Analysis of Smart Meter Data for Electricity Consumers

Grzegorz Dudek, Anna Gawlak, Mirosław Kornatka
Faculty of Electrical Engineering
Czestochowa University of Technology
Czestochowa, Poland
dudek(gawlak,kornatka)@el.pcz.czest.pl

Jerzy Szkutnik
Faculty of Management
Czestochowa University of Technology
Czestochowa, Poland
szkutnik@el.pcz.czest.pl

Abstract— **Smart meter systems are being deployed to improve grid reliability and promote energy efficiency while providing improved services to their customers. Smart metering which is installed in millions of households worldwide provides utility companies with real-time meaningful and timely data about electricity consumption and allow customers to make informed choices about energy usage. Smart meter data analytics has become an active area in research and industry. It aims to help utilities and consumers understand electricity consumption patterns. This paper provides analysis methods for load data including: analysis of daily load profiles and similarity between them, analysis of load density, and analysis of seasonal and irregular components in the load time series. We evaluate our approach by analyzing smart meter data collected from 1000 households in Poland at a 15-minute granularity over a period of one year.**

***Index Terms*-- Smart metering, Energy consumption analysis, Smart meter data analytics.**

I. INTRODUCTION

The combination of the smart meters, bi-directional communication network and data management system, constitute the advanced metering infrastructure (AMI). It plays an increasingly important role in modern power delivery systems by recording the load profiles of customers and facilitating two-way information flow, as well as improving grid reliability and promoting energy efficiency. The benefits of smart metering installations are numerous for many different stakeholders of the systems. Some of the benefits related to data obtained from smart meters are [1]: better access and data to manage energy use, more accurate and timely billing, improved outage restoration, power quality data, early detection of meter tampering and theft, data for improved efficiency, reliability of service, losses, and loading, improved data for efficient grid system design, power quality data for the service areas, and improved customer premise safety and risk profile.

There are serious challenges ahead of the future smart grid related to acquiring and analyzing massive amounts of data of integrated devices, such as distributed storage, intelligent loads, and distributed energy resources. Huge amounts of data from smart meters used for monitoring and control purposes need to be sufficiently managed to increase the efficiency, reliability and sustainability of the smart grid, to provide better understanding of customer behavior and assist in defining electric tariffs. This big data challenge requires advanced methods and infrastructure to deal with huge amounts of data and their analytics [2].

Analytics is the scientific process of transforming data into insights for making better decisions. Three categories of data analytics are defined [3]: descriptive analytics (what do the data look like), predictive analytics (what is going to happen with the data), and prescriptive analytics (what decisions can be made from the data). Following this three categories of analytics we identify the key application areas in smart grid as load analysis, load forecasting, and load management.

This work is focused on descriptive analytics in smart grid which leads to better understanding of the volatility and uncertainty of the load profiles. Descriptive analytics systems allow to describe specific characteristics of a household from its electricity consumption. Such a system based on supervised machine learning in [4] is proposed. It recognizes characteristics capturing socio-economic status of the household, dwelling properties, behavior and appliance stock. Other fields of applications of descriptive analytics in smart grid are [3]:

- bad or missing data detection and data imputation [5],
- energy thief detection. Two approaches are applied for this purpose: supervised [6] and unsupervised learning [7],
- load profiling referring to the classification of load curves or consumers according to electricity consumption behaviors. Clustering methods are used for these purpose: direct [8] and indirect [9] ones.

This paper provides analysis methods for household load data including: analysis of daily load profiles and similarity between them, analysis of load density, and analysis of seasonal and irregular components in the load time series.

II. DESCRIPTIVE ANALITICS OF SMART METERS DATA

The dataset used in this study includes smart meter data for 1000 household customers from the period of one year. The granularity of data is 15-min. The 15-min energies recorded by the smart meters are converted into load to facilitate further analysis. Fig. 1 shows 15-min loads for exemplary customer (customer X). As we can see from this figure the load time series for a household seems to have a strong irregular component manifested by many spikes. Also it reveals peaks and valleys associated with the daily activity of the customer. The strength of both components: irregular and seasonal, is analyzed in subsection II.C. Daily load profiles and similarity between them are analyzed in subsection II.A, and load density profiles are analyzed in subsection II.B.

