Available online at www.sciencedirect.com
ScienceDirect
Energy Reports 8 (2022) 612–618
www.elsevier.com/locate/egyr

2021 8th International Conference on Power and Energy Systems Engineering (CPESE 2021), 10–12 September 2021, Fukuoka, Japan

Medium-term load forecasting in isolated power systems based on ensemble machine learning models

Pavel Matrenin[a,d], Murodbek Safaraliev[b], Stepan Dmitriev[b,*], Sergey Kokin[b], Anvari Ghulomzoda[c], Sergey Mitrofanov[a,d]

[a] *Department of Power Supply Systems, Novosibirsk State Technical University, Novosibirsk, Russian Federation*
[b] *Department of Automated Electrical Systems, Ural Federal University, Yekaterinburg, Russian Federation*
[c] *Department of Automated Electric Power Systems, Novosibirsk State Technical University, Novosibirsk, Russian Federation*
[d] *Sirius University of Science and Technology, Sochi, Russia Federation*

Received 26 October 2021; accepted 8 November 2021
Available online xxxx

Abstract

Over the past decades, power companies have been implementing load forecasting to determine trends in the electric power system (EPS); therefore, load forecasting is applied to solve the problems of management and development of power systems. This paper considers the issue of building a model of medium-term forecasting of load graphs for EPS with specific properties, based on the use of ensemble machine learning methods. This paper implements the approach of identification of the most significant features to apply machine learning models for medium-term load forecasting in an isolated power system. A comparative study of the following models was carried out: linear regression, support vector regression (SVR), decision tree regression, random forest (Random Forest), gradient boosting over decision trees (XGBoost), adaptive boosting over decision trees (AdaBoost), AdaBoost over linear regression. Isolation of features from a time series allows for the implementation of simpler and more overfitting-resistant models. All the above makes it possible to increase the reliability of forecasts and expand the use of information technologies in the planning, management, and operation of isolated EPSs. Calculations of the total forecast error have proved that the characteristics of the proposed models are high quality and accurate, and thus they can be used to forecast the real load of a power system.

Peer-review under responsibility of the scientific committee of the 8th International Conference on Power and Energy Systems Engineering, CPESE, 2021.

Keywords: Medium-term forecasting; Electric power system; Ensemble models; Isolated power system

1. Introduction

Operational planning and efficient management of the EPS operation mode is impossible without reliable load forecasting, which is carried out for an onward period from several minutes to several days. Load forecasting is

* Corresponding author.
E-mail address: sa.dmitriev@urfu.ru (S. Dmitriev).

https://doi.org/10.1016/j.egyr.2021.11.175
2352-4847/

Peer-review under responsibility of the scientific committee of the 8th International Conference on Power and Energy Systems Engineering, CPESE, 2021.

necessary to optimize onward modes and adjust the current ones, to consider real-time dispatcher requests related to decommissioning of electric power equipment for repairs, tests, etc. [1].

Load forecasting is carried out based on statistical processing of load data for several preceding days [2]. In terms of forecast depth, forecasts fall into four different categories:

- Very short-term forecasting: Forecasts are made for a few minutes ahead, and the forecast values are transmitted to the control unit for sending in real-time. This type of forecasting is used to quickly respond to intraday fluctuations in electricity demand [3].
- Short-term forecasting: Forecasts are made for the period from several hours to several days ahead, and the results are used for a wide range of decisions related to the technical and economic optimization of the functioning of electric power systems, and related to the commitments for high-quality and reliable electricity supply to consumers [4]. Currently, the importance of on-the-spot and short-term forecasting is increasing for the reasons associated with the complication of electricity transportation and distribution systems, growing requirements for the quality and reliability of power supply, as well as the development of new equipment with automated control systems.
- Medium-term forecasting: In this case, forecasts are made for the period from several hours to several weeks ahead. The resulting forecasts provide information on weekly fluctuations, and this information is mainly used for planning network maintenance, setting electricity prices and agreeing on the power distribution mechanism, etc. [5].
- Long-term forecasting: Forecasts are made for the period from several months to several years ahead, and this information is usually used to estimate the capacity or analyze the need for new transmission lines [6].

