Theoretical and Applied Informatics
Vol. 27 (2015), no. 1, pp. 13–24
DOI: 10.20904/271013

Classification of LPG clients using the Hurst exponent and the correlation coefficient

Krzysztof Domino[1*] Przemysław Głomb[1†] Zbigniew Łaskarzewski[2‡]

[1]Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, 44-100 Gliwice, Poland

[2]AIUT Sp. z o.o., Wyczółkowskiego 113, 44-109 Gliwice

Abstract In this paper we present the analysis of the gas usage for different types of buildings. First, we introduce the classical theory of building heating. This allows the establishment of theoretical relations between gas consumption time series and the outside air temperature for different types of buildings, residential and industrial. These relations imply different auto-correlations of gas usage time series as well as different cross-correlations between gas consumption and temperature time series for different types of buildings. Therefore, the auto-correlation and the cross-correlation were used to classify the buildings into three classes: housing, housing with high thermal capacity, and industry. The Hurst exponent was calculated using the global DFA to investigate auto-correlation, while the Kendall's τ rank coefficient was calculated to investigate cross-correlation.

Keywords building heating; Hurst exponent; LPG distribution

Received 11 JAN 2016 **Revised** 28 JAN 2016 **Accepted** 03 FEB 2016

1 Introduction

Today, gas stock management uses telemetry to collect clients' daily or hourly gas usages and tank levels. This data is used both for day-to-day decisions (LPG tanker destinations) and strategic planning (supply contracts, tariff definition). Effective management at the scale of tens of thousands of users requires the development of advanced software tools that help specialists and partially automate data processing and analysis tasks.

*E-mail: kdomino@iitis.pl
†E-mail: przemg@iitis.pl
‡E-mail: zbigniew.laskarzewski@aiut.com

One of the important tasks of client management is profile creation. It groups clients according to their usage characteristics. Traditionally, this classification is based on the mean usage level, and it is performed by a human operator. The automation of this process reduces the need for manual data processing, allows for more flexible profile creation, and preparation of personalized options for clients.

In this paper, we investigate an approach for automatic classification of users based on temperature dependent gas usage. The dependence on temperature is one of the most important factors, as most users use LPG gas for house heating. The gas consumption time series were examined using the Hurst exponent and Kendall's τ cross-correlation coefficient. The first one was used to examine auto-correlations of time series and the second one to investigate cross-correlations between gas consumption and temperature time series. These parameters implied the classification into three classes: residential objects, residential objects with high thermal capacity, and industry objects.

There are many methods of the customer classification, such as SVM (Support Vector Machines), clustering, tree methods [1], fuzzy modelling [2, 3], etc. In our research we introduced a method based on the Hurst exponent. The Hurst exponent was used in the time series prediction / classification [4, 5, 6] and in the object classification [7, 8, 9]. To evaluate the Hurst exponent approach, we assume that the gas consumption time series are modelled by the stochastic process that is discussed below.

1.1 The heat consumption

The analysis of the gas fuel consumption in the heating process is an important issue especially now as energy efficiency has become an important issue in economy and industry. Let us review a few basic concepts concerning the use of energy, in order to analyze gas consumption time series in comparison to the outside air temperature time series.

The energy lost by the heat conduction Q_c in the given time period t can be represented as:

$$Q_c = \sum_{i=1}^{N} U_i t \Delta T. \tag{1}$$

The value of $U_i \left[\frac{W}{K}\right]$ corresponds to the energy transfered through the i-th outer wall at the unit time and the unit temperature. The index i counts all outer walls. The temperature difference between the outdoors and the indoors is represented by $\Delta T = T_{\text{inside}} - T_{\text{outside}}$. Similarly, the energy lost by the ventilation process is proportional to the temperature difference:

$$Q_v = c\phi\Delta T, \tag{2}$$

where ϕ is the ventilation air stream and the c coefficient involves the heat capacity of the air.

