# Automatic Segmentation of Electricity Consumption Data Series with Jensen-Shannon Divergence

István Pintér [#1], Lóránt Kovács [#2], András Oláh [#3] , Rajmund Drenyovszki[#4], Dávid Tisza[#5] and Kálmán Tornai[#6]

[#1 ,2 ,3, 4, 5,6] *Department of Informatics, Kecskemét College*
*Izsáki út 10, H-6000 Kecskemét, Hungary*

[#3, 5, 6] *Faculty of Information Technology, Pázmány Péter Catholic University*
*Práter utca 50/a, H-1083 Budapest, Hungary*

[#4] *Department of Computer Science and Systems Technology, University of Pannonia,*
*Egyetem u. 10, H-8200 Veszprém, Hungary*

[1] pinter.istvan@gamf.kefo.hu
[2] kovacs.lorant@gamf.kefo.hu
[3] olah.andras@itk.ppke.hu
[4] drenyovszki.rajmund@gamf.kefo.hu
[5] tisza.david@itk.ppke.hu
[6] tornai.kalman@itk.ppke.hu

***Abstract*—In Smart Grids the Information and Communication Technologies (ICT) could be used to better manage both consumption and production of electricity. The increasing presence of renewable energy sources in production and the permeation of novel consumption types (e.g. Plug-in Hybrid Electric Vehicles (PHEV)) will obviously cause the increase the fluctuation of electrical energy. One possible solution to these problems is development of novel methods for investigating electrical power consumption data series. As the existing learning algorithms of pattern classification are suitable for discovering internal structures of large datasets, it is important to generate a training/testing/validation learning database from existing measurements (e.g. from smart meters), actually via segmentation and labeling by hand. In this paper we propose a novel method for the automatic segmentation with a predefined confidence level. The algorithm is based on the generalized Jensen-Shannon divergence (JSD), and it estimates the change-points (CPTs) in electrical power consumption data. Both the method and some recent results in segmenting one household's power consumption data are presented in this paper.**

***Keywords:*** **Smart grid, household electricity consumption, Jensen-Shannon divergence, change point, segmentation**

## I. INTRODUCTION

In developing technologies and systems for energy-efficient economy [1], the smart grid concept is very important. From system's point of view even the role of a household should be revised, because it can be characterized not only with its power consumption, but with electricity production as well. Novel power consumption demands appeared, e.g. charging PHEVs' batteries, also with the problems of V2H (Vehicle to Home) and V2G (Vehicle to Grid) infrastructures. Power generation has also appeared at household-level, e.g. co-generation system (CHP, combined heat and power), small wind turbines, photovoltaic panels, charged batteries of electric vehicles. Both the power consumption and the production obviously increase the fluctuations in electrical energy, which makes the balancing much more complex. Obviously, using a feedback-system it could be easier to develop and maintain and adaptive system for handling these problems, and the ICT (and smart grid) naturally comes into question [2]. Using smart meters it is possible to record and evaluate the power consumption in a fine time-scale (e.g. 1 seconds or 1 minute sampling intervals in case of households), which makes it possible to create a large enough database for developing sophisticated adaptive (learning) algorithms for managing household's electrical power demand. According to a recent article [3], using such a large database households can be categorized into a small number of typical consumption patterns based on their electricity consumption data, which fact is important from the point of view of demand side management.

The power grid (be it smart or not) is very complex system, and it can't be easily modeled mathematically. However, there are suitable techniques of pattern recognition for modeling signals or systems from measured data series. The most informative are the labeled data, in which a kind of description of objective background is also attached to measurement therefore supervised learning algorithms can be used. When building a learning database segmentation is an important sub-task. During segmentation an expert investigates the data series and marks the beginning and the end of interesting parts. As it could be a time-consuming work, an automatic segmentation algorithm is useful. Automatic segmentation procedure is also useful in case of unsupervised learning, where labels are not available. Our proposed automatic segmentation method is based on Jensen-Shannon divergence-based statistical similarity of subsequences of data series, and the endpoints of the segments are the so-called change points [4], [5]. The CPTs (namely those time-instants where statistical properties of time-series data change) can also be determined by computing the JSDs. The results obtained are presented with artificially created data and also

with a publicly available household power consumption database [6].

