

Original Research Article

Analysis of factors and forecasting of the daily maximum electric load of an automobile factory

Vishniakou, U. A., Paloska K. I.

Belarusian State University of Informatics and radioelectronics, Belarus

Abstract: The study examines the factors influencing the daily maximum electrical load of an automobile plant and develops an approach to forecasting this indicator over daily, monthly, and nine-month horizons. The study is based on operational data from the Minsk Automobile Plant for 2024–2025, which includes 731 daily observations, meteorological data, and monthly production indicators, including data from the foundry production. The calculations were carried out using the author's comprehensive statistical analysis (ACSA) system, developed in Python. The system provides data loading, verification, and preliminary processing, as well as correlation and factor analysis, training of predictive models, and adaptive adjustment of the forecast. The initial data revealed seven days with discrepancies in the production schedule and twelve days with a violation of the permissible ratio between electricity consumption and the maximum load. Factor analysis of the daily data ($n=485$, $KMO=0.653$, $p<0.001$) made it possible to identify three factors that explain 62.6% of the variance. Eleven configurations of predictive models, including random forest and ridge regression, have been compared. A scheme with division by day type has been adopted: Ridge regression is used for working days, random forest for non-working days, and adaptive adjustment is applied to both. This combination was chosen due to the statistically indistinguishable accuracy of competing models, as it is more interpretable. The mean absolute percentage error for working days was 4.75% for a one-day forecast and 7.14% for a monthly forecast. Organizational measures to smooth out load peaks have been proposed, and their potential effect has been assessed.

Keywords: daily maximum load; autovobile factory; factor analysis; principal component analysis; ridge regression; power consumption forecasting; declared capacity; comprehensive statistical analysis

1. Introduction

The daily maximum electrical load of an industrial enterprise affects not only the electricity consumption pattern but also the amount of the power charge. With a two-part tariff, a significant portion of the costs is determined by the declared power. Therefore, even a slight reduction in the monthly maximum load can yield an economic effect without reducing production volume. In this regard, it is important, on the one hand, to identify the factors that determine the maximum load, and, on the other hand, to ensure its forecasting: For 24 hours ahead for operational management and for a month ahead for planning the declared power.

Forecasting the load of an individual enterprise has its own specific features compared to forecasting the load of a power system. In a power system, the total load is formed by a large number of consumers and usually has a fairly stable daily and weekly dynamics. At an industrial enterprise, the nature of the load largely depends on the actual production schedule: Working and non-working days, shifts, the operation of energy-intensive equipment, the schedule of smelting operations in foundry production, and decisions made at the departmental level. Therefore, methods that work well for an energy system do not always provide the same level of accuracy when forecasting the load of an individual plant^[1-2].

Furthermore, a high value of the coefficient of determination does not always indicate a high-quality model in itself. In this case, it may largely be determined by the difference between working and non-working days. Therefore, when evaluating models, it is necessary to take into account not only general statistical indicators but also the accuracy of the forecast under conditions of a real production schedule.

In this article the problem is solved using the author's system of comprehensive statistical analysis developed in Python^[3]. The system combines the loading and verification of raw data, their preprocessing, correlation and factor analysis, the construction and comparison of predictive models, including random forest and ridge regression, as well as adaptive forecast correction based on actual data. This approach ensures reproducibility of calculations: When new data is added, the analysis can be repeated using a single algorithm, and all transformations performed are recorded in the log.

The object of research and the information base. The object of the study is the Minsk Automobile Plant, an industrial consumer of electric energy of category 750 kV·A and above. During the period under consideration, the daily maximum load varied from 1,926 to 63,158 kW, and the company's annual electricity consumption in 2025 amounted to 160.4 thousand MWh.

The analysis used data from the period from January 1, 2024, to December 31, 2025. The information base includes three groups of indicators:

- daily operational data for 731 days: Maximum electrical load, electricity consumption volume, and a characteristic of the working day;

- power values in four time zones of the day and the time at which they occur – for February, July, and December 2024 and 2025, a total of 181 days;

- monthly production indicators: Output volume, number of working days, foundry production data, including casting volume, electricity consumption, and maximum load.

- meteorological data for Minsk: Average daily, minimum and maximum air temperature, wind speed, atmospheric pressure and precipitation.

The target indicator is the daily maximum electrical load P_{max} , kW.

Methodology of comprehensive statistical analysis. The calculations were performed using the ACSA, developed in the Python language. The system combines modules for loading, quality control, and preliminary data processing, as well as modules for correlation, factor, regression analysis, and forecasting^[3-4].

At the first stage, the source data were loaded into the system and underwent quality control. The completeness and correctness of the data, permissible value ranges, the presence of duplicates, and adherence to logical relationships between indicators were checked. Subsequently, preliminary processing was carried out: Correcting identified errors, restoring anomalous values, and refining calendar indicators. All performed operations were recorded in the transformation log.