Load forecasting models fall into two main categories: the ones using parametric and non-parametric methods. While parametric methods are based on analytical models, nonparametric methods build on artificial intelligence. Although many demand forecast methods have been proposed recently, described below are some of the most widely used models, which the authors build upon in their current research [7–10].

This paper aims to compare machine learning models for medium-term load forecasting in isolated power systems.

The object of the study is an isolated EPS of the Gorno-Badakhshan Autonomous Oblast (GBAO) - a region of the Republic of Tajikistan. In GBAO, the main power consumers are the population, government organizations, and small businesses. The GBAO EPS is operated by the energy company ‘Pamir Energy’, which has 11 hydroelectric power plants (HPPs) with a total capacity of 43.5 MW, and of these, two HPPs with daily regulation are located in different districts of the region. Today, the isolated EPS of GBAO faces serious difficulties associated with a constant shortage of electricity in the winter period (November–March). The main reasons for the electricity shortage at the EPS of GBAO are as follows [11,12]:

- Isolated operation of the power system;
- Limited flow of water in rivers in the winter (low-water) period;
- Lack of seasonal energy storage (reservoirs);
- Growth of electricity consumption by the population during the winter heating season.

In [13], models were developed for short-term forecasting of electricity consumption based on time series for a given EPS; the implementation of those models for medium-term forecasting is problematic due to the fact that the available reservoirs can only provide daily regulation of uninterrupted power supply. The development of a medium-term forecasting model is an urgent task for the formation of planned indicators with a horizon of one week.

For medium-term forecasting, many different methods can be applied, from the simplest regression models to deep recurrent neural networks. Note that the selection of models should be based on the data analysis and feature selection results. Setting hyper-parameters of models can be automated using the random search or grid search. The models themselves and the respective learning algorithms can be found in open-source libraries. Thus, the top priority task is not the building of a prognostic model, but rather the preceding phase of data analysis and feature selection, as shown in this paper.

In this research, the task of forecasting is formulated as follows: to build a model that would transform the input data (features) X into a forecast of daily loads for a week ahead Y *:

$$Y^{*} = f\left(X\right) \tag{1}$$

In this paper, the mean absolute percentage error (MAPE) was chosen to be the indicator of the forecast accuracy:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^* - y_i}{y_i}\right| \tag{2}$$

Where n is the number of hours within the dataset; Y_i^* - forecast load on the ith day; Y_i is the true value.

2. Load forecasting data analysis

The dataset under study contains the maximum daily load values (by days of the week) within one month from November through March, inclusively, for the period of 5 years (2015–2019), 175 values in total (5 years * 5 months * 7 days of the week).

Before building a forecasting model, it is necessary to assess and select the most significant features. In addition to calendar data (year, month, and day of the week), the load of the previous period can also be used for forecasting.

Generally, the processing of retrospective data as a time series to identify complex dependencies can be efficiently performed using recurrent neural networks [14]; but in the task under consideration, the data of the time series is insufficient for the implementation of the recurrent neural networks. From this time series, certain features need to be selected using the analysis of correlation coefficients. For this purpose, there has been prepared a graph of changes in the correlation of the load between a certain specific day and the preceding days (Fig. 1).