For residential housing, the inside temperature is assumed to be equal to $T_{\text{inside}} = 20^0 C$. Moreover, it can be assumed that if the temperature outside is greater or equal to $16^0 C$ the residential housing is heated by internal heat gains (generated by devices in operation, people, sunshine, etc.). Finally, in residential buildings, the energy lost by heat conduction and ventilation Q_{cv} can be written as:

$$Q_{cv} = \alpha\Delta T\ \theta(16 - T_{\text{outside}}) = (-\alpha T_{\text{outside}} + \beta)\ \theta(16 - T_{\text{outside}}), \tag{3}$$

where α is the coefficient of proportionality and $\theta()$ is the Heaviside theta function. Hence the $T_{\text{outside}} \leq 16$ data follow the linear relation:

$$Q_{cv} = -\alpha T_{\text{outside}} + \beta. \tag{4}$$

To account for the total energy consumption, the energy used for water heating – Q_w has to be added as well. In general, it does not depend on the temperature T_{outside}, and can be represented as:

$$Q_w = f \sum_{i=1}^{n} c_i, \tag{5}$$

where f is the coefficient of proportionality involving the heat capacity of water and the temperature difference between cold and hot water, c_i is the water used by the i-th user in a given period of time, and n is the number of users. The coefficients c_i are constant at average, although the individual behavior of users generates some level of noise. However, this random amount of noise seems to be much smaller than the noise generated by the variation of the outdoor temperature which is crucial in the following analysis.

Finally, the general heat consumption $Q_G = Q_c + Q_v + Q_w$ is analyzed. Let us assume that a building is heated with gas only. In this case, gas usage G is proportional to the heat consumption $G = \mu Q_G$, where μ accounts for the efficiency coefficient of the heating system that can be approximated as constant. In this approximation the seasonal variation in the supplied cold water temperature is disregarded. Assuming that the hot water consumption is constant, we have

$$G = -\alpha' T_{\text{outside}} + \beta', \tag{6}$$

where α' and β' are constant coefficients.

Let us represent the outside temperature as a function of time $T(t)$ being a stochastic variable with correlated increases $W_H(t)$ – it is adequate since the correlated noise was used to analyze weather phenomena,

$$T(t) = T(t-1) + W_H(t). \tag{7}$$

The gas consumption can be modelled similarly:

$$G(t) = G(t-1) - \alpha W_H(t) = G(t-1) + W'_H(t), \tag{8}$$

where $W_H(t)$ and $W'_H(t)$ have the same Hurst exponent, since the DFA procedure described in Section 2.1 overlooks the linear scaling.

1.2 Industry objects

In the case of industrial buildings, the production process may influence the gas usage more significantly than the heating process. This may be caused by the following situations:

1. there may be large energy gains from production;
2. gas may be used for the industry process itself in the larger volume that for the heating process.

If the use of gas for the industrial process is significant, the equation (6) has to be generalized:

$$G(t) = -\alpha T_{\text{outside}}(t) + \beta + W_H^I(t). \tag{9}$$

Here we introduce the random industrial process $W_H^I(t)$, that may be a correlated process in the general case. If its influence dominates over the heating process, the gas consumption function can be approximated as:

$$G(t) \propto W_H(t). \tag{10}$$

If the DFA is used for such a process it will result in very low Hurst exponent values, and industrial objects can be distinguished from residential ones. Note that the proportional coefficient is irrelevant in the DFA procedure.

2 The algorithm

The classification algorithm is based on the following assumptions:

- For residential objects the gas consumption is mainly modelled by outdoor temperature $T_{\text{outside}}(t)$.
  1. Negative cross-correlation between gas usage $G(t)$ and temperature $T_{\text{outside}}(t)$ time series.
  2. Similar auto-correlations of $G(t)$ and $T_{\text{outside}}(t)$ time series.
- If one or both of the above conditions are not fulfilled the object is classified as an industrial one.

