Implementation of K-Nearest Neighbour Regression for Forecasting Electricity Demand in Power System of Republic of North Macedonia

Mitko Kostov*, Metodija Atanasovski, Mile Spirovski
Department of Electrical Engineering
Faculty of Technical Sciences, University St. Kliment Ohridski – Bitola, Republic of N, Macedonia
e–mails: mitko.kostov@uklo.edu.mk, metodija.atanasovski@uklo.edu.mk, mile.spirovski@tfb.uklo.edu.mk

ABSTRACT

Load forecast is an important factor for operational and development planning of power system. Factors that play key role in forecasting power load consumption are the air temperature, type of the day (weekday, weekend or holiday), geographical differences, people standard, gross domestic product, demographic information, energy efficiency etc. The air temperature is one of the factors, which has significant impact on electricity consumption and power system load. This paper analyses the correlation between the power system load and the air temperature in Republic of North Macedonia. Furthermore, forecasting of the power system load is investigated. The power system load forecast is performed by applying k-nearest neighbour machine learning model. The power load depends on two variables – air temperature and date. Results show that for power load forecasts, k-nearest neighbour regression outperforms polynomial and sinuses regressions.

KEYWORDS

Power system load, Load forecast, Machine learning, k-nearest neighbour regression.

INTRODUCTION

Power system load forecast is an important factor for operational and development planning of power system and future electrical energy consumption. Good forecast leads to significant savings in operating and maintenance costs and increased reliability of the electricity supply system. Factors that play key role in forecasting power load consumption are the air temperature, type of the day (weekday, weekend or holiday), geographical differences, people standard, gross domestic product, demographic information, energy efficiency, etc. This means that load forecasting is a complex multi-variable estimation problem where forecasting methods such as curve fitting using numerical methods do not provide accurate results, while at the same time some machine learning models perform better.

Air temperature is one of the factors, which has significant impact on electricity consumption and power system load. Air temperature impact on system load is especially important from the power system operational management aspects on short run. This fact is evident in power system of Republic of North Macedonia due to high variations of consumption and load in year seasons.

* Corresponding author

The annual number of scientific papers on load forecasting has increased from around a hundred in 1995 to more than a thousand in recent years. Accordingly, it is hard to make a thorough state of the art, and here only some of the papers are mentioned. There is a number of techniques which have been used for load forecast, such as: single or multiple linear or nonlinear regressions, stochastic time series, exponential smoothing, state space and Kalman filter, knowledge based approach, neural networks, wavelet transformations, semi parametric additive model, fuzzy logic etc. The first generation of the load forecast methods (also called analytical methods) includes time series analysis, regression methods [1-3], similar day method, Wavelet Transform [4-5], etc. Artificial intelligence methods, also known as the second generation of the load forecast methods, mainly comprises artificial neural networks (ANN) [6-8], including deep neural networks [9-10], random forests, gradient boosting [11], fuzzy logic [12]. The second generation compared with the first one has gained importance due to errors reduction. Some combination of methods (known as hybrid methods) that belong to the both generations is also possible [13-14]. Authors in [15] have concluded that all the previous methods appear to have at least one of the following three limitations:

- they might only work for a subset of days (i.e. load forecast is performed only for a given class of days, e.g., working days);
- simulation results are given for a small window of time (e.g., a couple of months);
- experiments are conducted on a single set of data, which might make a reader wonder whether the proposed methodology depends on the specific data-set, or can be actually adopted to predict the load in other countries as well.

This paper analyses the correlation between the power system load and the air temperature in Republic of North Macedonia. Furthermore, forecasting of the power system load consumption is investigated. The power system load forecast is performed 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 neighbour 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.

The proposed methodology tries to overcome some of the previous methods limitations mentioned, above. Namely, methodology works with all types of days and simulation results are given for a large window of time. Another contribution of the paper is comparative analysis between k-nearest neighbour regression and polynomial and sinuses regressions. The results show that for power load forecasts the proposed algorithm outperforms polynomial and sinuses regressions.

METHODS

Efficient and precise forecasts of energy requirements of a system are important for making decisions including decisions on purchasing and generating electric power, load switching and infrastructure development. The power load forecasting relies on historical data to determine how much power customers may need. Forecasting model inputs can include day of the week, holiday calendars, weather conditions and forecasts, geographical differences, demographic information, etc. Accurate models for electric power load forecasting are essential to the operation and planning of a utility company [16].

