K-Nearest Neighbor Regression for Forecasting Electricity Demand

Metodija Atanasovski, Mitko Kostov, Blagoja Arapinoski, and Mile Spirovski

Abstract – **Power system load forecasting plays a vital role in all aspects of power system planning, operation and control. It is a basic function for reliable and economical operation of power systems. This paper analyses the power system load forecast performed by applying k-nearest neighbour machine learning model, which is for the first time applied on real data of North Macedonia power system. The results are compared with polynomial and sinuses regressions.**

Keywords – **Power system load, Forecast, Machine learning, K-nearest neighbour, Correlation, Regression.**

I. Introduction

Electricity load forecasting is a process of projecting future electric energy demand in order to meet the increasing demand. Electric load forecasting will control which power plants should increase their output and which generators should be dispatched. In terms of the period of load forecasting, it can be divided into three categories: short-term load forecasting, medium-term load forecasting and long-term load forecasting [1].

Short-term load forecasting refers to forecasting electricity demand several days in advance every hour. An underestimation of the power load can result in lack of electricity production. On the other hand, in the case of overestimation of the load sub-optimal dispatch of power plants could be scheduled.

The factors that play a key role in forecasting energy consumption are temperature, type of the day (weekday, weekend, holiday), geographic differences, people standard, demographics, etc. The number of scientific articles about power forecasting is very large. Some of the used techniques include regression methods [2-4], discrete wavelet transform [5, 6], neural networks [7-11], fuzzy logic [12]. This paper studies the forecast of power load consumption by applying k-nearest neighbour (KNN) machine learning model, which is for the first time applied on real data of North Macedonia power system and the results are compared with polynomial and sinuses regressions. The implementation of the k-nearest neighbor machine-learning model is performed by using two independent variables: air temperature and date. It means the algorithm searches for/calculates suitable power load candidates around a certain period and temperature. The hourly data (8760 per year) for the temperatures and load are used for the years 2014-2018 as a training dataset, while 2019 (temperatures and load) data are used as a test dataset. The effectiveness of the model is evaluated and confirmed by cross-validation.

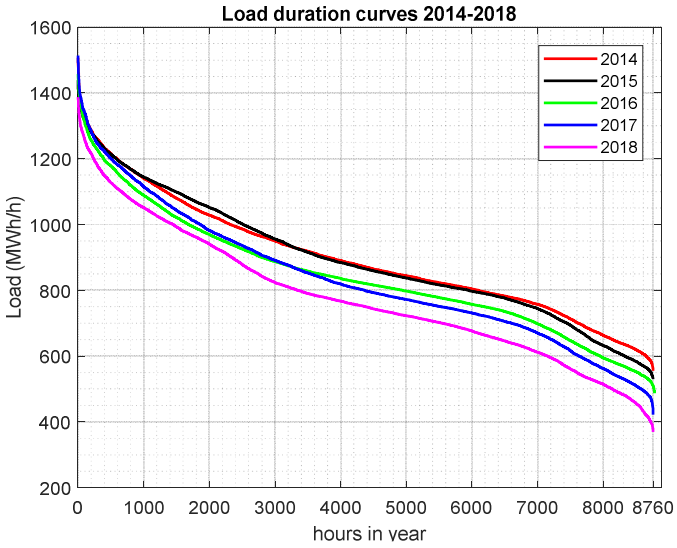


Fig. 1. Load duration curve for years 2014-2018 for power system of Republic of N.Macedonia

The remainder of this paper is structured as follows. Section II briefly describes the basic statistical analysis of power system load data and air temperature. The methodology used in the paper is introduced in Section III. Section IV elaborates the experimental results and the conclusions are given in Section V.

II. Basic Analysis

In this paper, dataset consists of power system load data for Republic of North Macedonia for the calendar years 2014-2019 on hourly basis (8760 per year) (Fig. 1) [13, 14] and the matching meteorological data about minimal, average and maximal air temperatures for the city of Skopje, North Macedonia [15].

