Improving Forecasting Accuracy Through the Influence of Time Series Representations and Clustering

Peter Laurinec*
Institute of Informatics, Information Systems and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
peter.laurinec@stuba.sk

Abstract

Time series data is the type of data that is produced in large amounts. Sensors, web traffic, the economy, and business are all resources of time series data. One of the main tasks of time series data mining is forecasting future values given current observations. The energetics domain is one of the many forecasting use cases. For more accurate decision making and control in the energetics domain, sensors called smart meters are deployed in countries that creates a large amount of time series data. These data have also characteristics, besides volume, of variety and velocity. A large number of patterns and the data acquisition rate are significant factors that influence a processing of smart meter data. For this reason, the thesis is focused on taking advantage of smart meter data as much as possible regarding above mentioned problems. For solving these problems, we are using and proposing new time series data mining methods as time series representations and clustering for improving forecasting accuracy of electricity consumption.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Time series analysis; H.2.8 [**Database Applications**]: Data mining; I.5.3 [**Clustering**]: Algorithms; G.3 [**Probability and Statistics**]: Correlation and regression analysis

Keywords

time series data mining, forecasting, cluster analysis, data streams, time series representations

*Recommended by thesis supervisor: Prof. Assoc. Mária Lucká
To be defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on [To be specified later].

1. Introduction

An accurate forecasting of electricity consumption is important for many companies and organisations around the world, due to economic, technical and environmental reasons. Practitioners and salespersons can be interested in the forecasting of a global aggregated consumption, or aggregated consumption in a small area, or even in the consumption of individual end-consumers. For some companies and producers of electricity, end-consumer consumption forecasting is the most important task.

The electricity consumption data from smart meters, with measurements coming every 15 or 30 minutes, has its own special characteristics and its forecasting is challenging for these reasons: a) High-dimensionality - long time series; b) Double-seasonality - daily and weekly pattern of consumption; c) There is also yearly pattern - holidays, vacations, change of annual seasons; d) Large number of variable consumers in the grid; e) Lot of other stochastic factors influencing consumption - weather, unexpected vacations, blackouts, market changes, special events (hockey, football) etc. One more important note concerns the character of smart meter data. Data from smart meters are coming fast and can evolve (change) dramatically. For these reasons, they can be referred as data streams. Therefore, forecasting methods have to adapt to changes automatically and be fast to compute.

There are a number of possible different forecasting methods that can be used on our problem. The choice of the most appropriate one is not a trivial task. A no less challenging task is the feature selection and engineering for the most appropriate model creation. Possible solutions to these problems give a combination of multiple models from multiple forecasting methods. This procedure is called ensemble learning and we analysed it further in this work.

There is another interesting question. Can we take advantage of the large number of data from all consumers to improve the forecasting accuracy of electricity consumption? Since there is available a large number of long time series from smart meters, time series data mining methods can be also used here. Specifically, we focused on the analysis of time series clustering and representations. The clustering was used for the creation of a more predictable (forecastable) groups of consumers. On the other hand,

the time series representations help us perform more accurate clustering.

Improving accuracy of forecasting aggregate electricity consumption through consumers clustering was already studied. The work of Ilic et al. proofs that the size of the customer base has an impact on the accuracy of forecasting methods [10]. Shahzadeh et al. deal with clustering of consumers in three different ways of feature extraction from time series and its impact on the accuracy of the forecast of energy consumption [19]. The best results achieved the clustering with regression coefficients, which showed significant improvements in the accuracy of the forecast with the help of clustering. Wijaya et al. are using for clustering correlation-based feature selection as the representation of consumers [21]. They want to investigate the impact of aggregation on the accuracy of the forecast. As forecasting methods, linear regression, multi-layer perceptron and support vector regression were used. Smart meter data clustering for improving forecasting accuracy with the ODAC method [17] and the neural network was used in work of Rodrigues et al. [9].

In the thesis, we are interested in improving the data mining workflow for forecasting a large amount of time series or forecasting a time series with a help of a large amount of time series data. For this task, beyond previous works, we are using time series representations, ensemble learning methods for forecasting, and multiple data streams clustering.

