Detection of Anomalous Days in Energy Demand using Leading Point Multi-Regression Model

Krzysztof Karpio[1][0000-0003-0621-3499], Piotr Łukasiewicz[1][0000-0002-1169-3282],

[1] Institute of Information Technology, Warsaw University of Life-SGGW, Nowoursynowska 159, 02-787 Warsaw, Poland
krzysztof_karpio@sggw.edu.pl, piotr_lukasiewicz@sggw.edu.pl

Abstract. Leading Point Multi-Regression Model was utilized to detect days with abnormal energy consumption profiles. They were identified based on the statistical analysis of relative errors of the model. Two ranges of error values were identified: above 4.98% and 3.88% – 4.98%. All days with anomalous energy consumption profiles were identified as major religious holidays in Poland: Easter, All Saints, and Christmas Eve, as well as days related to a celebration of the new year: New Year's Eve and New Year.

Keywords: Leading Point Multi-Regression Model, energy consumption, untypical daily profiles

1 Introduction

In recent years, new methods of electricity consumption analysis have been proposed. They go beyond standard methods such as Holt-Winters or ARIMA [1,2]. More advanced methods, such as nonlinear machine learning (ML), have been utilized in order to forecast power consumption. Among those methods are KNN (K-nearest neighbors) [3], SVM (support vector machine) [4], GBM (gradient boosting machine) [5], RF (random forest) [6], and ANN (artificial neural networks) [7]. An analysis of energy consumption in order to identify sources of energy demand and other factors influencing the consumption, including weather, season, and economics gained importance [8]. One of the most important aspects of the analysis of energy consumption is peak identification. Artificial neural networks were used in [7], while extended CART trees and the K-Nearest Neighbors classifiers were utilized in [3]. In turn, generalized combined additive models and deep ANN were adopted in [9]. Another important issue in a modeling of the power demand is an identification of outliers [10]. In [11] the hybrid model combining Long Short Term Memory (LSTM) and the K-means algorithm was used, while in [12], the authors used a combination of the deep learning model Transformer and a clustering approach based on K-means. Other advanced methods were described in [13,14].

In this paper, we deal with a detection of outliers in terms of daily energy consumption profiles. In contrast to the majority of studies [14, 15], we do not start with the possible reasons for an untypical profile and do not verify them. We use the Lead-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

ing Points Multi-Regression model (LPMR) [16], which is solely based on the energy consumption during a few chosen hours. It does not use any other variables, and it's precision is very high, regardless of other factors, like weather conditions, season. In order to detect outliers, we used the error measure of the model and limit values of errors that were defined precisely.

2 Leading Points Multi-Regression Model

The purpose of LPMR is to model hourly energy demands for the whole day using only a few variables, such as energy consumption at chosen hours. However, in this work, we use the model for other purposes: to detect untypical days from the energy consumption point of view. The model's details are presented and discussed in [16].

2.1 Data

These studies were carried out based on the data regarding total electricity consumption in the Polish power system [17]. The consumption was denoted in MWh on an hourly basis from 1 Jan 2008 to 31 Dec 2020. The data being analyzed contained 4,749 days, which corresponded to 113,976 hours. While our model was solely based on the energy consumption, additional factors were used in the discussion of the results to distinguish separate days and hours, such as a day of the week, specific dates, working hours, etc. The data set was divided into two subsets: the training set and the testing set, consisting of 1583 and 3166 days, respectively.

2.2 Model, errors and variable selection

Data is analyzed on the daily basis: 24 time series of hourly electricity consumptions. We use 24 variables: $E(h_m) = \big(E_1(h_m), \dots, E_i(h_m), \dots, E_N(h_m)\big)$, where $h_m \in \{h_1, h_2, \dots, h_{24}\}$, $E_i(h_m)$ is an electricity consumption at hour h_m in the *i*-th day, and *N* is the total number of analyzed days. The energy consumption is described by the multiple equation linear regression model of the form:

$$E(h_p) = a_{0p} + a_{1p}E(h_1) + a_{2p}E(h_2) + \cdots + a_{kp}E(h_k) + \xi_p \tag{1}$$

where $p \in \{1, \dots, 24\} \setminus \{1, \dots, k\}$ and $a_{0p}, a_{1p}, \dots, a_{kp}$ are model parameters. The number of equations is related to the number of describing variables. In the case of *k* variables, the model consists of 24 – *k* equations.