The organization of the paper is the following. Section II gives a short summary of the computation method and the properties of the generalized Jensen-Shannon divergence. The concept of JSD-contour also introduced with its maximum's index as estimation of a CPT. We demonstrate the method using artificially created symbol sequences and in Section III the results of automatic segmentation of household data series are presented. After conclusions and acknowledgements the paper ends with references.

## II. CPT ESTIMATION IN SYMBOLIC SEQUENCES USING GENERALIZED JENSEN-SHANNON DIVERGENCE

Mainly based on [4] this section gives a short summary of JSD, generalized JSD, JSD contour, index of maximum of JSD contour as CPT-estimate and the recursive segmentation algorithm for determining CPTs at a given significance level.

### A. *The generalized Jensen-Shannon divergence and the JSD-contour*

The JSD is an information theoretic measure to determine the similarity of two discrete probability distributions [7],[8]:

$$D_{JS}(P,Q) = \frac{D_{KL}(P,A)+D_{KL}(Q,A)}{2}, \quad (1)$$

where $P = (p_1, \dots, p_k, \dots, p_K)$ and $Q = (q_1, \dots, q_k, \dots, q_K)$ are the distributions in question, $A = (a_1, \dots, a_k, \dots, a_K)$ the average distribution $\left(a_k = \frac{p_k+q_k}{2}\right)$, and $D_{KL}(.,.)$ denotes the Kullback-Leibler (K-L) divergence [7],[8]. It is important to note, that authors of [3] also used information theoretic measure in their method, namely the symmetrized generalized K-L-divergence.

Useful properties of $D_{JS}(P,Q)$ are the symmetry and the non-negativity. Moreover $\sqrt{D_{JS}(P,Q)}$ satisfies the triangle inequality, therefore it can be used as a generalized distance between the two distributions.

The JSD can also be formulated using the notion of Shannon-entropy:

$$D_{JS}(P,Q) = H(A) - \frac{H(P)+H(Q)}{2} \quad (2)$$

where $H(.)$ denotes the Shannon-entropy of the discrete probability distribution [7].

By introducing the non-negative weights $\omega_P, \omega_Q$ for which $\omega_P + \omega_Q = 1$ the average distribution becomes $A = \omega_P \cdot P + \omega_Q \cdot Q$ and the JSD can be generalized as [4]:

$$D_{JS}(P,Q) = H\left(\omega_P P + \omega_Q Q\right) - \left[\omega_P H(P) + \omega_Q H(Q)\right] \quad (3)$$

This definition can be generalized further for $M > 2$ distribution, for details see [4].

The generalized JSD can be used for CPT estimation in symbol-sequences by introducing the notion of JSD-contour. The symbols are from a finite alphabet $ABC = \{\alpha_1, \dots, \alpha_k, \dots, \alpha_K\}$. Let the symbol sequence $S = [S_1|S_2]$ be concatenation of the left sequence $S_1$ and the right sequence $S_2$, and these sequences are generated by $P$ and $Q$ probability distribution. That is, for the left sequence the probability that in a given position one can find a symbol $\alpha_k$ is $p_k$. The connection between the symbols of right sequence and the probability distribution $Q$ is similar. When the number of symbols in $S, S_1, S_2$ are $, N_1, N_2$ , respectively, we can say, that there is a CPT in $S$ at position $N_1$. The task of estimation of CPT now is the following: with given alphabet and symbol-sequence estimate the position of the CPT. Intuitively speaking, the estimated CPT cuts the original sequence into two sub-sequences, namely the left and right ones. Therefore the above mentioned CPT-estimation procedure can be continued until a suitable stopping criterion. This procedure leads to the recursive CPT-estimation algorithm, which has been used in our work. Following [4], the proposition for a CPT estimate is the maximum value of the so-called JSD-contour. The symbol sequence $S$ can be cut into left and right sequences in positions $= 1, .., N-1$ . At a given position $m$ the number of symbols in left sequence is $m$, whilst the symbols' number is the right sequence is $N-m$. We can now estimate the $P$ and $Q$ distribution for the left and right sequences using the symbols relative frequencies, and by choosing the weights as $\omega_P = \frac{m}{N}$ , $\omega_Q = \frac{N-m}{N}$ the generalized JSD can be computed. Repeating this procedure for all $m$ positions, we get the function $JSD(m)$, which we call JSD-contour. The estimation of CPT is the index of maximum value of the JSD-contour.