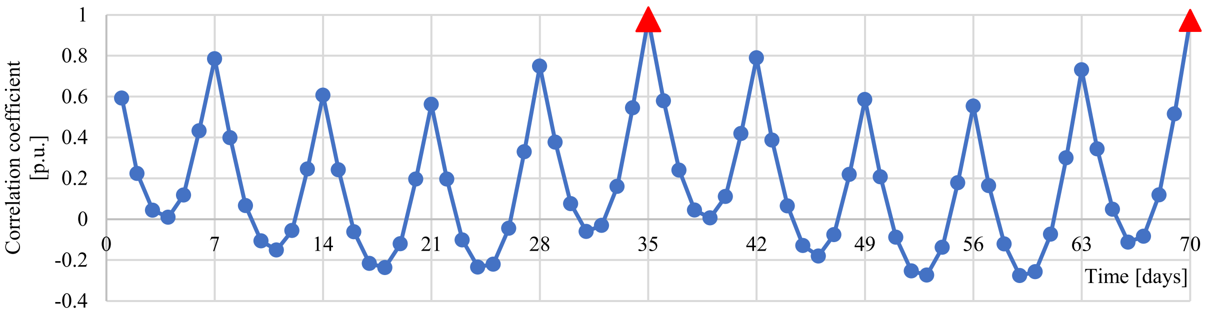

Fig. 1. Correlation between the load on the reference day and the load during the preceding days.

Fig. 1 shows that the highest correlations are observed at the interval of 7 days, which means the similarity between the same days of the week. Note that the two maximum values marked in Fig. 1 fall on the 35th and the 70th days. These are the points where both days of the week and days of the month coincide. For example, for forecasting the value of Tuesday, December 2019, the data of Tuesday, December 2018, and 2017 will be the most useful.

Fig. 2 shows the correlation coefficients between the features in use (y - year, m - month, d – day of the week, p_1 *and* p_2 are the load values for the same day of the last year and the year before last, p is the forecasted load).

The resulting dataset structure and an example of one of its entries are shown in Table 1. Months are numbered from 1 (November) to 5 (December), years are converted to the number of years from the last year of the dataset. The selected features limit the size of the dataset since the preparation of a dataset requires the data for the previous two years. Therefore, the data for the two earlier years (2015, 2016) are used only as features, while the target column p contains the data for the years 2017–2019.

Table 1. The dataset structure and an entry example.

y	m	d	p_1, MW	p_2, MW	p, MW
3	1	1	29.689	30.456	32.452

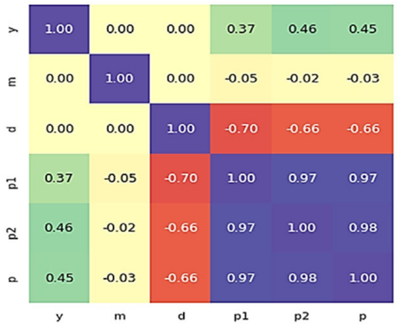

Fig. 2. Correlation coefficients between the dataset features.

3. Building and discussion of forecast models

Since the used dataset is relatively small in terms of both the number of items and the number of features, the models should not be too complex in terms of the number of learning parameters (model size).

The dataset is divided into two parts: the training part and the test part, the ratio between them is 90% and 10%. The test dataset was not used to set up the models; it was used for the final assessment of the already trained models. To set up the hyper-parameters of the models, cross-validation was used: the training dataset is divided into 8 parts, each of them in its turn is used as a validation dataset, while the remaining 7 parts are used as a training dataset.

Thus, each of the models was trained 8 times on different sub-datasets, then checked on validation datasets. The results obtained were averaged. That way, it is possible to reduce the influence of the dataset splitting method on the selection of hyper-parameters of models.

It is important to note that the dataset was not mixed, so the testing was carried out under realistic conditions when the model is trained on historical data and then used in the future on new data.

A comparative study of the following models was carried out: Linear regression; Regression based on the Support Vector Regression (SVR) [15]; Regression Decision Tree [15]; Random Forest [15,16]; Gradient boosting over decision trees (XGBoost) [17]; Adaptive boosting (AdaBoost) [15,18] over decision trees; Adaptive boosting (AdaBoost) [15,18] over linear regressions.

Each of these models is aimed at constructing a regression dependence f between the input parameters $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i = \{year_i, month_i, day_i, p_{1i}, p_{2i}\}$ and the forecasted output power $Y = \{y_1, y_2, \ldots y_n\}$, $y_i = p_i$:

$$Y^* = f\ (w, X)$$

where w are model parameters (regression coefficients, parameters of dividing rules of decision trees, etc.).