Model (1) takes electricity consumptions at hours $h_1, h_2, \dots, h_k$ and use them to describe electricity consumptions at the remained hours. The selection of variables is based on the analysis of the random components ξ_p. For each model equation, one calculates the standard deviation of the residuals:

$$\sigma(h_p) = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N}\big(E_i(h_p) - \hat{E}_i(h_p)\big)^2} \tag{2}$$

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

where, $\hat{E}_i(h_p)$ denotes theoretical value and $E_i(h_p) - \hat{E}_i(h_p) = \xi_p$. The quality of the model regressions is measured by means of the relative standard deviation:

$$\nu(h_p) = \sqrt{\frac{1}{N-k-1}\sum_{i=1}^{N}\left(E_i(h_p) - \hat{E}_i(h_p)\right)^2} \Big/ \overline{E(h_p)} \qquad (3)$$

where, standard deviation of residuals is divided by the mean electricity consumption. The quality of the whole model (1) is measured based on the mean values of the measure (4) calculated for all 24 – k equations in the model:

$$MRSD = \frac{1}{24-k}\sum_{p=1}^{24-k}\nu(h_p) \qquad (4)$$

The independent variables are selected by the algorithm in successive steps. In each step, one describing variable $E(h_i)$ is chosen, the new model is built and its precision is evaluated by the $\sigma(h_p)$ and $MRSD$ measures. The procedure can be stopped after reaching a desired precision. An algorithm of variable selection is described precisely in [16]. During the model's construction, we observed a strong decrease of errors in steps 1 – 4. Already in step two, the error decreased below 3%, and in step four, it was below 2%. Subsequent declines are not so significant.

2.3 Application of the model

We concluded that four variables are sufficient to describe a data with reasonable quality, $MRSD = 1.74\%$. They corresponded to energy consumptions at hours: 14, 20, 2, 18. The model quality was evaluated for each data point (hour) and for every day in the testing data set, using the relative measure:

$$RSD(i) = \sqrt{\frac{1}{20}\sum_{p=5}^{24}\left(E_i(h_p) - \hat{E}_i(h_p)\right)^2} \Big/ \frac{1}{20}\sum_{p=5}^{24}E_i(h_p) \qquad (5)$$

where in the sum, the following hours are omitted: from h_1 to h_4, i = 1, 2, …, N, and N denotes the number of analyzed days. For illustration of the model precision one shows empirical and theoretical daily time series for selected six days in Figure 1.

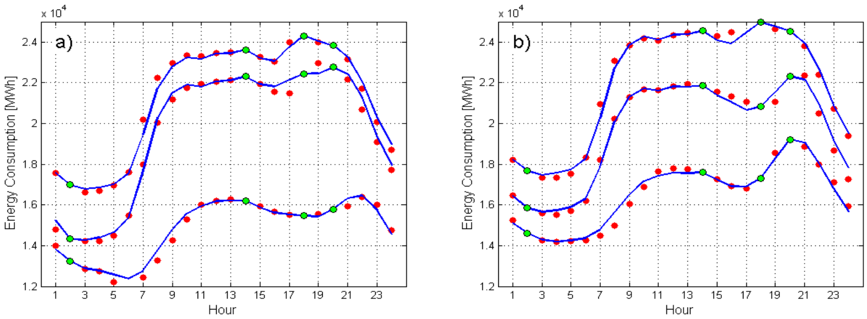

Fig. 1. Empirical (red dots) and theoretical (blue lines) time series for six selected days. From the top: a) 2016-01-14, 2016-02-22, 2016-07-17; b) 2017-12-14, 2017-09-12, 2017-10-08. Green dots denote independent variables of the model.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

For the days presented, the error was between 0.0118 and 0.0178. The model exhibits very good agreement with the data; the mean *RSD* is 0.0175 and for about 90% of days it does not exceed 0.0251.