However, there is a question of 'goodness' of CPT estimations. In [4] an approximation of probability distribution function *F(x)* of $JSD_{max}$ has been published, that is the CPT-estimation can be checked at a given $p_0$ confidence level. The computed index is accepted as CPT-estimate, if $F(JSD_{max}) > p_0$ is hold.

### B. *Simulation results using artificially generated symbol-sequences*

In order to verify the method summarized in the previous subsection, detailed simulations were performed in case of $K = 2$ and $K = 12$ symbols, respectively.

In first experiment the JSD-contours and the average has been investigated using binary symbols and two probability distributions. The change point was at position 500 (at the end of left sequence), and the length of right sequence was 2000. Altogether 1000 such 2500 symbols-long binary sequences have been generated using $P = (0.3, 0.7)$ and $Q = (0.7, 0.3)$ probability distributions. The results can be seen on Fig 1.

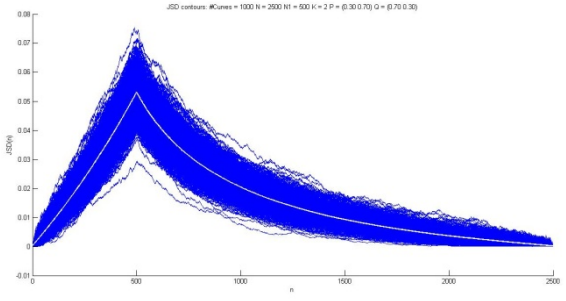

Fig.1. JSD-contours (blue) and the average (white).

There is only one CPT in the above case. The histogram of CPT-estimates has also been computed and can be seen on Fig. 2.

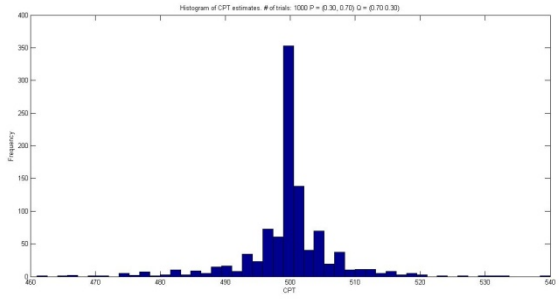

Fig 2. The histogram of CPT estimates (binary case, two concatenated sequences, 1000 experiments).

In case of $K = 12$ symbols there were three subsequences and the CPTs have been estimated using the recursive algorithm with confidence level of 95%. Fig. 3. shows the resulting histograms of the CPT estimates, also with the medians.

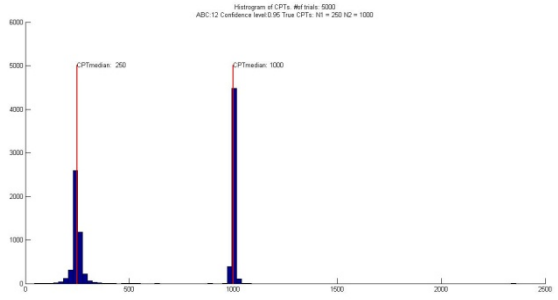

Fig. 3. The histogram of CPT estimates with medians (alphabet size 12, number of experiments 5000, 3 concatenated sequences, true CPTs are at positions 250 and 1000 respectively).

## III. Automatic Segmentation of Household Data Series using the JSD-contour

In this section the application of the method described in Section II is presented for automatic segmenting of a household's power consumption data series.

The source of data was the "Individual household electric power consumption data set" from UCI Machine Learning Repository [6]. In the database there are seven time-series from December 2006 to November 2010. The sampling interval was 1 minute, the measured parameters are global active power, global reactive power, voltage, intensity, sub-metering 1 (kitchen), sub-metering 2 (laundry), sub-metering 3 (water heater, air-conditioner).

### *A. Alphabet generation from measurement values*

The CPT-estimation algorithm is based on the symbols from a known alphabet, so it is necessary to elaborate a mapping from measured electrical power consumption values to a suitable alphabet. It was found experimentally that linear quantization of measured data into $K = 12$ quantization levels gives acceptable results, therefore the symbols are the indices of quantization levels. However, we have to note, that to find other mapping schemes with the corresponding probability distribution functions are open questions.