2.1 Auto-correlation

DFA procedure To measure the auto-correlation, the standard procedure of the global DFA was performed [4, 6]. The DFA starts with the time series e.g. $x(t)$. Next, the scaling of the mean squared difference between data $x(t)$ and their polynomial trend approximation $r_n(t)$ is examined. In this research the linear trend approximation was used as in [4, 6] and the references therein. The DFA procedure is performed as follows [10]:

1. the time series $x(t)$ is divided into i non-overlapping time windows of length T',
2. for each window
   - linear regression prediction $r(t)$ is calculated,
   - detrended variance $F_j^2(T') = \frac{\sum_{t=1}^{T'}(x(t)-r(t))^2}{T'}$ is evaluated,
3. for i-th window the detrended variance is averaged $\left\langle F^2(T')\right\rangle = \frac{\sum_{j=1}^{i} F_j^2}{T_i'}$,
4. the relation, $\left\langle F^2(T')\right\rangle \propto T'^{2H}$ is used to examine the Hurst exponent – H.

To perform last step the logarithm is applied:

$$\log \left\langle F^2(T') \right\rangle = 2H \log T' + b, \tag{11}$$

and finally, the linear regression fit is used to determine the Hurst exponent. In our research the T' were chosen such as $T' = \lceil \frac{T}{i} \rceil$ and $i = 2, 3, 4, ..., 15$. This data choice allowed to reflect the linear scaling region of $\log \left\langle F^2(T') \right\rangle$ vs $2H \log T'$. The examples of the linear fitting are presented in Figure 1.

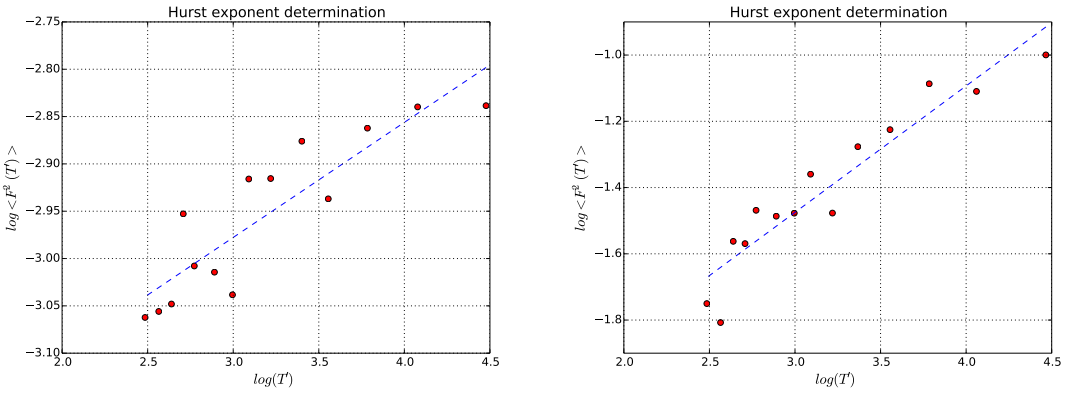

Figure 1 The linear fit of $\log \left\langle F^2(T') \right\rangle$ vs $2H \log T'$, the slope equals to $2H$.

Finally, the auto-correlations of time series can be deduced by using the Hurst exponent [4, 6]

1. if $H > 0.5$ data posses long range auto-correlations,
2. if $H = 0.5$ there are no auto-correlations,
3. if $H < 0.5$ there are negative auto-correlations.

Auto-correlation data analysis. At the beginning the temperature data were processed. For 15 samples of data the following Hurst exponent values were observed $H \in [0.31, 0.45]$. Therefore, it can be concluded that temperature data are negatively auto-correlated. For residential objects gas consumption is supposed to have negative auto-correlations reflected by $H < 0.5$. However, if the Hurst exponent is very low – the threshold value $H < 0.2$ is set – the relation of the type of eq. (10) is suggested to model the gas consumption for industrial objects. See Figure 2 for the Hurst exponent threshold value justification. Finally, if the gas consumption is positively auto-correlated or has no auto-correlation, some factor must be responsible for this situation. The most natural explanation is high thermal capacity of the building (e.g. old type brick construction). The choice for the negative auto-correlation would be $H < 0.5$ – see Figure 2 for justification. We suggest three classes for data classification.