3 Analysis of anomalyous profiles

3.1 Daily errors of the model

In Figure 2, logarithms of the daily *RSD* errors are shown for the whole testing data set. They do not exhibit any trend and are symmetric around the mean value.

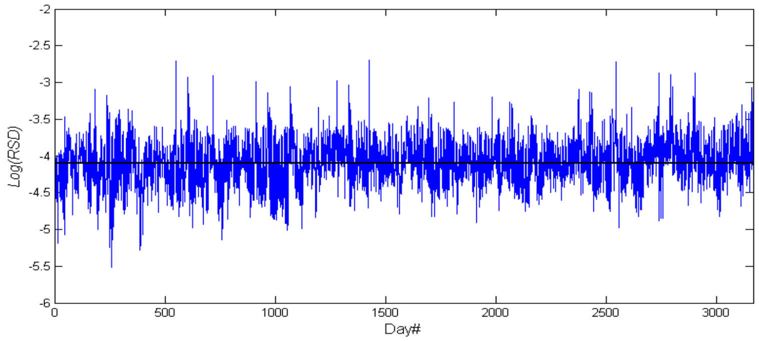

Fig. 2. Logarithms of the daily errors of the model for testing the data set. The mean value of the errors is indicated by the horizontal line at −4.102.

A comparison of the distribution of log-*RSD* with a normal distribution is shown in Figure 3 a) in vertical log scale. The Q-Q plot in Figure 3 b) demonstrates a good agreement between empirical and Gaussian distributions.

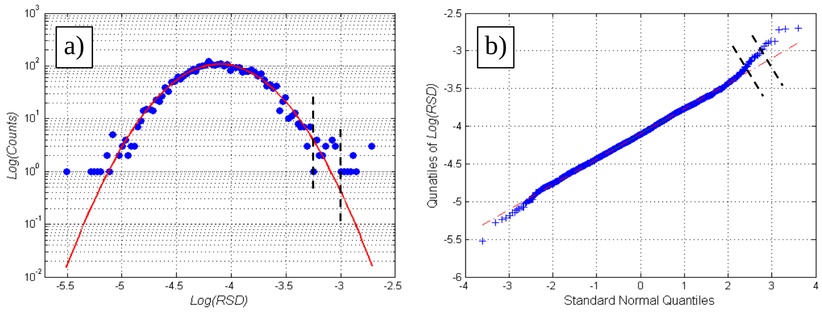

Fig. 3. The compatibility of error and normal distributions: a) distributions of log-errors with normal curve on a logarithmic vertical scale; b) comparison of empirical and theoretical quantiles. The black dashed lines indicate threshold values of errors.

However, there is a visible deviation at the right tail, the number of counts is higher than for the theoretical distribution. Those days are anomalous, characterized by their

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

untypical daily profiles of energy consumption. Moreover, the distortions from normal distribution may indicate the existence of additional factors influencing data apart from statistical fluctuations.

3.2 Identification of unusual daily energy consumption profiles

We identified two ranges of big values of relative errors, based on the analysis of their distribution.
(1) $0.0498 < RSD$ $(-3 < \log(RSD))$ and
(2) $0.0388 < RSD \leq 0.0498$ $(-3.25 < \log(RSD) \leq -3)$.
Boundaries between ranges are indicated by vertical dashed lines in Figure 4. We observe deviations from the normal distributions in both ranges. However, in the first range, the normal distribution is negligible, while in the second range, we may expect some days distributed according to Gaussian. All days from both ranges are listed in Table 1. Weekdays are numbered from 1 (Monday) to 7 (Sunday).