Fig. 2 shows the normalized daily average power load diagram and normalized daily air temperature diagram from 2014 to 2019, both of which were normalized in the interval [0 1]. An average load is average of all 24-hourly loads on a daily diagram. The analysis of the power load data shows that there is a high variation between hourly loads in the power system on a year basis. Fig. 2 confirms the strong negative correlation between the power load and the air temperature

Metodija Atanasovski, Mitko Kostov, Blagoja Arapinoski and Mile Spirovski are with the Faculty of Technical Sciences-Bitola, University St. Kliment Ohridski, Republic of North Macedonia, E-mail: metodija.atanasovski@uklo.edu.mk

978-1-7281-7142-5/20/$31.00 ©2020 IEEE

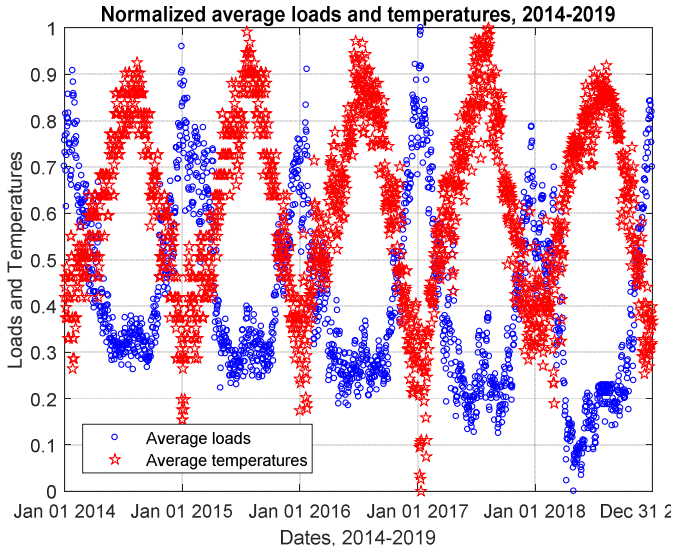


Fig. 2. Normalized average daily loads and air temperatures for the years 2014-2019

which is typical for this region: 1) the electric energy consumption is large in winter; 2) the positive air temperature peaks correspond to negative power load peaks and vice versa.

The regression analyses over power load and temperature datasets in 2014 and 2015 performed in [16] and [17] examined the approximations parameters, determination coefficients (R^2) and correlation coefficients [18]. According to [16] and [17], the determination coefficients for polynomial regression and sinuses regression of the maximal, average and minimal daily load due to the average temperature are very high, which means that the regression analysis shows high prediction degree of the daily typical loads from the air temperature. The correlation coefficients for polynomial regression and sinuses regressions have values in a range between –0.90 to –0.95 what implies very close negative relation between all the combinations of typical daily loads and air temperatures.

III. Methodology

K-nearest neighbour machine learning model [19] is considered for power system load forecast. The power load as a dependent variable depends on two independent variables – average air temperature and date. This means the algorithm will search for suitable k power load candidates around a certain period and temperature. After defining the training dataset (temperatures, dates and power load for the years 2014-2018), the model is tuned by setting an appropriate parameter for the number of neighbours k, and then it is trained on the training dataset.

The effectiveness of the model is evaluated by 10-fold and Leave-one-out cross-validations. Cross-validation is a method for obtaining reliable estimates of model performance using only training data. To predict the performance of a model on a new dataset, it is needed to assess its performance on a dataset that plays no part in the formation of the model – the test dataset. By comparing the test performance and training

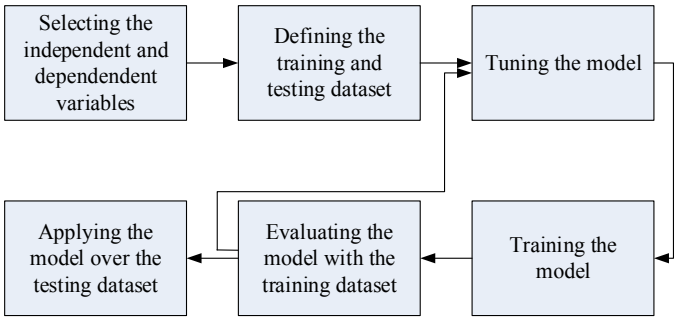


Fig. 3. Phases of the defined machine learning model

performance, overfitting can be avoided. If the model performs well on the training data, but poorly on the test data, then it is overfitted. Besides with the correlation coefficient R and determination coefficient R^2, the performance can be measured in other ways; two of those ways are through the root-mean-squared error *RMSE* and mean-absolute error *MAE*:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}, \quad (1)$$