A. Analysis of Daily Load Profiles

The daily load profiles of the household is characterized by high variability. This is illustrated in Fig. 2 where daily profiles of the exemplary household (household X) from annual period are shown. As we can see the median profile (marked in white in Fig. 2) does not reflect the load nature, ignoring characteristic spikes which correspond to turning on the devices. In Fig. 2 also 0.05 and 0.95 quantiles are shown which give a fuller picture of the profile variability.

As a similarity measure between daily load profiles two measures are used:

- Pearson correlation coefficient r:

$$r = \frac{\sum_{i=1}^{m}(L_{A,i} - \bar{L}_A)(L_{B,i} - \bar{L}_B)}{\sqrt{\sum_{i=1}^{m}(L_{A,i} - \bar{L}_A)^2}\sqrt{\sum_{i=1}^{m}(L_{B,i} - \bar{L}_B)^2}} \quad (1)$$

- Euclidean distance d:

$$d = \sqrt{\sum_{i=1}^{m}(L_{A,i} - L_{B,i})^2} \quad (2)$$

where m = 96 is the number of 15-min periods in the daily period, A and B are the daily period symbols, $L_{A,i}$ is the load at i-th 15-min period of the day A, and $\bar{L}_A$ is the average load for the day A.

Fig. 3 shows distributions of similarity measures r and d between load profiles for two successive Mondays, …, Sundays (customer X). The greatest similarity (highest r, lowest d) between Tuesday profiles is observed and the lowest for Fridays, Saturdays and Sundays. For Mondays the results of both methods are not conclusive. According to correlation coefficient Mondays are less similar to each other when comparing to other working days Tuesday-Thursday. But according to the Euclidean distance similarity between Mondays does not differ from similarity between Tuesday-Thursday and is higher than similarity between Fridays, Saturdays and Sundays.

The Euclidean distance is dependent on the scale, so we cannot compare consumers differing in the level of electricity consumption using this measure.

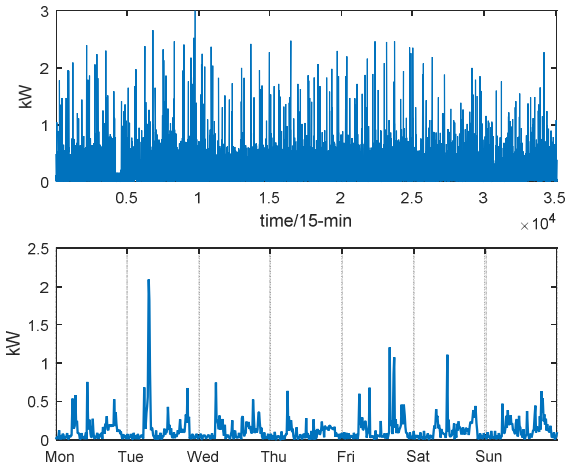

Figure 1. Load of the customer X in one year and one week periods.

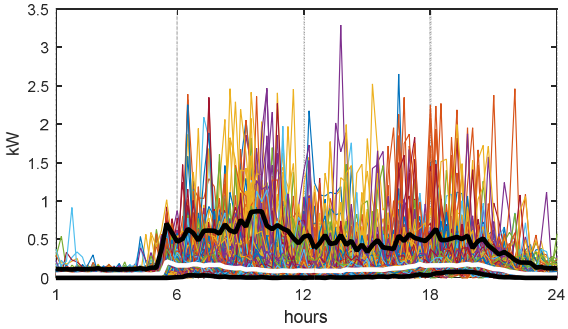

Figure 2. Daily load profiles of customer X (median in white, 0.05 and 0.95 quantiles in black).