K-nearest neighbour machine learning model [17] is considered for power system load forecast. The power load is selected as a dependent variable that 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 getting a reliable estimate 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 performance, overfitting can be avoided. If the model performs well on the training data, but poorly on the test data, then it is overfitted. The performance can be measured in various ways, and one way is through the root-mean-squared error:

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

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. 1.

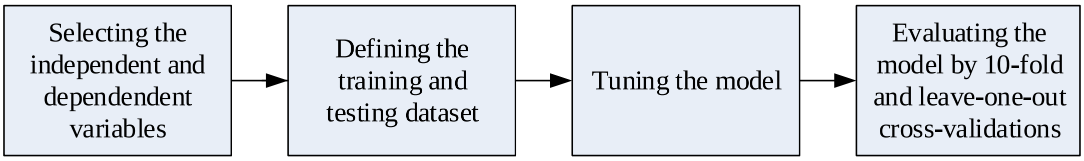
Selecting the independent and dependendent variables
Defining the training and testing dataset
Tuning the model
Evaluating the model by 10-fold and leave-one-out cross-validations

Fig. 1 Phases of the defined machine learning model.

Dataset overview and basic analysis

In this paper, a case study dataset consists of hourly power system load data (8760 per year) for Republic of N. Macedonia for the calendar years 2014-2019 [18, 19] and the corresponding meteorological information about minimal, average and maximal air temperatures obtained from the internet [20] (Figure 2). An analysis of the power load data shows that it depends on the period of the year, day of the week and hour of the day and there is a high variation between hourly loads in the power system on year basis.

Figure 3 depicts a daily diagram of power system of Republic of N. Macedonia for the day Jan 08, 2014. Three typical points (minimal power system load Pmin=836MW, average power system load Pavg=1100MW, maximal power system load Pmax=1293MW) are marked on the diagram and they are used in the analysis from each daily diagram. An average load is average of all 24-hourly loads on a daily diagram. In [21] and [22] regression analyses were performed over the dataset of power loads and air temperatures in order to estimate dependence curves of these three typical loads (Pmin, Pavg, Pmax) from the independent variable – the average temperature Tavg for the years 2014 and 2015. It was shown that there is a strong negative correlation between the power load and the air temperature. The regression analyses examined the approximations parameters, determination coefficients (R^2) and correlation coefficients. The determination coefficient shows the proportion of the variance in the dependent variable that is predictable from the independent variable (it ranges from 0 to 1, the coefficient 0 means the dependent variable cannot be predicted from the independent variable, while 1 means the dependent variable can be predicted without error from the independent variable) [23]. According to [21] and [22], the determination coefficients for polynomial regression and sinuses regression of the maximum, average and minimum daily load due to the average daily temperature are very high, which means that the regression analysis shows high prediction degree of the daily typical loads from the air temperature.

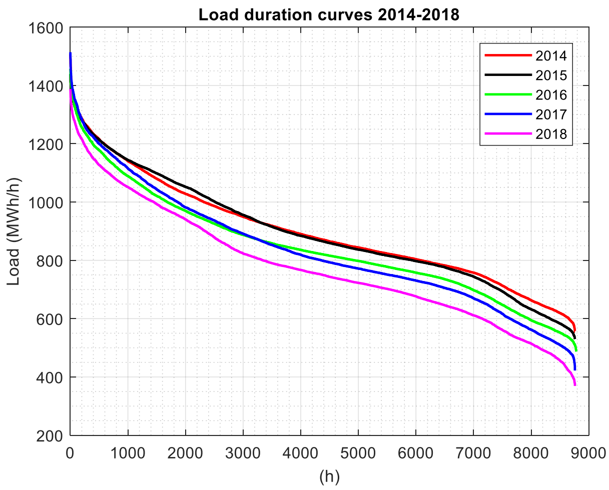

Figure 2. Load duration curve for years 2014–2018 for power system of Republic of North Macedonia.