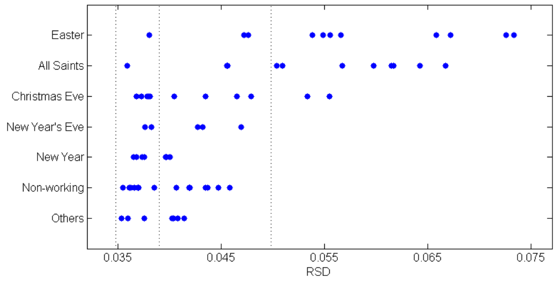

Fig. 4. Identified days corresponding to the biggest errors. Three ranges of error values are indicated by dashed lines.

All days discussed below are shown in Figure 4 with their *RSD* errors. The third group of days with slightly smaller errors, in the range $0.0353 < RSD \leq 0.0388$, were also added to the plot. The whole plot contains a total of 45 days, including all the: Easters, All Saints Days, Christmas Eves, New Year's Eves, and New Year's Days. There are 11 days in the first range with the greatest errors, above 0.0498. The second group contains 19 days with errors between 0.0388 and 0.0498. All the points in the first group are related to the three biggest religious holidays in Poland: Easter, All Saints, and Christmas Eve. The last one is working day, but the working hours are usually shortened. Those days also exist in the second group, predominantly in its upper region. There are also New Year's Eve and New Year in the second group. We got some non-working days that are not holidays but are not random, e.g., a day before Christmas Eve, which was a Sunday; a second day of Christmas, which is a paid holiday in Poland; or Easter Monday. Figures 5 and 6 show daily profiles for days with the greatest values of errors, which are All Saints, Easter, New Year's, and Christmas Eve. Untypical profiles are accompanied by profiles for adjacent or corresponding days. Solid lines denote theoretical values, and green dots correspond to the four independent variables.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

Table 1. Days from both ranges of the big measure values (see text).

No	Date (Weekday)	*RSD*	Description
1	2016-03-27 (7)	0.0672	Easter
2	2013-11-01 (5)	0.0667	All Saints
3	2019-04-21 (7)	0.0658	Easter
4	2019-11-01 (5)	0.0567	All Saints
5	2020-04-12 (7)	0.0566	Easter
6	2012-04-08 (7)	0.0555	Easter
7	2019-12-24 (2)	0.0555	Christmas Eve
8	2014-04-20 (7)	0.0548	Easter
9	2013-12-24 (2)	0.0533	Christmas Eve
10	2015-11-01 (7)	0.0509	All Saints
11	2014-11-01 (6)	0.0504	All Saints
12	2015-12-24 (4)	0.0479	Christmas Eve
13	2015-04-05 (7)	0.0472	Easter
14	2019-12-31 (2)	0.0469	New Year's Eve
15	2020-12-24 (4)	0.0465	Christmas Eve
16	2013-12-25 (3)	0.0458	Non-working
17	2018-11-01 (4)	0.0456	All Saints
18	2012-11-01 (4)	0.0455	All Saints
19	2018-12-23 (7)	0.0437	Non-working
20	2014-12-24 (3)	0.0435	Christmas Eve
21	2018-12-31 (1)	0.0432	New Year's Eve
22	2013-12-31 (2)	0.0427	New Year's Eve
23	2012-12-23 (7)	0.0420	Non-working
24	2012-04-09 (1)	0.0419	Non-working
25	2020-04-06 (1)	0.0414	Other
26	2017-10-05 (4)	0.0408	Other
27	2016-12-24 (6)	0.0404	Christmas Eve
28	2020-04-07 (2)	0.0404	Other
29	2019-12-20 (5)	0.0402	Other
30	2019-01-01 (2)	0.0400	New Year

We limit a discussion to only the days mentioned above. The daily profiles for each type of day are very similar to one another, so the presented profiles in Figures 5, 6 can be treated as representative. (1) Easter: when compared to other Sundays, a profile is more flattened. There is a clear maximum between 9:00 and 11:00, followed by a long slow decrease. (2) All Saints: is compared to adjacent days. The first maximum is moved to the left; we also observe a significantly more flattened profile before 17.