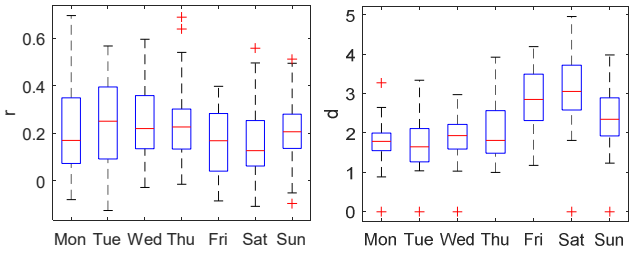

Figure 3. Box-and-whisker plots of the correlation coefficients and Euclidean distances between load profiles for two successive Mondays, …, Sundays (customerX).

The distribution of the mean values of correlation coefficient calculated for 1000 customers according to the above described procedure in Fig. 4 is shown. As we can see from this figure the correlation between profiles of successive days representing the same day of the week is not high for most customers (the median of r is 0.21, and its mean is 0.25). This indicates very weak or no linear relationship between profiles. Therefore, one can infer low similarity between load profiles.

In Fig. 5 similarity measures are shown between different days of the week for customers X and Y (mean values of r or d between day of type P (Monday, ..., Sunday) and following

it the nearest day of type Q (Monday, ..., Sunday)). For customer X the most similar are load profiles of Mondays and Tuesdays, and Tuesdays and Wednesdays. The least similar profiles are for Fridays and Saturdays (according to d) or Mondays and Saturdays (according to r). In the case of customer Y the greatest similarity in load profiles is for Monday and Wednesday, and the smallest similarity is for Saturdays and Sundays, and also for Sundays and Thursdays (according to r).

B. Analysis of Load Density Profiles

To describe the variability of the consumer in the given time period T we compute the distribution of his loads in 15-min intervals (calculated on the basis of 15-min consumption measured by the meter) and visualize it using a histogram. There are the load ranges on the histogram x-axis and the frequency on the y-axis. The frequency is calculated as the number of 15-min periods in which load falls in the given load range divided by the total number of 15-min periods in T. The histogram gives cumulative information about the X customer load density in the period T, so we call it a load density profile.

Histograms can be used for comparison density profiles of the customer in different time periods, e.g. neighboring years/months, the same months of different years or summer and winter periods. The latter case for two households (X and Y) in Fig. 6 is shown. The differences between histograms shown in the figure as Δh are defined as follows:

$$\Delta h = \sum_{i=1}^{n} | h_{1,i} - h_{2,i} | \tag{3}$$

where $h_{k,i}$ is the height of the i-th bar of the k-th histogram.

The maximum value of Δh, indicating the maximum discrepancy of histograms, is 2. If the histograms are identical $\Delta h = 0$. As we can see from Fig. 6 the household Y expresses greater differences in summer and winter profiles. In winter, it uses more load in range of 0.35-1.0 kW, and less in range of 0.1-0.3 kW.

The distribution of the differences Δh between winter and summer profiles for all 1000 customers in Fig. 7 is shown. The median and mean values of Δh are 0.29 and 0.36, respectively.

Typically, daily load profiles show different nature for working days, Saturdays and Sundays/public holidays. The difference Δh calculated for:

- working days and Saturdays was 0.19 for X customer and 0.14 for Y customer,
- working days and Sundays was 0.18 for X customer and 0.14 for Y customer,
- Saturdays and Sundays was 0.06 for X customer and 0.07 for Y customer.

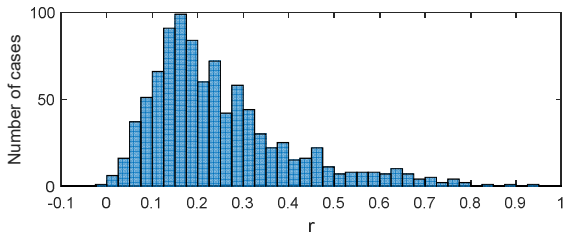

Figure 4. Histogram of the mean value of the correlation coeficient between load profiles for two successive Mondays, …, Sundays calculted for 1000 customers.