After processing, a single daily data set was formed. Meteorological observations and production indicators were added to the electrical load indicators by date. Monthly indicators were used with a time lag, so only the information available on the date the forecast was compiled was used in building the forecast.

The correlation analysis module made it possible to assess the relationships between the daily maximum electrical load and calendar, production, and meteorological factors. To achieve this, Pearson and Spearman correlation coefficients were calculated, along with statistical significance tests, as well as partial correlation coefficients, taking into account seasonality and the overall level of load.

The factor analysis module was used to reduce the number of interrelated indicators and identify generalized factors. Prior to the analysis, the indicators were standardized, after which the data suitability for factorization was assessed using the Kaiser–Meyer–Olkin (KMO) and Bartlett criteria^[5]. Next, the main components were identified, the required number of factors was determined, and varimax rotation was applied to simplify the interpretation of the results.

The regression analysis and forecasting module provided training and comparison of forecast models for the daily maximum load. A naive model with the load value seven days before the forecast date, a calendar-group model, multiple linear regression, regression based on factor scores, ridge regression with a regularization parameter $\alpha = 0.3$ ^[6], random forest^[7], gradient boosting, and the MLP neural network model were considered. In addition, ensemble models combining regression models with random forest were evaluated. The quality of the models was compared using forecasting accuracy metrics; in cases of similar results, preference was given to the more interpretable model.

To improve accuracy, an adaptive feedback mechanism was used. After obtaining the actual daily maximum value, the forecast bias was adjusted, and new observations were taken into account in subsequent calculations.

The models were evaluated using a sequential retraining scheme without using future information. At the beginning of each month, the models were trained on all data available up to that point and generated a forecast for the following month. The assessment covered the period from February to December 2025 – 334 days, including 222 working days.

Data quality control results. Comparison of daily and monthly data showed that the main aggregated indicators are consistent with each other. The sums of daily electricity consumption and monthly load peaks fully matched the data from the monthly reporting for all 24 months. The number of working days also matched between the daily and monthly data. At the same time, four groups of errors and inconsistencies were identified in the source information.

Descriptive analysis. The distribution of the daily maximum electrical load has a clearly defined two-group structure, related to the division into working and non-working days. On working days, the average maximum is 52,111 kW with a standard deviation of 4,688 kW, while on non-working days it drops to 10,907 kW with a standard deviation of 7,814 kW.

The differences between these groups are statistically significant according to Welch's t-test ($t=72.4$, $p<0.001$). The value $\eta^2=0.914$ indicates that the working day indicator explains more than 90% of the variability in the daily maximum load. Therefore, when evaluating predictive models, it is not sufficient to rely solely on the coefficient of determination calculated for all observations. Its high value may be primarily due to the correct accounting for working and non-working days, rather than to the accuracy of predicting the actual load level within each group.

Correlation analysis. The calculation was performed separately for the entire dataset and for a subset of working days. For the entire sample, the main factor associated with the daily maximum load is the working day indicator: The correlation coefficient is $r=0.957$. This value is higher than the threshold of 0.7 adopted in the system, which corresponds to a strong relationship. Consequently, the difference between working and non-working days largely determines the level of maximum load.

However, when considering only working days, the nature of the relationships changes. The maximum load of the previous workday has the greatest impact: The correlation coefficient is $r = 0.703$. The relationship with past load values gradually weakens as the time lag between days increases: The autocorrelation coefficient is 0.704 for the previous workday.

Factor analysis. To simplify and generalize the original data, factor analysis was carried out separately for two sets of indicators: Daily and monthly data.

Daily data. The daily analysis included 11 indicators calculated based on 485 working days. These included the average daily air temperature, wind speed, atmospheric pressure, amount of precipitation, the maximum load of the previous working day, the average maximum for the last working days, the power consumption of the previous day, and four production indicators.

Initially, 14 indicators were considered. In addition to the variables listed, the set included the minimum and maximum air temperatures, as well as the daily temperature amplitude. However, these indicators were closely related to the average daily temperature. Because of this, the correlation matrix turned out to be practically degenerate, and conducting factor analysis became некорректным. Therefore, the minimum, maximum temperature, and daily amplitude were excluded from further analysis.

After reducing the set of indicators, the quality of the factor model improved. The KMO measure value increased from 0.561 to 0.653, and the determinant of the correlation matrix was 0.00648. This made it possible to use factor analysis for further research.

The suitability of the data is confirmed by the results of statistical criteria: $KMO=0.653$, and the partial indicators of variable adequacy ranged from 0.544 to 0.887. The Bartlett sphericity test also showed the presence of statistically significant relationships between the indicators: $\chi^2=2416.2$, degrees of freedom – 55, $p<0.001$. The KMO value corresponds to an acceptable level. Its relatively low value is due to the fact that some meteorological indicators have weak relationships with other variables.

Based on the analysis results, three factors were identified with eigenvalues of 3.456, 1.872, and 1.556. Together, they explain 62.6% of the total variability of the original indicators. The obtained factors can be interpreted as follows.