The process of training the model comes down to solving an optimization problem as follows:

$$w = \arg\max(Err(Y, f(w, X)))$$

where $Err(Y, f\ (w, X))$ is the forecast error metric used in the learning algorithm.

The hyper-parameters of the models are shown in Table 2, the results obtained for solving the task (1) with criterion (2) are shown in Table 3 and in Fig. 3.

Except for SVR which was used for comparison purposes, all other models are based either on regression equations of the type $f\ (X) = AX + b$, or on decision trees, which are an hierarchy of logical rules. The accuracy of one regression equation or one decision tree is often insufficient. This problem is solved by moving from one model to an ensemble, that is, a weighted sum of models:

$$Y^* = f\ (X) = \sum_{j=1}^{m} c_j f_j\ (X)$$

where c stands for the weights of the models, m is the number of models in the ensemble.

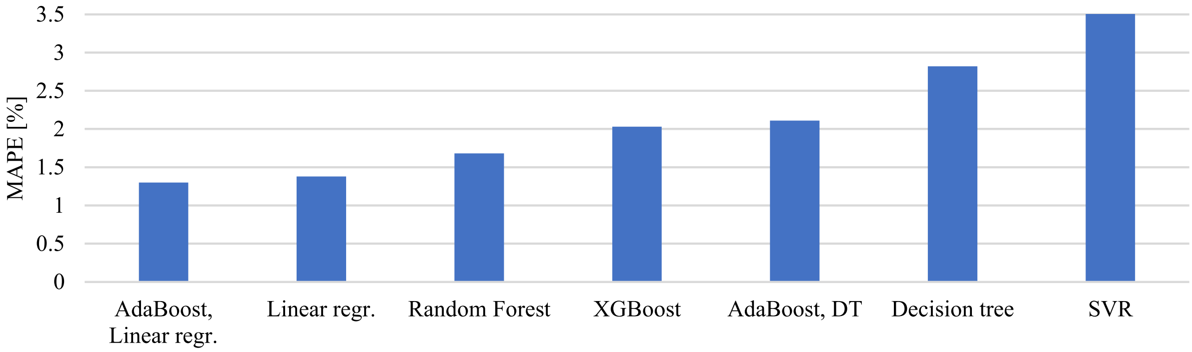

Fig. 3. Results for the test dataset.

Table 2. Hyper-parameters of models.

Model	Hyperparameters
Linear regression	Regularization — no; Offset — yes
SVR	Kernel — Gaussian radial basis function; γ (kernel coefficient) = 0.2; C (regularization) = 1.0
Decision tree	Split strategy — choosing the best; Maximum depth = 5
Random Forest	Ensemble size = 30; Maximum tree depth = 4
XGBoost	Ensemble size = 60; Learning step = 0.1; Maximum tree depth = 2
AdaBoost over decision trees	Ensemble size = 40; Learning step = 0.1; Maximum tree depth = 4
AdaBoost over linear regression	Ensemble size = 4; Learning step = 0.1 Linear regression regularization — no; Linear regression offset — yes

Table 3. Results by model.

Model	MAPE on the training set, %	MAPE on validation set, %	MAPE on the test set, %
Linear regression	1.29	1.40	1.38
SVR	2.89	3.28	3.52
Decision tree	0.45	1.96	2.82
Random Forest	0.73	1.68	1.68
XGBoost	0.71	1.66	2.03
AdaBoost over decision trees	0.56	1.64	2.11
AdaBoost over linear regression	1.32	1.40	1.30

Fig. 4 shows a comparison of the model that produced the best result (AdaBoost over linear regression) and the true values for the entire period of 2017–2019; Fig. 5 shows the same for the year of 2019.