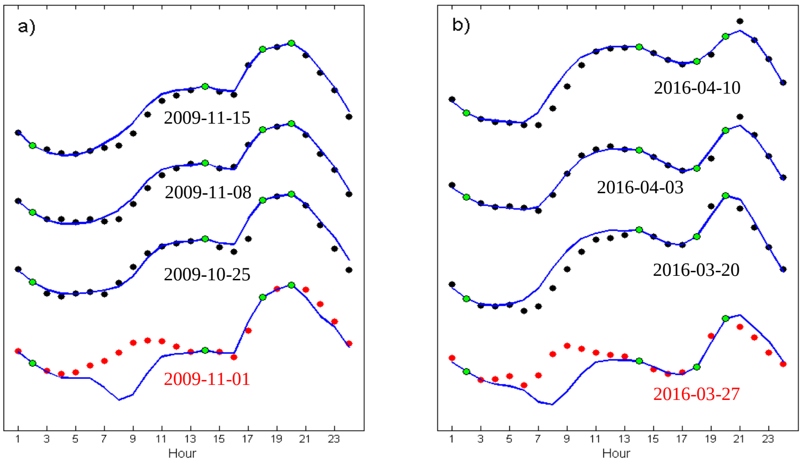

Fig. 5. Daily energy consumption profiles for a) All Saints (red dots *RSD* = 0.0642) and adjacent days, b) Easters (red dots *RSD* = 0.0672) and other Sundays.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

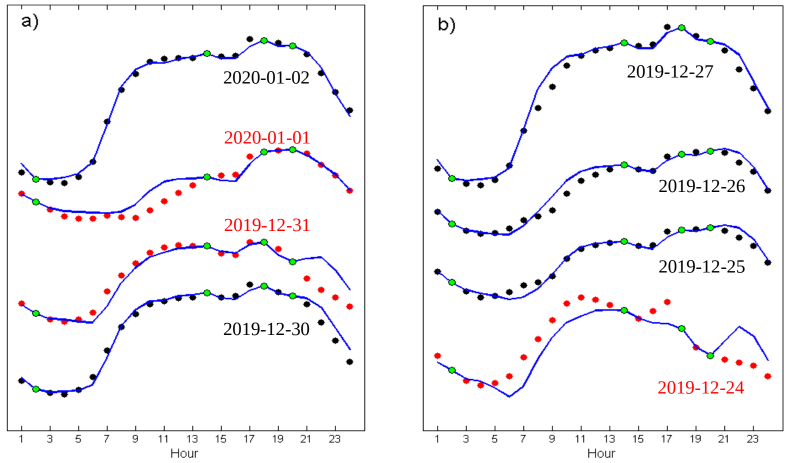

Fig. 6. Daily energy consumption profiles: a) New Year's Eve (red dots, *RSD* = 0.0469) and New Year (red dots, *RSD* = 0,0375) and adjacent days; b) Christmas Eve (red dots, *RSD* = 0.0555) and the following days.

(3) Christmas Eve: compared to the next three days. The analyzed profiles exhibit clear peaks around 10-12 and 17, followed by an anomalous drop of convex shape. (4) New Year's Eve: compared to the profiles for Dec 30th and Jan 2nd. Profile is in general similar to them. However, we observe a different drop in consumption after around 18-19 in both cases. (5) New Year: compared to the profiles for Dec 30th and Jan 2nd. There is a semi-flat shape with no maximum before 17 o'clock, completely different than for other days. For all the days in the upper range, the social factors related to the Easter, All Saints, and Christmas Eve holidays exist and significantly influence energy consumption profiles. Due to the common celebration of those holidays in Poland, one can assume that these factors are related to the short-term, intensive migration of people.