The first factor, F1, is production activity; it explains 26.95% of the total variability of the original indicators. The largest contributions to it come from the volume of casting output, the number of trucks

produced, and the value of output per working day. This factor characterizes the intensity of the company's production program, which is usually planned in advance for a month.

The second factor, F2, is seasonality and stability of the load level; it explains 20.36% of the variance. It includes the maximum load of previous working days and the air temperature, with the temperature having an inverse relationship with the load: During the cold period, the maximum load is usually higher. This factor is most closely related to the daily maximum electrical load: The correlation coefficient is $r=0.633$, whereas for the production factor it is $r=0.258$. Therefore, when making forecasts, it is especially important to take into account the achieved load level on previous days and the temperature conditions.

The third factor, F3, is weather conditions not related to temperature; it explains 15.28% of the total variability in the indicators. It is formed by precipitation, wind speed, and low atmospheric pressure. However, this factor is practically not related to the daily maximum load ($r = -0.076$). This means that for the enterprise in question, weather conditions other than air temperature do not have a noticeable impact on the magnitude of the daily maximum.

2. Conclusion

(1) An analysis of the maximum electrical load of the Minsk Automobile Plant for 2024–2025 was carried out. Data for 731 days were studied. For this purpose, the author's KSA system was used, which allows to sequentially upload data, check it, perform calculations, build forecasts, and generate recommendations.

(2) During the data quality check, errors of two types were identified: In 7 cases, the production calendar did not match the actual work schedule; in 12 cases, the data on daily electricity consumption did not correspond to the maximum load. All errors were identified, corrected, and recorded. The factors that have the greatest impact on the maximum load on working days are: The load level on previous days; air temperature and heating mode. In cold weather, the load increases due to heating and ventilation.

(3) The analysis showed that three main groups of factors influence the daily maximum load: The production program; seasonality, temperature, and accumulated load level; sudden weather changes. Together, these factors explain about 63 % of the differences in daily maximum loads. At the monthly level, the maximum is associated with temperature and the operation of the foundry. However, no direct confirmed link has been identified between the monthly maximum and the total output volume. This does not mean that there is no link at all: The available 24 months of observations are insufficient for a definitive conclusion.

(4) Eleven forecast model variants were compared. The combined model showed the best results: A regression model is used for working days; a random forest model is used for non-working days; and a mechanism for adapting to the latest load changes is used for the daily forecast. The average forecast error for the next working day was 4.75%, or approximately 2,275 kW. For the monthly maximum forecast, the error was 1.14%, or about 350 kW.

(5) Important conditions for the model's operation have been identified: Monthly production data must be used with a two-month delay so that this information is already available at the time of the forecast; the fast adaptation mechanism works well for the next-day forecast but is not suitable for the monthly forecast; the forecast accuracy hardly changes in the absence of weather data. This means that the system can operate without constant dependence on external weather services.

(6) Six measures have been proposed, most of which can be implemented organizationally without major investments: To start equipment gradually in the morning, to shift part of the loads to the night, to limit secondary equipment on high-risk days, to manage heating and ventilation in winter, to improve the accuracy of capacity requests, and to enhance the quality of the source data. According to preliminary estimates, these measures may reduce the reported capacity by 1.4–5.0 %. This could result in savings of 0.34 to 1.17 million rubles per year (\$ 0,111–\$0,39 million), solely due to reduced power charges. However, before implementation, each measure must be checked taking into account the actual technological modes and equipment capabilities.

References

- [1] Hong T., Fan S. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*. – 2016. – Vol. 32, N 3: 914–938.
- [2] Nadtoka, I. I., Sedov A. V. Forecasting electricity consumption of industrial enterprises using neural networks. *Proceedings of Higher Educational Institutions. Electromechanics*. – 2018. – No. 4: 78–85.
- [3] Vishniakou, U. A., Paloska K. I. Software tools for automating calculations for factorial, regression, and correlation analyses. *System Analysis and Applied Informatics*. – 2026. – No 1: 82–87.
- [4] Vishniakou, U.A., Khatskevich G.A., Paloska K.I. Comprehensive analysis for assessing the use of Internet of Things networks in the educational process at the University of Informatics and Radioelectronics. *Statistics and Economics*. – 2025. – Vol. 22, No. 4: 52–60.
- [5] Kaiser, H. F., Kennard R. W. An index of factorial simplicity. *Psychometrika*. – 1974. – Vol. 39, No 1: 31–36.
- [6] Hoerl, A. E. Ridge regression: biased estimation for nonorthogonal problems / A. E. Hoerl, // *Technometrics*. – 1970. – Vol. 12, No 1: 55–67.
- [7] Breiman, L. Random forests. *Machine Learning*. – 2001. – Vol. 45, No 1: 5–32.
- [8] Diebold, F. X., Mariano R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*. – 1995. – Vol. 13, No 3: 253–263.