Since AdaBoost is a linear sum of models, the final AdaBoost combining four linear regressions can be collapsed to a linear form $Y^* = f(X) = AX + B$:

$$y_i^* = f(x_i) = a_1 year_i + a_2 month_i + a_3 day_i + a_4 p1_i + a5p2_i + b$$

Obviously, this task does not require complex models, since even the simplest linear regression has produced an accuracy close to the best result. The implementation of more complex models using ensembles of decision trees for the solution of this task has produced the effect of overfitting; their accuracy for the test dataset was significantly lower than that for the training dataset. It can be assumed that the reason is that the problem under consideration

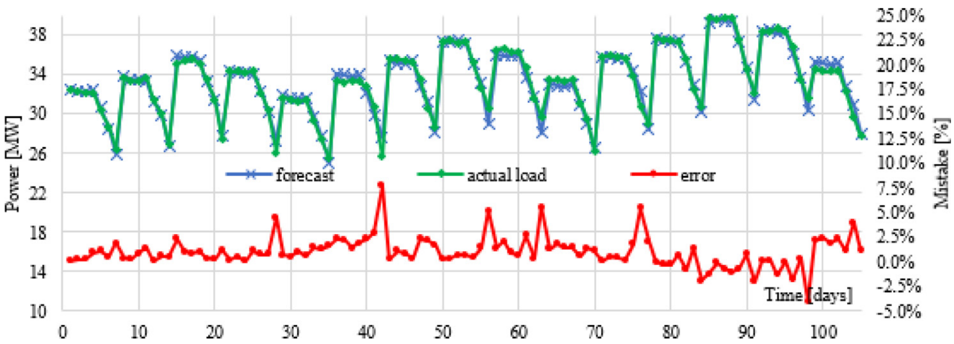

Fig. 4. Comparison of the forecast and the true load values, 2017–2019.

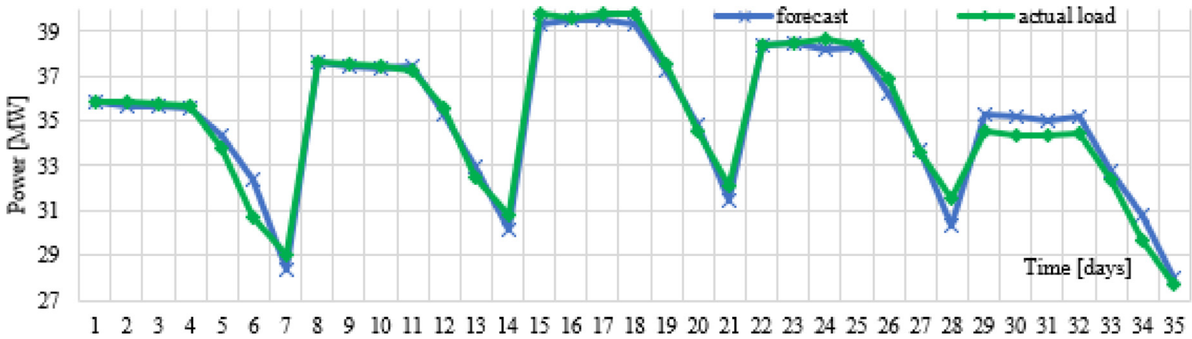

Fig. 5. Comparison of the forecast and the true load values, 2019.

requires rather a continuous form of the dependence of the output power consumption on input factors, primarily power consumption in the past, than a discrete form or dependence in the form of rules for which regression trees and ensembles of trees can be are more effective. At the same time, the best result was nevertheless obtained by the ensemble model, which combined four linear regressions in one model. The resulting error of 1.3% is low enough for the model to be used in planning the operation of the power system. Also, it should be noted that the forecast result was over 2% underestimated for a mere two days from the dataset.

4. Conclusion

This study presents a forecasting method that is based on the selection of the most significant features from the time series and on the ensemble machine learning models, which allows for forecasting the load in the isolated EPS of GBAO for a week ahead. Seven different models were analyzed to forecast the load of the area under study based on previous load consumption data. The best result was obtained with an ensemble model (AdaBoost) that combined four linear regressions in one model.

The results obtained in the course of this study can be used to improve the quality of load forecasting of the GBAO EPS when making informed decisions about the structure of load in the region. In addition, the proposed methods can be recommended for other power supply companies that manage isolated EPSs. The use of the ensemble forecasting model is aimed at increasing the reliability of forecasting in order to ensure high-quality and reliable power supply to consumers in similar isolated EPSs.