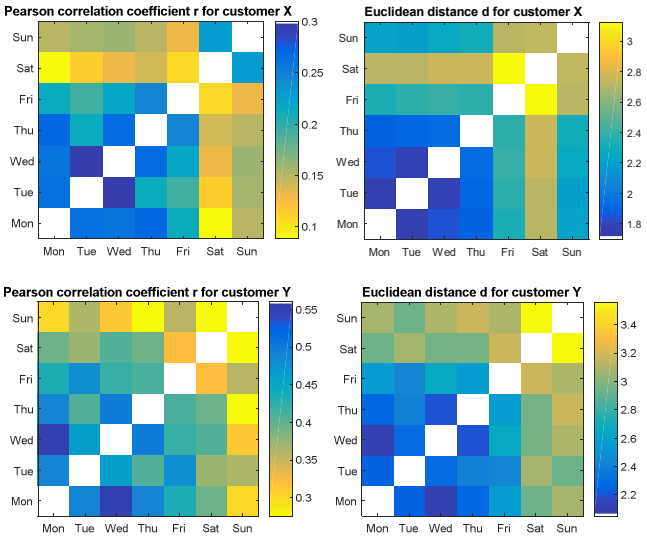

Figure 5. Similarity in daily load profiles for days of the week.

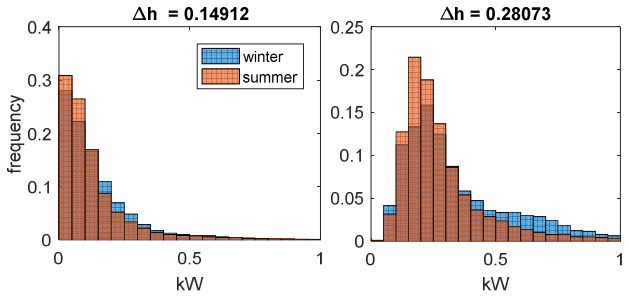

Figure 6. Comparison of density profiles of winter and summer periods for two customers (left – customer X, right – customer Y).

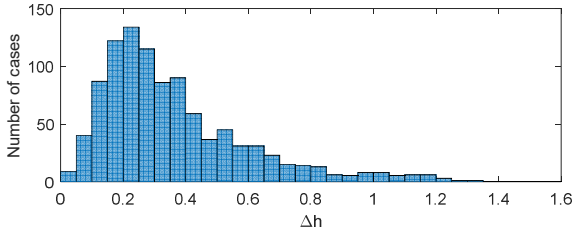

Figure 7. Histogram of the differences Δh between winter and summer profiles for 1000 customers.

Thus, working days are much less similar in density profiles to Saturdays and Sundays than Saturdays to Sundays. More detailed results in Fig. 8 are shown, where differences Δh between each two days of the week are presented. Interestingly, for X customer Wednesday and Thursday density profiles are similar to each other but not much similar to other working days. All days except Wednesday and Thursday show a high degree of similarity to each other. Another picture we get for customer Y, where the most similar in density profiles are: Monday to Sunday and Tuesday to Friday. And the least similar are: Wednesday and Thursday to Tuesday, Friday and Saturday.

We can also compare two households using their density profiles as well. Fig. 9 shows the distribution of the differences Δh between annual density profiles calculated for customer X and each of the 999 other customers included in our database. In this case median of Δh is 0.61 and its mean is 0.69.

C. *Analysis of Seasonal and Irregular Components*

The seasonal variations can be illustrated using autocorrelation plot. Sample autocorrelation function for customer X in Fig. 10 is shown. The strongest autocorrelation for daily lag is observed.

The main frequencies of the measurement data time series can be detected using harmonic analysis. In Fig. 11 an example of the periodogram for customer X is shown obtained using the Fourier transform. The peaks indicate periodic components in the input data. We can observe three dominant peaks: for 12 hours, 24 hours and one week. The first two are significantly higher than the last one meaning that this customer expresses mainly the daily periodicity. Among 1000 customers, in 735 cases the highest peak was for daily period, in 149 cases for half daily period and only in one case for weekly period.