1. The industrial class, where $H < 0.2$ – this Hurst exponent value divides data in Figure 2 into two peaks.

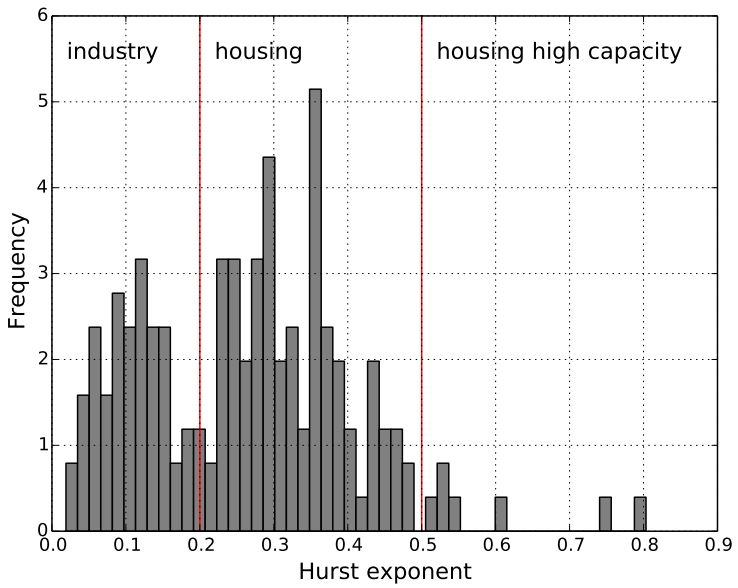

Figure 2 The frequency distribution of the Hurst exponent values – bins areas are normalized frequencies.

2. The housing class, where $0.2 \leq H < 0.5$ – the anti-correlation region in Figure 2.
3. The housing class with the large heat capacity of the building, where $H \geq 0.5$ – the correlation (tail) region in Figure 2.

2.2 Cross-correlation.

Kendall's τ. The cross-correlation analysis can be used to improve the classification. The Kendall's τ correlation coefficient [11, 12] is suitable, since it does not require the normal frequency distribution and makes the investigation easily expandable to the copula approach [13]. Let us suppose that (X_1, X_2) and $(\tilde{X_1}, \tilde{X_2})$ are random vectors from two time series, numbered as 1 and 2 – gas consumption and temperature time series in our case. The Kendall's τ rank coefficient is defined as [12]:

$$\tau = E\Big(\text{sign}\big((X_1 - \tilde{X_1})(X_2 - \tilde{X_2})\big)\Big), \tag{12}$$

where $E()$ is the expecting value. The Kendall's τ as a rank coefficient is independent from the frequency distribution of data and, unlike the Pearson's one, it can be used for non-Gaussian data. This issue is important since the temperature data may not have the normal frequency distribution [14, 15].

Cross-correlation analysis. The Kendall's τ rank coefficient can be used to determine if two sets of data are cross-correlated. It is important to remember that the gas consumption

and the outside temperature are expected to be negatively cross-correlated for housing units – strong negative cross-correlation is suggested since the equation (6) is supposed to hold at least approximately. Let us propose $\tau < -0.3$ to classify the gas consumption for housing units – see Figure 3 for the justification, the value from the middle of the $[-0.2 - 0.4]$ range was chosen that divides data into two peaks.