4 Summary

The LPMR-based method of identification of days with untypical daily profiles of energy consumption was presented herein. The following conclusions have been drawn out: (1) Analyzed data could be described with high precision using four independent variables; (2) the distribution of model's errors follow a Gaussian distribution with a high accuracy; (3) days with untypical energy consumption profiles were precisely defined, as days with errors deviating from Gaussian distribution; (4) untypical days were identified as the major religious holidays in Poland: Easter, All Saints, and Christmas Eve; (5) there were also New Year's Eve and New Year identified in the range of smaller errors; and (6) the main factor causing anomalies in the daily profiles was related to the short-term migration of people. The future research will focus on the quantitative description of anomalies in the daily energy consumption profiles as

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53

well as the investigation of their reasons. The studies presented herein can be easily extended to other countries and regions. This subject of studies is of great interest because faults in the energy consumption predictions cause non optimal energy production.

References

1. Chicco, G., Mazza, A.: Load profiling revisited: Prosumer profiling for local energy markets. In: Pinto, T., Vale, Z., Windergrean, S. (eds.) Local Electricity Markets, Cambridge, MA, USA, 215-242, (2021).
2. Karpio, K.; Łukasiewicz, P.; Nafkha, R. Regression Technique for Electricity Load Modeling and Outlined Data Points Explanation. In: Peja´s, J., El Fray, I., Hyla, T., Kacprzyk, J. (eds.) Advances in Soft and Hard Computing; Advances in Intelligent Systems and, Computing Springer: Cham, Switzerland, Volume 889, 56–67 (2019).
3. Gajowniczek, K., Nafkha, R., Ząbkowski, T.: Seasonal Peak Demand Classification with Machine Learning Techniques, International Conference on Applied Mathematics & Computer Science (ICAMCS), Paris, France, 101-1014, (2018).
4. Niu, D., Wang, Y., Wu, D.D.: Power load forecasting using support vector machine and ant colony optimization. Expert Syst. Appl. 37, 2531–2539 (2010).
5. Massaoudi, M., Refaat, S.S., Chihi, I., Trabelsi, M., Oueslati, F.S., Abu-Rub, H.: A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting. Energy 214, 118874 (2021).
6. Huang, N., Lu, G., Xu, D.: A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. Energies 9, 767 (2016).
7. Gajowniczek, K., Nafkha, R., Ząbkowski, T.: Electricity peak demand classification with artificial neural networks, Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 307-315 (2017).
8. Hong, T., Fan, S.: Probabilistic electric load forecasting: A tutorial review. Int. J. Forecast. 32, 914–938 (2016).
9. Berrisch, J., Narajewski, M., Ziel, F.: High-resolution peak demand estimation using generalized additive models and deep neural networks. Energy and AI 13, 100236, (2023).
10. Berthold, M.R.; Borgelt, C.; Höppner, F.; Klawonn, F.; Silipo, R. Guide to Intelligent Data Science: How to Intelligently Make Use of Real Data; Springer: Cham, (2020).
11. Chahla, C., Snoussi, H., Merghem, L., Esseghir, M.: A Novel Approach for Anomaly Detection in Power Consumption Data. In: De Marsico, M., Sanniti di Baja, G., Fred, A. (eds.) International Conference on Pattern Recognition Applications and Methods - ICPRAM, 2019, vol. 1, February 19-21, Prague, Czech Republic, 483-490 (2019).
12. Zhang, J., Zhang, H., Ding, S., Zhang, X.: Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. Frontiers in Energy Res. 9, 779587 (2021).
13. Fu, T., Zhou, H., Ma, X., Hou, ZJ., Wu, D.: Predicting peak day and peak hour of electricity demand with ensemble machine learning. Front. Energy Res. 10, 944804 (2022).
14. Zhang, W., Dong, X., Li, H., et al.: Unsupervised Detection of Abnormal Electricity Consumption Behavior Based on Feature Engineering. IEEE Access 8, 55483-55500 (2020).
15. Hu, M., Ji, Z., Yan, K., et al.: Detecting Anomalies in Time Series Data via a Meta-Feature Based Approach. IEEE Access 6, 27760–27776 (2018).
16. Karpio, K., Łukasiewicz, P., Nafkha, R.: New Method of Modeling Daily Energy Consumption. Energies 16(5), 2095 (2023).
17. Polskie Sieci Elektroenergetyczne, http://www.pse.pl, last accessed 2023/01/15

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_53