To detect components of the measurement time series, Seasonal and Trend decomposition using Loess (STL) is used. STL is an algorithm that was developed to divide up a time series into three components namely: the trend, seasonality and remainder [10]. In Fig. 12 an example of decomposition is shown for customer X. There is a bar at the right hand side of each graph to allow a relative comparison of the magnitudes of each component. Each bar represents the same length. The smallest bar in the bottom panel shows that the variation in the remainder component is much greater compared to the variation in the seasonal and trend components. Thus, the random component in this time series plays the most important role.

To measure the relative strength of the irregular component in relation to the seasonal component we define the ratio:

$$s = \frac{IQR(remainder)}{IQR(seasonal)} \tag{4}$$

where $IQR(remainder)$ is the interquartile range of the remainder component and $IQR(seasonal)$ is the interquartile range of the seasonal component.

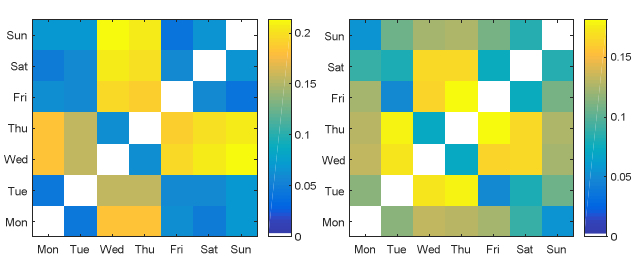

Figure 8. Differences in density profiles for days of the week (left – customer X, right – customer Y).

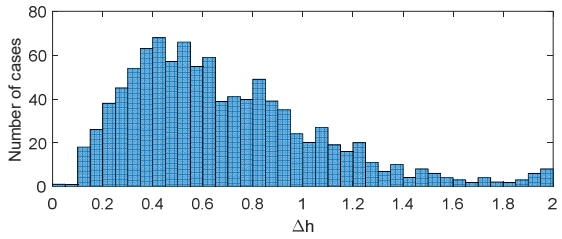

Figure 9. Histogram of the differences Δh between annual density profiles calculated for customer X and each of the 999 other customers.

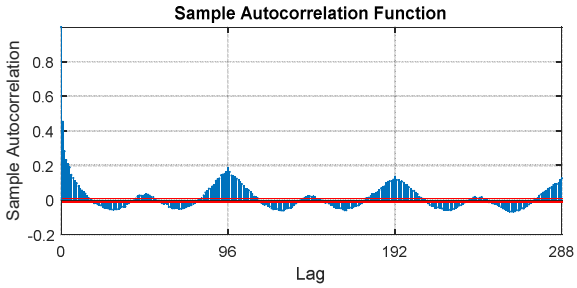

Figure 10. Autocorrelation plot for customer X.

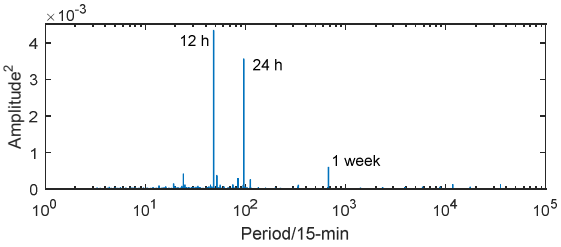

Figure 11. Periodogram for customer X.

The interquartile range is a measure of statistical dispersion, being equal to the difference between upper and lower quartiles $IQR = Q_3 - Q_1$. Ratio (4) informs about how strong is the irregular component compared to the seasonal component. Fig. 13 shows that in most of the analyzed measurement time series it is stronger than the seasonal component (median of s was about 1.11, and its mean was 1.29).