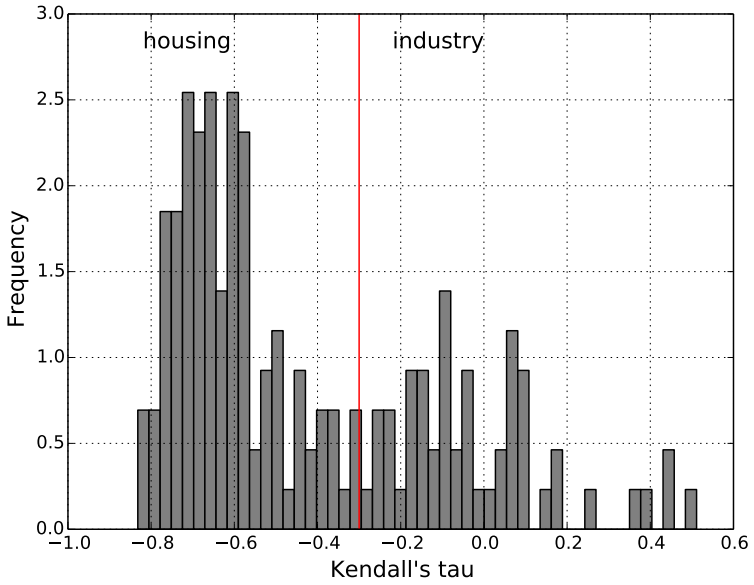

Figure 3 The frequency distribution of Kendall's τ values – bins areas are normalized frequencies.

The final classification outcomes:

1. Industrial class: $H < 0.2$ or $\tau \geq -0.3$.
2. Housing class: $0.2 \leq H < 0.5$ and $\tau < -0.3$.
3. Housing with high thermal capacity class: $H \geq 0.5$ and $\tau < -0.3$.

In Figure 4, mean, variance, median, mode, asymmetry, and kurtosis of the gas usage of these three classes are presented. For the industrial class, mean, mode, variance and median are generally smaller than for housing classes. On the other hand, asymmetry and kurtosis are generally larger for the industrial class than for both housing classes. We believe that this differences are due to different dynamic of the factors influencing housing classes (the outside temperature being the main factor), and factors influencing industrial class (the industry process dynamics being the main factor). However, there are some overlaps making the classification inaccurate. Moreover, the housing with high thermal capacity cannot be distinguished from the housing class by using the descriptive analysis.