$$MAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n}, \quad (2)$$

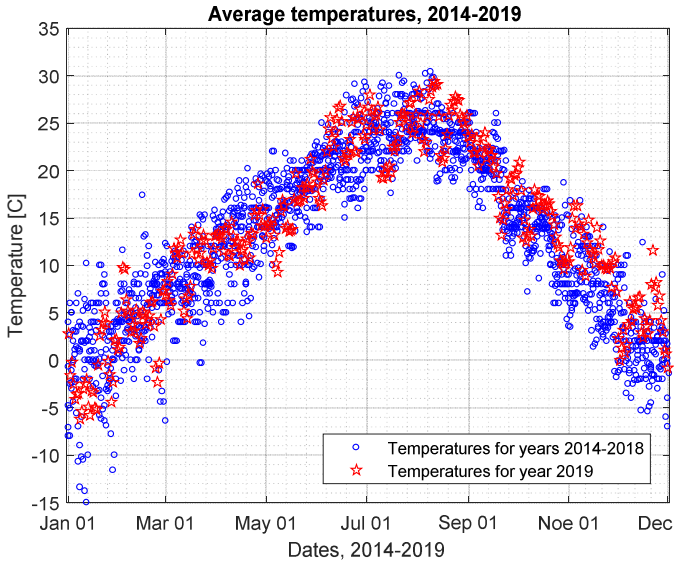
where p_i and a_i are the predicted and actual values, respectively, while n is the total number of the test instances.

All the phases of the machine learning model are illustrated in Fig. 3.

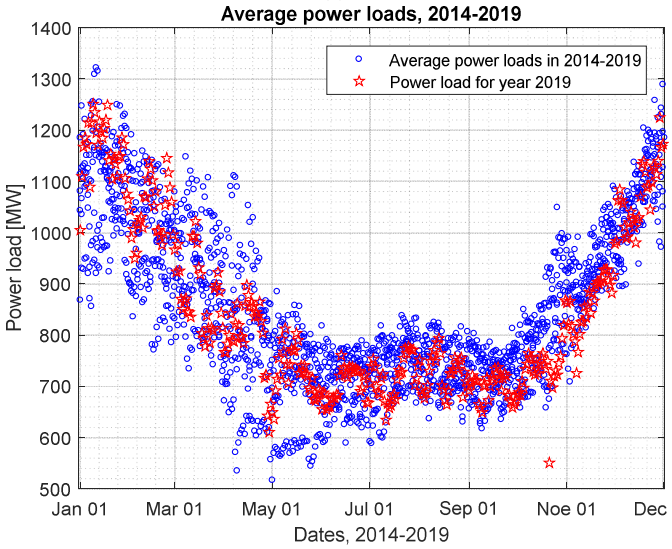
IV. Case Study

This section presents the results obtained by using the presented methodology. The temperatures, dates and power load for the years 2014-2018 are used as a training dataset, while corresponding data for 2019 as a testing dataset. Figure 4 illustrates distributions of the average air temperatures and the average power loads for the years 2014-2018 through 365 days, respectively. The blue circles in Figures 4 denotes the temperatures and power loads in the period 2014-2018, respectively, while the red stars denote the temperatures and power loads in the forecast period 2019. Figure 5 illustrates distribution of the real average power load in the period Mar. 1-21 2014-2019.

According to the above methodology the average power system load for the year of 2019 is forecasted on the basis of the defined training dataset. KNN machine learning model is used over the two independent variables, average air temperature and date, and the distance between neighbours is measured by Euclidean distance function as a commonly used distance metric. The experiments for this case study show that a suitable number of neighbours in the model is 30. The minimum and the maximum of the variable average air temperature data are –15 and +30(C), respectively, while the minimum and maximum of the variable date are 1 and 365 (the first and the last day in a year), respectively. Variables measured on different scales contribute differently to the analysis, which may eventually lead



(a)



(b)

Fig. 4. Distribution of average air temperatures and average power loads for the years 2014-2019

to bias. The variable date (due to the larger range) will outweigh the variable air temperature, i.e. the variable date will have a bigger weight in an analysis compared to the variable air temperature. This means the date will have higher influence on the calculated distance than the air temperature will do. Transforming the data to a comparable scale can avoid this problem. Normalization is a way of standardizing a set of numbers so each one is between 0 and 1. Hence, both the variables in this model are normalized in the range [0–1].

The model is evaluated by 10-fold and Leave-one-out cross-validations through the root-mean-squared error and the mean-absolute error (30 neighbours used over the 2014–2018 dataset). Results given in Table 1 show that the errors are smaller when the variables are normalized.