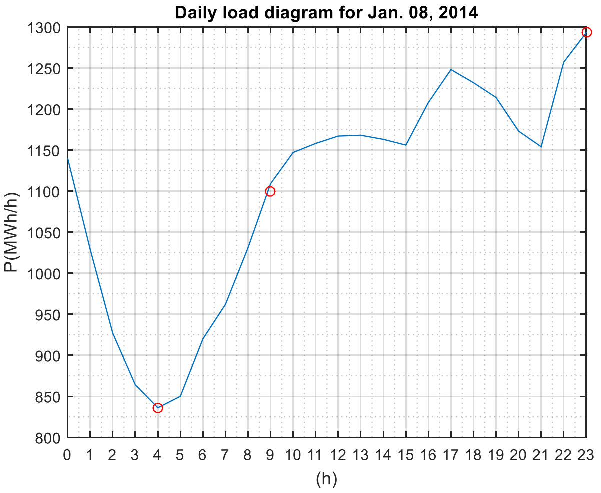

Figure 3. Daily load diagram for Jan. 08, 2014 and typical load points.

The correlation coefficient is a statistical measure that calculates the strength of the relationship between two variables [23]. Coefficients of correlation can have values in range from –1 (negative relation) to +1 (positive relation). There is not a significant relation if the correlation coefficient is less than 0.3. The correlation is with practical importance when the correlation coefficient is between 0.5 and 0.7. Correlation coefficient between 0.7 and 0.9 shows close correlation, while correlation coefficient greater than 0.9 shows very close correlation. According to [21] and [22] 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.

RESULTS AND DISCUSION

Presented methodology is used for investigation the correlation between the power system load and the air temperature in Republic of North Macedonia. In this case study, the temperatures, dates and power load for the years 2014-2018 are used as a training dataset. Figure 4 and Figure 5 illustrate the distributions of the average air temperatures and the average power load for the years 2014-2018 through 365 days, respectively. In these graphics, the number 1 in the apsis means the date Jan. 01, the number 32 is used for the date Feb. 01, etc. The blue circles in Figure 4 denote the temperatures in the period 2014-2018, while the red stars denote the temperatures in the forecast period Mar. 01-21, 2019, which was analysed in [21] and [22]. The blue circles in Figure 5 denote the real average power load in the period 2014-2018, while the red stars denote the real average power load in the forecast period Mar. 01-21, 2019.

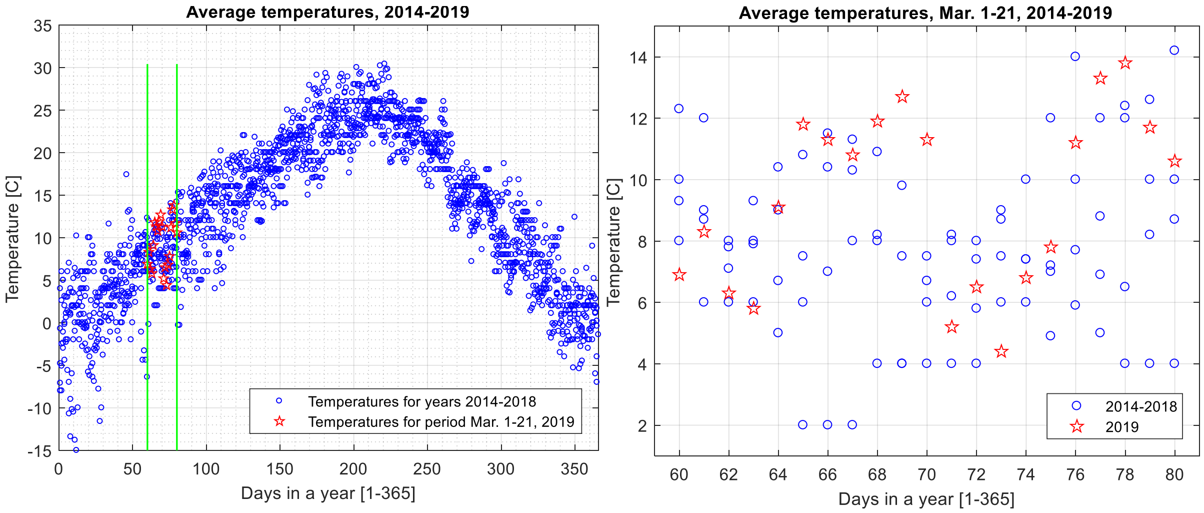

Figure 4. Distribution of average air temperatures for the years 2014-2019:
a) period 365 days, b) period Mar. 01-21, 2014-2019.