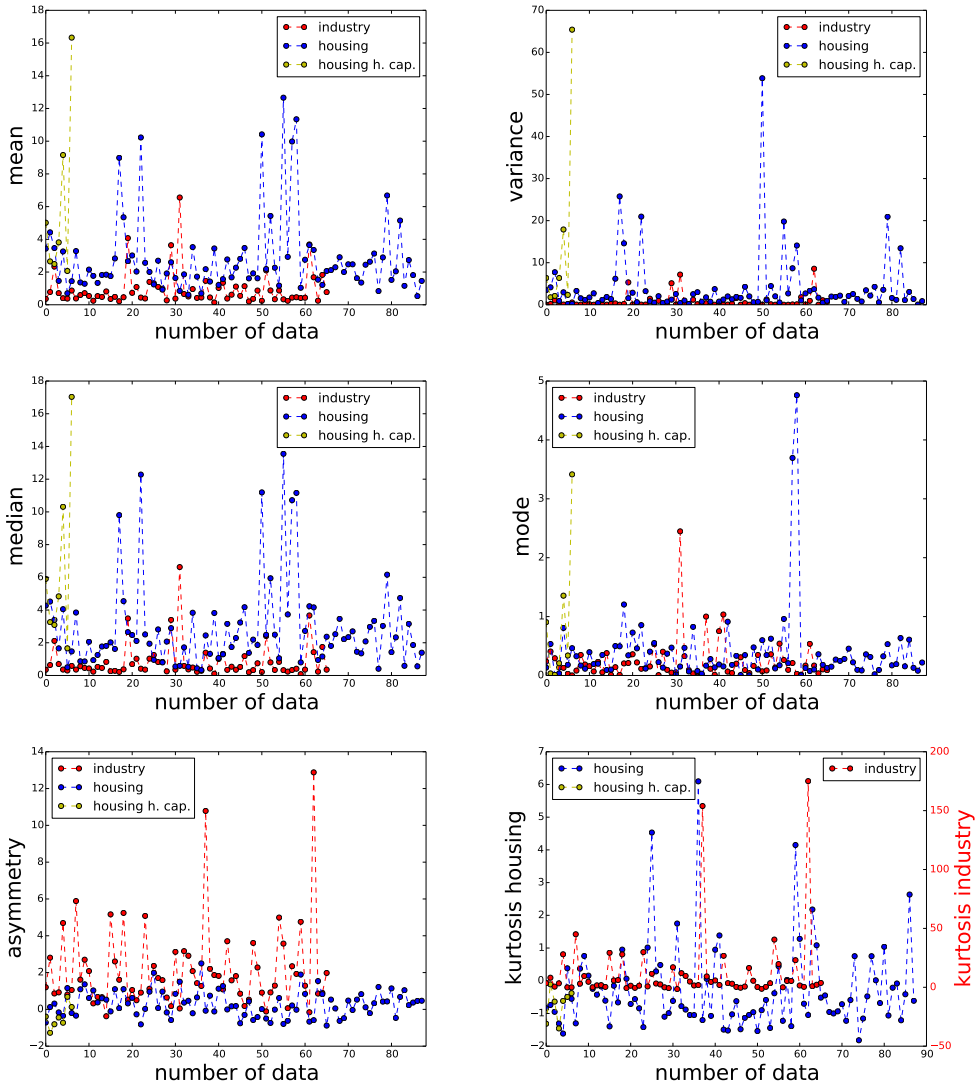

Figure 4 Descriptive analysis of the gas usage.

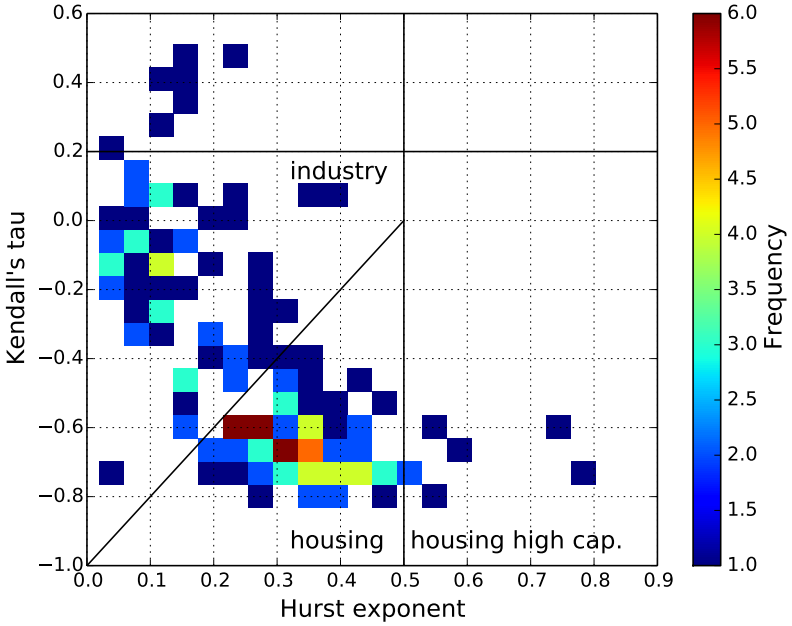

Figure 5 The frequency distribution of the Hurst exponent and Kendall's τ values.

3 Data processing

Gas usage values are collected from telemetry modules installed onto various types of gas meters. Reed switch or Hall effect sensor of telemetry module detects the rotation of the last digit disk of a mechanical counter with an installed magnet. The counter of impulses is stored in a module and typically once per day it is transmitted to the telemetry server using GSM, SMS, or GPRS. The telemetry server is a high-performance platform of data acquisition that decodes data from various types of modules. The decoded measurements are stored in a database.

The experiments were performed on gas usage series from thousands of telemetry modules installed in the locations for 2-3 years. The algorithm was used to analyze the counter values from gas meters recalculated to gas usage with measurements timestamps (mostly there wes one event per day), outside temperature, and an auxiliary column used to determine interpolated data that have to be overlooked for the DFA procedure. The timestamps were used to index data for the DFA. Only data where temperature was lower than 16^0C were concerned, and the reason for this restriction was discussed in Section 1.1.