Results of applying the model over the testing dataset (2019, 30 neighbours used) and forecasting the corresponding power load are given in Table 2 and Figure 6. Table 2 gives an overview of *MAE* and *RMSE* errors and correlation coefficient

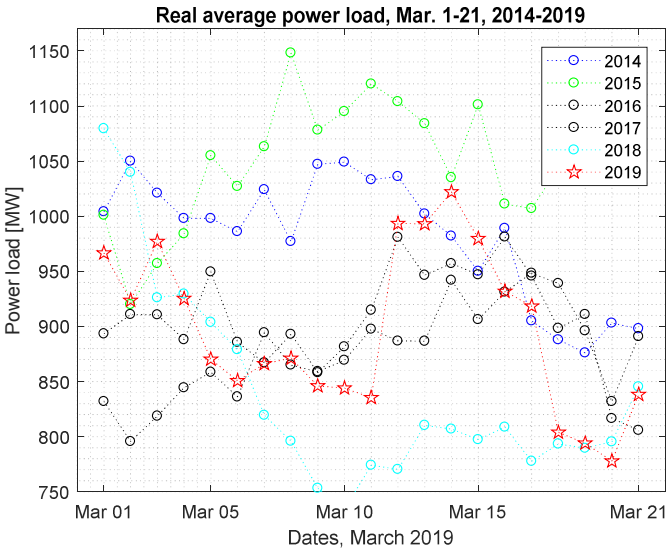


Fig. 5. Distribution of average power load for the period Mar. 01-21, 2014-2019

TABLE I
PERFORMANCE OF THE MODEL MEASURED ON THE TRAINING DATASET [MW]

Cross-validation	normalized variables	non-normalized variables
10-fold	62.2467	66.4218
Leave-one-out	65.8212	62.2133

for power load forecast with different models. Fig. 6 graphically illustrates the deviation between forecasted and real power load data. These results show that for power load forecasts, k-nearest neighbour regression outperforms polynomial and sinuses regressions.

V. SUMMARY

This paper is the first one using k-nearest neighbour machine-learning model for investigation the forecast of power system load in correlation with air temperature and date on real data of power system of Republic of North Macedonia. On the basis of statistical analysis, it can be noticed that there is a close time matching in appearance of power system maximum load and minimum air temperature. The same time matching is noticed between power system summer maximum load and registered maximum temperature in analysed years. The presented results show that for power load forecasts the proposed algorithm outperforms polynomial and sinuses regressions. The effectiveness of the model is evaluated by 10-fold and Leave-one-out cross-validations. Methodology works with all types of days and simulation results are given in a large time frame.

ACKNOWLEDGEMENT

This research is supported by the EU H2020 project TRINITY (Grant Agreement no. 863874) This paper reflects only the author's views and neither the Agency nor the

TABLE II
COMPARISON OF *MAE*, *RMSE* ERRORS AND CORRELATION COEFFICIENT *R* OF DIFFERENT MODELS FOR POWER LOAD FORECAST FOR 2019 DATA

Algorithm	***MAE***	***RMSE***	***R***
k-nearest neighbour (normalized variables)	38.4046	50.6231	0.9614
k-nearest neighbour (non-normalized var.)	39.1885	51.7435	0.9564
polynomial (order 4)	41.3541	56.4752	0.9488
sinuses (order 4)	40.7253	55.3602	0.9523
sinuses (order 4) + wavelet transform	41.1788	54.2211	0.9584

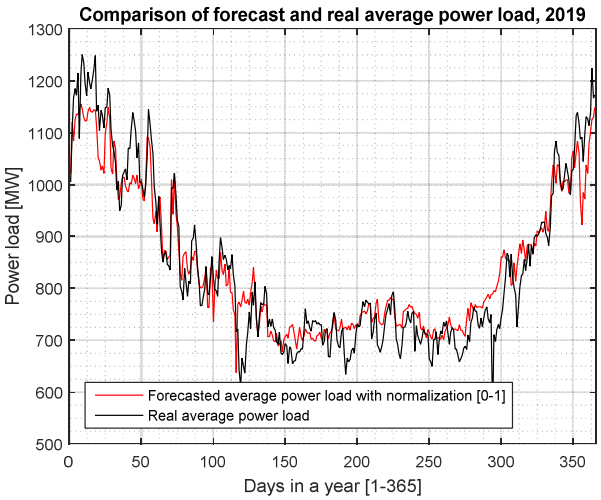


Fig. 6. Comparison of power load KNN forecast with normalized variables and real average power load for the year 2019

Commission are responsible for any use that may be made of the information contained therein.