The load forecasting method presented in this study opens up the possibility of planning the loading mode of generators in an isolated power system and, consequently, optimizing the operating mode of the generators taking into account the available hydro resources. Based on the daily maximums and the nature of the load changes, one can determine the expected demand for power generation, which can be further redistributed between hydro generators depending on the available hydro resources. The provided results allow us to recommend the ensemble model for medium-term forecasting to optimize the load of generating plants.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The reported study was funded by RFBR, Sirius University of Science and Technology, JSC Russian Railways and Educational Fund “Talent and success”, project number 20-38-51007.

References

[1] Piotrowski P. Forecasting in power engineering in different time horizons. Warsaw: Warsaw University of Technology Publishing House; 2013.
[2] Bartholomew PI, Panikovskaya TYu. Optimization of power systems modes: A tutorial. Ekaterinburg: GOU VPO USTU – UPI; 2007.
[3] Guan C, Luh PB, Michel LD, Friedland PB, Wang Y. Very short-term load forecasting: Wavelet neural networks with data pre-filtering. IEEE Trans Power Syst 2013;28(2013):30–41.
[4] Jiang H, Ding F, Zhang Y. Short-term load forecasting based automatic distribution network reconfiguration. In: Proceedings of the IEEE power & energy society general meeting. 2017.
[5] Ilseven E, Göl M. Medium-term electricity demand forecasting based on MARS. In: Proceedings of the IEEE PES innovative smart grid technologies conference Europe. 2017.
[6] Carvallo JP, Larsen PH, Sanstad AH, Goldman C. Long term load forecasting accuracy in electric utility integrated resource planning. Energy Policy 2018;119:410–22, 2018.
[7] Kuster C, Rezgui Y, Mourshed M. Electrical load forecasting models: A critical systematic review. Sustain Cities Soc 2017;35:257–70, 2017.
[8] Dudek G. Pattern-based local linear regression models for short-term load forecasting. Electr Power Syst Res 2016;130:139–47, 2016.
[9] Verdejo H, Awerkin A, Becker C, Olguin G. Statistic linear parametric techniques for residential electric energy demand forecasting. A review and an implementation to Chile. Renew Sustain Energy Rev 2017;73:4–1122, 2017.
[10] Baliyan A, Gurav K, Mishra SK. A review of short term load forecasting using artificial neural network models. Proc Comput Sci 2015;48:121–5, 2015.
[11] Kirgizov AK, Dmitriev SA, Safaraliev MK, Pavlyuchenko DA, Ghulomzoda AH, Ahyoev JS. Expert system application for reactive power compensation in isolated electric power systems. Int J Electr Comput Eng 2021;11(5):3682–91, (2021).
[12] Matrenin P, Safaraliev M, Dmitriev S, Kokin S, Eshchanov B, A. Rusina. Adaptive ensemble models for medium-term forecasting of water inflow when planning electricity generation under climate change. Energy Rep 2021;7. 2021.
[13] Khudzhasaidov JKh, Rusina AG, Matrenin PV, Dmitriev SA, Safaraliev MKh. Development of models for forecasting power consumption based on time series in isolated power systems. Electrotech Syst Complexes 2020;48(3):23–7, (2020).
[14] Matrenin PV, Manusov VZ, Khalyasmaa AI, Antonenkov DV, Eroshenko SA, Butusov D. Improving accuracy and generalization performance of small-size recurrent neural networks applied to short-term load forecasting. Mathematics 2020;12(8):2169, (2020).
[15] Machine learning in Python. 2020, [Web source]. Available online: https://scikit-learn.org.
[16] Breiman L. Random forests. Mach Learn 2001;4:5–32, (2001).
[17] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016, p. 1–13, [Web source]. Available online: https://arxiv.org/abs/1603.02754.
[18] Drucker H. Improving regressors using boosting techniques. 1997, p. 1–9, [Web source]. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.314&rep=rep1&type=pdf.