Figure 4 presents the two-dimensional frequency distribution of data. There are two distinct areas that can be associated with the housing class (low Kendall's τ and high Hurst exponent) and the industry class (low Hurst exponent and high Kendall's τ). There is also the third "tail" area, where $H \geq 0.5$, that may be a subclass of the housing class due to negative cross-correlation – low Kendall's τ. The buildings with high thermal capacity may fit into the third class. The algorithm presented in Section 2 roughly separates data into three classes.

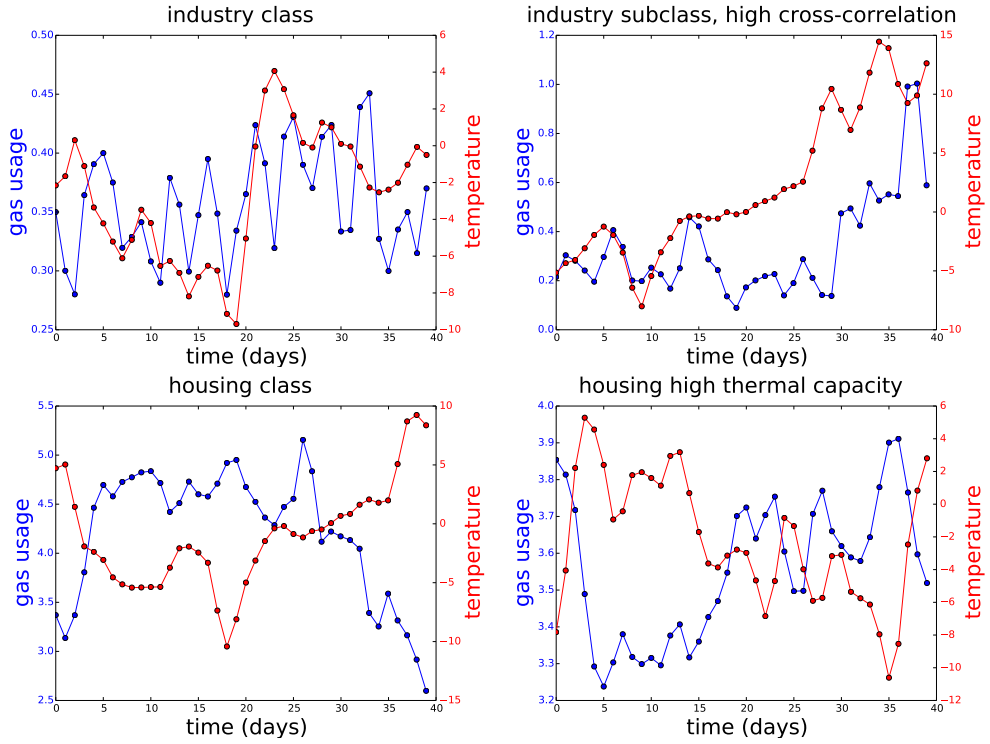

Figure 6 Exemplary randomly chosen fragments of time series for industrial and housing classes.

However, the analysis of Figure 5 revealed that the linear combination of the Hurst exponent and Kendall's τ values may present a better classification tool. Further analysis is necessary to justify this proposal. Furthermore, there is a possibility of the occurence of the fourth class – a subclass of the industrial one, where Kendall's τ values are high (e.g. $\tau > 0.2$). We call it the industrial subclass with high cross-correlation, with the industrial objects where the gas consumption is cross-correlated to temperature. Further analysis is necessary to investigate this type of industrial process.

Regarding the "improved" classification scheme (presented in Figure 5), the examples of randomly chosen fragments of time series are presented in Figure 6. Complete time series cannot be presented since data are confidential. For housing classes the negative cross-correlation of outside temperature and gas usage is evident. For the industrial classes this is not the case. The analysis of Figure 6 leads to the conclusion that it is difficult to distinguish between the housing class and the housing class with high thermal capacity, as well as between the industrial class and the industrial subclass with high cross-correlation without the use of the Hurst exponent and the cross-correlation coefficient.