REFERENCES

[1] E. A. Feinberg, D. Genethliou, Chapter 12: Load Forecasting, *Applied Mathematics for Power Systems*, pp. 269-282.

[2] W. Charytoniuk, M. S. Chen, and P. Van Olinda, "Nonparametric Regression Based Short-Term Load Forecasting", IEEE Trans. Power Syst., vol. 13, no. 3, pp. 725-730, Aug. 1998.

[3] S. Rucic, A. Vuckovic, and N. Nikolic, "Weather Sensitive Method for Short Term Load Forecasting in Electric Power Utility of Serbia", IEEE Trans. Power Syst., vol. 18, no. 4, pp. 1581-1586, Nov. 2003.

[4] T. Hong, M. Gui, M. E. Baran, and H. L. Willis, "Modeling and Forecasting Hourly Electric Load by Multiple Linear Regression with Interactions", in Proc. IEEE PES Gen. Meeting, Jul. 2010, pp. 1-8.

[5] Y. Chen, P. B. Luh, C. Guan, Y. Zhao, L. D. Michel, M. A. Coolbeth, P. B. Friedland, and S. J. Rourke, "Short-Term Load Forecasting: Similar Day-Based Wavelet Neural Networks", IEEE Trans. Power Syst., vol. 25, no. 1, pp. 322-330, Feb. 2010.

[6] G. Sudheer and A. Suseelatha, "Short Term Load Forecasting using Wavelet Transform Combined with Holt_Winters and Weighted Nearest Neighbor Models", Int. J. Electr. Power Energy Syst., vol. 64, pp. 340-346, 2015.

[7] W. Charytoniuk, M.-S. Chen, "Very Short-Term Load Forecasting using Artificial Neural Networks", IEEE Trans. Power Syst., vol. 15, no. 1, pp. 263-268, Feb. 2000.

[8] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation", IEEE Trans.Power Syst., vol. 16, no. 1, pp. 44-55, Feb. 2001.

[9] P. Mandal, T. Senjyu, N. Urasaki, and T. Funabashi, "A Neural Network Based Several-Hour-Ahead Electric Load Forecasting using Similar Days Approach", Int. J. Electr. Power Energy Syst., vol. 28, no. 6, pp. 367-373, 2006.

[10] H. Shi, M. Xu, and R. Li, "Deep Learning for Household Load Forecasting - A Novel Pooling Deep RNN", IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 5271-5280, Sep. 2018.

[11] M. Cai, M. Pipattanasomporn, and S. Rahman, "Day-ahead Building-Level Load Forecasts using Deep Learning vs. Traditional Time-Series Techniques", Appl. Energy, vol. 236, pp. 1078-1088, 2019.

[12] *Dispatching and Operational Reports for Power System Load Data* obtained from MEPSO.

[13] H. H. Çevik and M. Çunka³, "Short-Term Load Forecasting using Fuzzy Logic and ANFIS", Neural Comput. Appl., vol. 26, no. 6, pp. 1355-1367, 2015.

[14] D. Bajs, M. Atanasovski, *Longterm Forecast Study of Electrical Energy and Power Balance and Adequacy Analysis of Transmission Network of Republic of Macedonia*, Zagreb/Skopje EIHP, 2016.

[15] https://www.wunderground.com/history

[16] M. Atanasovski, M. Kostov, B. Arapinoski, and I. Andreevski, Correlation between Power System Load and Air Temperature in Republic of Macedonia, Int. Scientific Conf. on Information, Communication and Energy Systems and Technologies, pp. 213-216, Sozopol, Bulgaria, Jun. 2018.

[17] M. Kostov, M. Atanasovski, G. Janevska, and B. Arapinoski, "Power System Load Forecasting by using Sinuses Approximation and Wavelet Transform", Int. Scientific Conf. on Information, Communication and Energy Systems and Technologies, pp. 273-276, Ohrid, North Macedonia, Jun. 2019.

[18] S. Vukadinovic, *Elements of Probability Theory and Mathematical Statistics*, Belgrade, 1973.

[19] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Burlington, USA, Morgan Kaufmann, 2017.