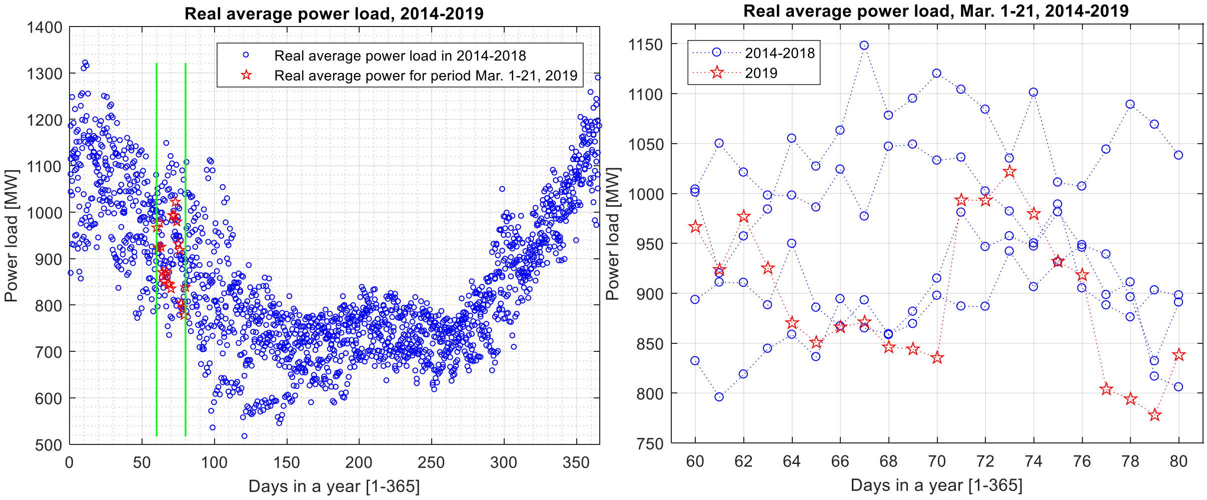

Figure 5. Distribution of average power load for the years 2014-2019:
a) period 365 days, b) period Mar. 01-21, 2014-2019.

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. The experiments for this case study show that a suitable number of neighbours in the model, which distance is measured by Euclidean distance function as a commonly used distance metric, 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 at different scales do not contribute equally to an analysis and this might end up creating a 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 comparable scales can prevent 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 root-mean-squared errors when 10-fold cross-validation and Leave-one-out cross-validation are used for the KNN model with 30 neighbours over the 2014–2018 dataset are given in Table 1. These results show that the errors are smaller when the variables are normalized.

Table 1. 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

In this case study, the following two periods are analysed, and corresponding power load is forecasted: 1) the period of Mar. 01–21, 2019 and 2) the period of the year 2019 (365 days). The forecasted average power loads for these periods are compared to the corresponding real average -power loads and the root-squared-mean errors are given in Table 2.

Table 2. Comparison of root-mean-squared errors of different models (MW)

Period	k-nearest neighbour regression		polynomial and sinuses regressions [21-22]		
	normalized variables	non-normalized variables	polynomial order 4	sinuses order 4	sinuses order 4 + wavelet transform
Mar.1-21,19	32.339	39.166	161.086	162.885	152.627
Year 2019	50.623	51.744	–	–	–

These results show that for power load forecasts, k-nearest neighbour regression outperforms polynomial and sinuses regressions. The graphics in Figure 6 depict comparison of forecasts of average power load with and without of normalization of variables (30 neighbours used) against real average power load for the period Mar. 01–21, 2019 (Fig. 6a) and for year 2019 (Fig. 6b). The graphics show that forecasts obtained with normalized variables are very close to the real average power load.