4 Conclusions

The method based on the auto-correlation and cross-correlation analysis was used to classify the gas consumption data. The Hurst exponent calculated by the global DFA was used to measure auto-correlation. The Kendall's τ rang correlation coefficient was applied to measure cross-correlation. A simple classification algorithm was proposed. The two-dimensional frequency distribution of the Hurst exponent and Kendall's τ values analysis was performed and its results are presented in Figure 5. The algorithm roughly distinguishes the presented classes, but its improvement is possible. Finally, it was suggested that the use of the descriptive analysis (Figure 4) or the observation of time series (Figure 6) without the use of the Hurst exponent and the cross-correlation coefficient will not allow distinguishing the data between all suggested classes.

Acknowledgemnts The research work was realized in cooperation with the company AIUT sp. z o.o. and supported by Polish agency The National Research and Development Centre grant INNOTECH-K2/HI2/6/182421/NCBR/13.

References

[1] W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, and J. Kunze, editors. *Challenges at the Interface of Data Analysis, Computer Science, and Optimization. Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V., Karlsruhe, July 21 - 23, 2010*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-24466-7.

[2] M. Sugeno and T. Yasukawa. A fuzzy-logic-based approach to qualitative modeling. *IEEE Trans. Fuzzy Syst.*, 1(1):7–31, 1993. DOI: 10.1109/TFUZZ.1993.390281.

[3] T.-L. Hu and J.-B. Sheu. A fuzzy-based customer classification method for demand-responsive logistical distribution operations. *Fuzzy Sets Syst.*, 139(2):431–450, 2003. DOI: 10.1016/S0165-0114(02)00516-X.

[4] K. Domino. The use of the hurst exponent to predict changes in trends on the warsaw stock exchange. *Physica A*, 390(1):98–109, 2011. DOI: 10.1016/j.physa.2010.04.015.

[5] K. Domino. The use of the hurst exponent to investigate the global maximum of the warsaw stock exchange WIG20 index. *Physica A*, 391(1):156–169, 2012. DOI: 10.1016/j.physa.2011.06.062.

[6] K. Domino, T. Błachowicz, and M. Ciupak. The use of copula functions for predictive analysis of correlations between extreme storm tides. *Physica A*, 413:489–497, 2014. DOI: 10.1016/j.physa.2014.07.020.

[7] H. Zghidi, M. Walczak, T. Błachowicz, K. Domino, and A. Ehrmann. Image processing and analysis of textile fibers by virtual random walk. *Sci. Educ.*, 44:100, 2015. DOI: 10.15439/2015F40.

[8] T. Błachowicz, A. Ehrmann, H. Zghidi, and K. Domino. Optical determination of hemp fiber structures by statistical methods. *Proceedings of Aachen-Dresden International Textile Conference*, 2015.

[9] A. Ehrmann, T. Błachowicz, K. Domino, S. Aumann, M.O. Weber, and H. Zghidi. Examination of hairiness changes due to washing in knitted fabrics using a random walk approach. *Text. Res. J.*, page 0040517515581591, 2015. DOI: 10.1177/0040517515581591.

[10] G.L. Vasconcelos. A guided walk down wall street: an introduction to econophysics. *Braz. J. Phys.*, 34(3B):1039–1065, 2004. DOI: 10.1590/S0103-97332004000600002.

[11] H. Abdi. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510, 2007.

[12] M.G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938. DOI: 10.1093/biomet/30.1-2.81.

[13] U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004. DOI: 10.1002/9781118673331.

[14] R.-G. Cong and M. Brady. The interdependence between rainfall and temperature: copula analyses. *Sci. World J.*, 2012:405675, 2012. DOI: 10.1100/2012/405675.

[15] C. Schoelzel and P. Friederichs. Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlin. Processes Geophys.*, 15(5):761–772, 2008. DOI: 10.5194/npg-15-761-2008.