III. CONCLUSIONS

This paper presents descriptive analytics methods for smart meter data. At the household and building levels the

data are much more random and volatile than those at aggregate levels. Analysis of the daily load profiles using correlation coefficient and Euclidean distance shows low level of similarity between them. This is due to spikes corresponding to the switching on and off electrical appliances such as a cooker, kettle, iron, microwave, washing machine etc. Moments of switching on and off of these devices change from day to day.

Load density profiles inform about the distribution of the customer load in a given time period. They can be used for comparison the variability of the consumer in different period of the year or in different days of the week. We can also compare different customers using their density profiles. The outlying density profile in relation to the profiles of customers representing the same tariff and similar contracted power can be used to detect bad or missing data or energy thief.

The analyzed smart meter data for 1000 households express daily and half daily cycles and much weaker weekly cycle. Decomposition of the measurement time series using STL method shows that the irregular component in time series is even stronger than seasonal component.

REFERENCES

[1] Smart Meters and Smart Meter Systems: A Metering Industry Perspective. Edison Electrical Institute/EEI and AEIC Meter Committees, EEI-AEIC-UTC White Paper, March 2011.

[2] A.A. Munshia, and Yasser A.-R. I. Mohamed, "Big data framework form analytics in smart grids," *Electric Power Systems Research*, vol. 151 pp. 369–380, 2017.

[3] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," arXiv:1802.04117v2, 2018.

[4] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397-410, 2014.

[5] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in *Proc. 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conf.*, pp. 1–5.

[6] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Industrial Informatics*, vol. 12(3), pp. 1005–1016, 2016.

[7] L.A.P. Junior, C.C.O. Ramos, D. Rodrigues, D.R. Pereira, A.N. de Souza, K.A.P. da Costa, and J.P. Papa, "Unsupervised nontechnical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016.

[8] R. Granell, C.J. Axon, and D.C. Wallom, "Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles," *IEEE Trans. Power Systems*, vol. 30(6), pp. 3217–3224, 2015.

[9] E.D. Varga, S.F. Beretka, C. Noce, and G. Sapienza, "Robust realtime load profile encoding and classification framework for efficient power systems operation," *IEEE Trans. Power Systems*, vol. 30(4), pp. 1897–1904, 2015.

[10] R.B. Cleveland, W.S. Cleveland, J.E. McRae and I.J. Terpenning, "STL: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6(1), pp. 3–73, 1990.

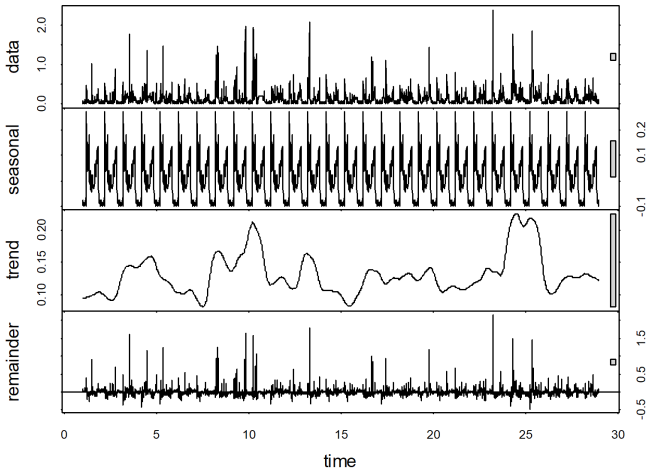

Figure 12. Measurement time series decomposition using STL for customer X.

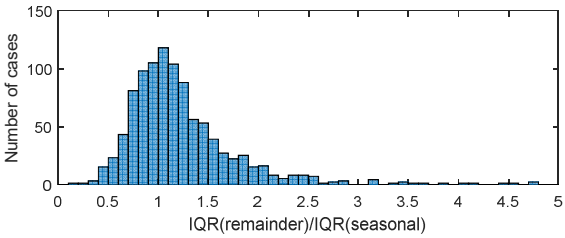

Figure 13. Histogram of the ratio *s* for 1000 customers.