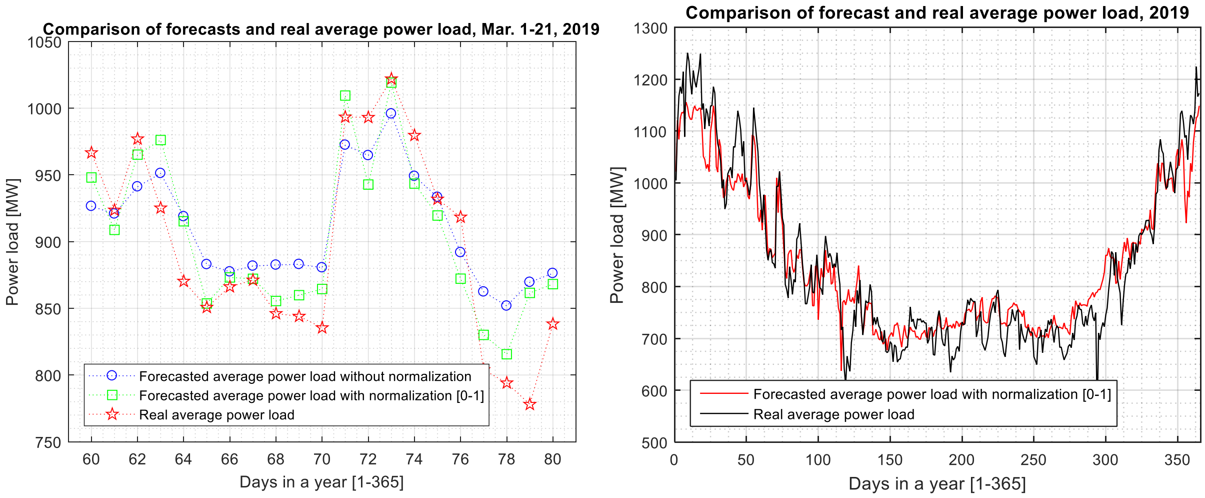

Figure 6. Comparison of: a) power load KNN forecast without normalized variables, power load KNN forecast with normalized variables, real average power load for the period Mar. 01–21, 2019, b) power load KNN forecast with normalized variables, real average power load for 2019.

CONCLUSION

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. Also it has to be emphasized that in North Macedonia electricity is widely used for heating of residential and commercial buildings and houses. This fact explains the very close correlation between power system load and air temperature in North Macedonia power system.

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 cross-validation and Leave-one-out cross-validation. Methodology works with all types of days and simulation results are given for a large window of time.

ACKNOWLEDGMENT

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

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. W. Charytoniuk and 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. S. Papadopoulos and I. Karakatsanis, “Short-term electricity load forecasting using time series and ensemble learning methods,” in *Proc. IEEE Power Energy Conf. Illinois (PECI)*, Feb. 2015, pp. 1-6.
12. 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.
13. P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, “Short-term residential load forecasting: Impact of calendar effects and forecast granularity,” *Appl. Energy*, vol. 205, pp. 654 - 669, 2017.
14. M. Ghayekhloo, M. B. Menhaj, and M. Ghofrani, “A hybrid short-term load forecasting with a new data preprocessing framework,” *Electr. Power Syst. Res.*, vol. 119, pp. 138-148, 2015.
15. M. Tucci, E. Crisostomi, G. Giunta, and M. Raugi, “A Multi-Objective Method for Short-Term Load Forecasting in European Countries,” *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3537 - 3547, Sep. 2016.
16. E.A Feinberg, D Genethliou, *Chapter 12 Load forecasting, Applied Mathematics for Power Systems*, pp.269-282.
17. I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques, Burlington*, USA, Morgan Kaufmann, 2017.
18. Dispatching and operational reports for power system load data obtained from MEPSO.
19. 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.
20. https://www.wunderground.com/history.

21. M. Atanasovski, M. Kostov, B. Arapinoski, 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*, ISSN: 2603-3259 (Print) ISSN: 2603-3267 (Online), pp. 213-216, Sozopol, Bulgaria, Jun. 2018.
22. M. Kostov, M. Atanasovski, G. Janevska, B. Arapinoski, Power System Load Forecasting by using Sinuses Approximation and Wavelet Transform, *Int. Scientific Conf. on Information, Communication and Energy Systems and Technologies*, ISSN: 2603-3259 (Print) ISSN: 2603-3267 (Online), pp. 273-276, Ohrid, North Macedonia, Jun. 2019.
23. S.Vukadinovic, *Elements of Probability Theory and Mathematical Statistics*, Belgrade, 1973.