The original data and the result of linear quantization can be seen on Fig. 4.

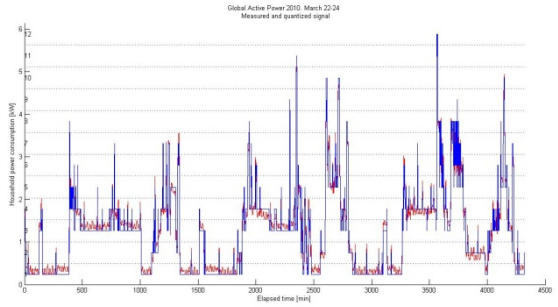

Fig. 4. Original and quantized data (period: March 22-24, 2010, global active power data).

### *B. JSD-contour and the results of automatic segmentation*

On Fig. 5. a JSD-contour can be seen, which has been computed from the same data as of Fig. 4. For better visibility the first JSD-contour has been scaled up to the data values, and the estimated CPT at 95% confidence level has also been marked.

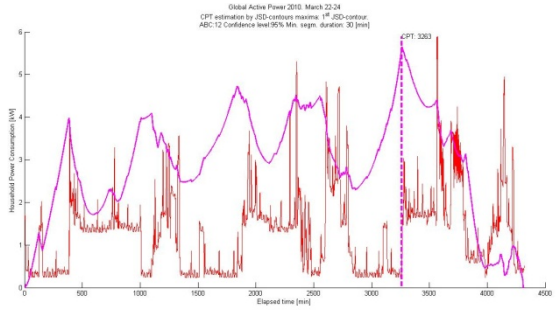

Fig. 5. Original data and the 1st JSD-contour (period: March 22-24, 2010, global active power data, confidence level: 95%).

As the available database contains no labels, but the time-stamps are available, a reasonable input parameter for the automatic segmentation algorithm is the minimal sequence length, which stops the recursion. The following figures illustrate the effect of this parameter on the resulting segmentation.

Fig. 6 presents the case of the shortest sequence with duration of 30 minutes.

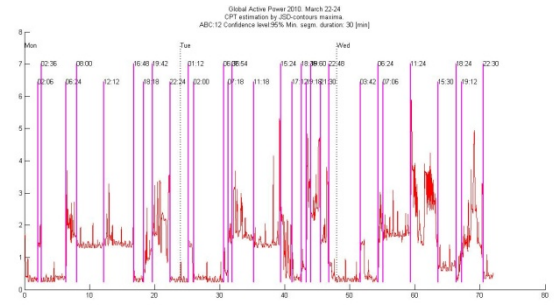

Fig. 6. Automatic segmentation for at least 30 minutes long sub-sequences. (period: March 22-24, 2010, global active power data, confidence level: 95%)

On Figs. 7., 8., 9. the cases of minimal sequence-durations of 1 hour, 3 hours and 6 hours can be seen, respectively.

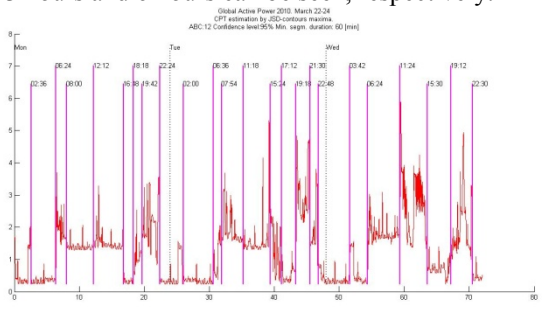

Fig. 7. Automatic segmentation for at least 60 minutes long sub-sequences. (period: March 22-24, 2010, global active power data, confidence level: 95%)

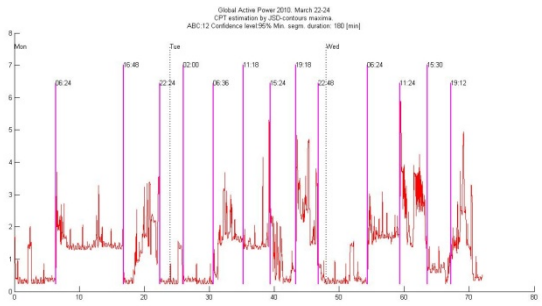

Fig. 8. Automatic segmentation for at least 180 minutes long sub-sequences. (period: March 22-24, 2010, global active power data, confidence level: 95%)

