Renewable and Sustainable Energy Reviews*, Vol. 31, pp 53-60, 2014.
2. Madsen, H. and Holst, J., Estimation of continuous-time models for the heat dynamics of a building, *Energy and Buildings*, Vol. 22, pp 67–79, 1995.
3. Nielsen, H. A. and Madson, H., Modelling the heat consumption in district heating system using a grey-box approach, *Energy and Buildings*, Vol. 38, pp 63-71, 2006.
4. Kusiak, A., Li, M. and Zhang, Z., A data-driven approach for steam load prediction in buildings, *Applied Energy*, Vol. 87, pp 925-933, 2010.
5. Arvastson, L., Stochastic modelling and operational optimization in district-heating systems. *PhD Thesis*, Mathematical Statistics, Lund University, Lund, Sweden, 2001.
6. Yang, H. et al., Heat Load Forecasting of District Heating System Based on Numerical Weather Prediction Model, *2nd International Forum on Electrical Engineering and Automation IFEEA*, December 26-27, 2015, pp 1-5.
7. Streckiene G. et al, Feasibility of CHP-plants with thermal stores in the German spot market, *Applied Energy*, Vol. 86, pp 2308-2316, 2009.
8. Andersen, A.N. and Lund H., New CHP partnerships offering balancing of fluctuating renewable electricity productions, *Journal of Cleaner Production*, Vol. 15, pp 288-293, 2007.
9. Harrington, P., Machine Learning in Action, Manning Publications Co, Greenwich, CT, 2012.
10. Zentralanstalt für Meteorologie und Geodynamik (ZAMG), http://www.zamg.ac.at
11. Draper, N.R. and Smith, H.; Applied Regression Analysis, John Wiley & Sons, New York, 1998.
12. Breiman, L., Random Forests, *Machine Learning*, Vol. 45, pp 5-32, 2001.
13. Dudek, G., Short-Term Load Forecasting Using Random Forests, *Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014*, September 24-26, 2014, pp 821-828.
14. Vapnik, V., Golowich, S. and Smola, A., Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, Cambridge, MA, pp 281-287, 1997.
15. Chen, B.-J., Chang, M.-W. and Lin C.J., Load Forecasting using Support Vector Machines: A Study on EUNITE Competition 2001, *IEEE Transactions on Power Systems*, Vol. 19, No. 4, pp 1821-1830, 2004.
16. Bishop, C. M., Neural Networks for Pattern Recognition, Claredon Press, Oxford 1997.
17. Hayati, M. and Shirvany, Y., Artificial Neural Network Apporach for Short Term Load Forecasting for Illam Region, *International Journal of Electrical, Computer, and Systems Engineering*, Vol. 1, No. 2, pp 121-125 ,2007.
18. Yakowitz, S., Nearest-Neighbour Methods for Time Series Analysis, *Journal of Time Series Analysis*, Vol. 8, No. 2, pp 235-247, 1987.
19. Al-Qathani, F. H. and Crone, S. F., Multivariate k-Nearest Neighbour Regression for Time Series data – a novel Algorithm for Forecasting UK Electricity Demand, *Proceedings of International Joint Conference on Neural Networks*, August 4-9, 2013, pp 228-235.