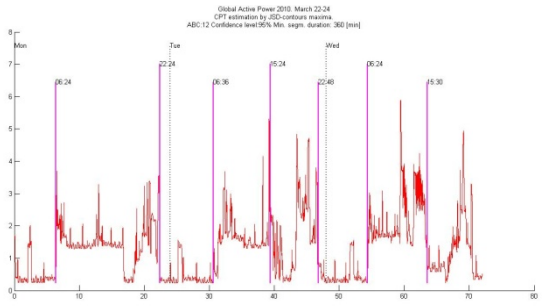

Fig. 9. Automatic segmentation for at least 360 minutes long sub-sequences. (period: March 22-24, 2010, global active power data, confidence level: 95%)

It is worthwhile to mention that the set of CPTs of 6 hour duration case is subset of the CPTs of the 3 hours case and so on unto the 30 minutes case. Although the data set is not labeled, based mainly on Figs. 7., 8. and 9 a kind of regularity in daily power consumption activities of the resident can be deduced by investigating the CPTs (e.g. power consumption activities at morning/afternoon/night). It can also be seen that the borders of the activities are not exactly matching with days. In spite of this, these properties discovered by the automatic segmentation algorithm strengthens the results on typical electricity consumption patterns published in [3]. Our further results have been published in [9] on detecting seasonal changes and home appliance changes in artificially concatenated real measurement-based data series.

## IV. Conclusion

The smart grid concept – namely the integration of power grids with ICT – is very important from the point of view of energy-efficient economy. The management of households' electrical energy consumption became a recent topic, so the developing of methods for analyzing and modeling households' power consumption is important. Both for the supervised and unsupervised learning approach large datasets are necessary from the real world.

As the proposed automatic segmentation algorithm partitions the non-stationary data series into stationary sub-sequences at a given confidence level it could support the machine learning approach in smart grid developments. In this paper we presented a JSD-based method for this purpose and illustrated it on artificially created and also on real data. The most important open questions for further development are the mapping from measurements to symbols-set, the verification using labeled data, and to elaborate a sliding window-based version intended for real-time applications in smart grids.

## Acknowledgment

This research and publication have been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004: National Research Center for the Development and Market Introduction of Advanced Information and Communication Technologies. This source of support is gratefully acknowledged.

## References

[1] Karen Ehrhardt-Martinez et. al.: *Advanced Metering Initiatives and Residental Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities.* June 2010, Report Number E105, American Council for Energy-Efficinet Economy

[2] Christensen, T. H., Gram-Hanssen, K., & Friis, F. (2013). *Households in the smart grid – existing knowledge and new approaches*. In L. Hansson, U. Holmberg, & H. Brembeck (Eds.), Making Sense of Consumption. Chapter 20., pp. 333-348. Centre for Consumer Science, University of Gothenburg.

[3] H. Hino, H. Shen, N. Murata, S. Wakao, Y. Hayashi: A Versatile Clustering Method for Electricity Consumption Pattern Analysis in Hoseholds. *IEEE Transactions on Smart Grid*, Vol. 4. No. 2., June 2013, pp. 1048 – 1057.

[4] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Holdán, J. Oliver and H. E. Stanley: Analysis of symbolic sequences using the Jensen-Shannon divergence, *Physical Review E*, Volume 65, 2002, pp. 65-81.

[5] Y. Kawahara, M. Sugiyama Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation. *Proceedings of the SIAM International Conference on Data Mining*, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA (2009), pp. 389-400

[6] Bache, K., Lichman, M.: „UCI Machine Learning Repository. Individual household electric power consumption Data Set ", Irvine, University of California, School of Information and Computer Science, [http:/archive.ics.uci.edu/ml], 2013

[7] I. Csiszár, J. Körner: *Information Theory - Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest, 1981.

[8] I. Deák: *Random Number Generators and Simulation.* Akadémiai Kiadó, Budapest, 1990.

[9] I. Pinter, L. Kovacs, A. Olah, R. Drenyovszki, D. Tisza and K. Tornai: Application of Jensen-Shannon Divergence in Smart Grids. 5th International Scientific and Expert Conference TEAM 2013, Prešov, 4th to 6th November 2013, Slovakia