This extended abstract is organised as follows. Section 2 presents various model-based time series representations. In Section 3, new unsupervised ensemble learning methods are proposed, while in Section 4, new multiple data streams clustering method is proposed. Section 5 covers proposals of new clustering-based forecasting method for individual consumers. This abstract ends with conclusion in Section 6.

2. Model-based Time Series Representations

One of our main contributions are proposals of new time series representations that are suitable for double-seasonal time series of electricity consumption. Time series representations are crucial for accurate and effective clustering of consumers, as is the extraction of typical consumption patterns for improving forecasting accuracy.

Various model-based representations were proposed that are based on regression models or time series analysis methods [13]. By model-based representation the seasonal profile(s) is (are) extracted. Multiple linear regression (LM), robust linear regression (RLM), quantile regression (L1), generalized additive model (GAM), Holt-Winters exponential smoothing (HW), and simple average or median seasonal profile were all used for an extraction. In Figure 1, the comparison of seven different model-based representations on randomly picked Irish residential consumer are shown. The source code of all implemented representation methods is available online[1].

Clustering time series representastions by K-means [15] was used for a creation of new clustered time series as training set to forecasting methods. Experiments were performed on datasets from Slovakia and Ireland. We

[1] https://github.com/PetoLau/TSrepr

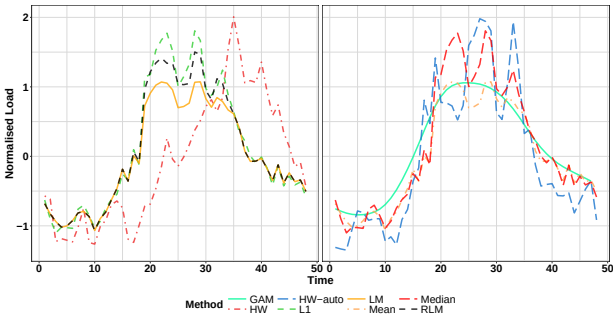

Figure 1: Comparison of seven model-based representations. On the vertical axis is normalised consumption and on the horizontal axis is the measurement during the day.

proved that the accuracy of forecasting aggregate electricity consumption with clustering significantly outperformed forecastig with simple aggregation of all consumers. We also identified that model-based representations are more suitable for this task than most nondata adaptive and data adaptive representations [8]. Created clusters of consumers preprocessed by median daily profile representation can be seen in Figure 2. The 16 aggregated time series within clustering created in previous figure are shown in Figure 3.

3. Unsupervised Ensemble Learning

There are several suitable methods for electricity consumption forecasting such as time series analysis and regression methods. Both types of them have their limitations such as an inability of adaptation to sudden changes (concept drift) and the noisy behaviour of time series. Therefore to find and choose the most suitable forecasting method is difficult. To overcome this problem, an ensemble learning is used that combines forecasts from various models (methods). Moreover, forecasting methods themselves can be tuned by the simplest of all ensemble methods - bootstrap aggregating (bagging) [3]. We have evaluated two types of methods of bagging that are based on a type of the used forecasting method: a) time series analysis method; b) regression tree. Three types of bootstrapping methods were implemented to observe their usefulness for time series analysis methods: a) moving block bootstrap with the combination of STL decomposition and Box-Cox transformation proposed by Bergmeir et al. [1]; b) our proposed smoothed version of the previous one; c) and our proposed K-means based bootstrapping [14]. We have proposed also several new ideas to unsupervised ensemble learning approaches relying on a proper combination of multiple bootstrap forecasts. The source code of all implemented ensemble learning methods is available online[2].

There are other widely used ensemble learning methods, which are an error based [12], so they weight each prediction method by its performance. We proposed ensemble learning methods, which are structure based, so it uses unsupervised learning to create a final ensemble forecast. As we have found and experimentally proven, only unsupervised approaches are suitable for time series created

[2] https://github.com/PetoLau/UnsupervisedEnsembles

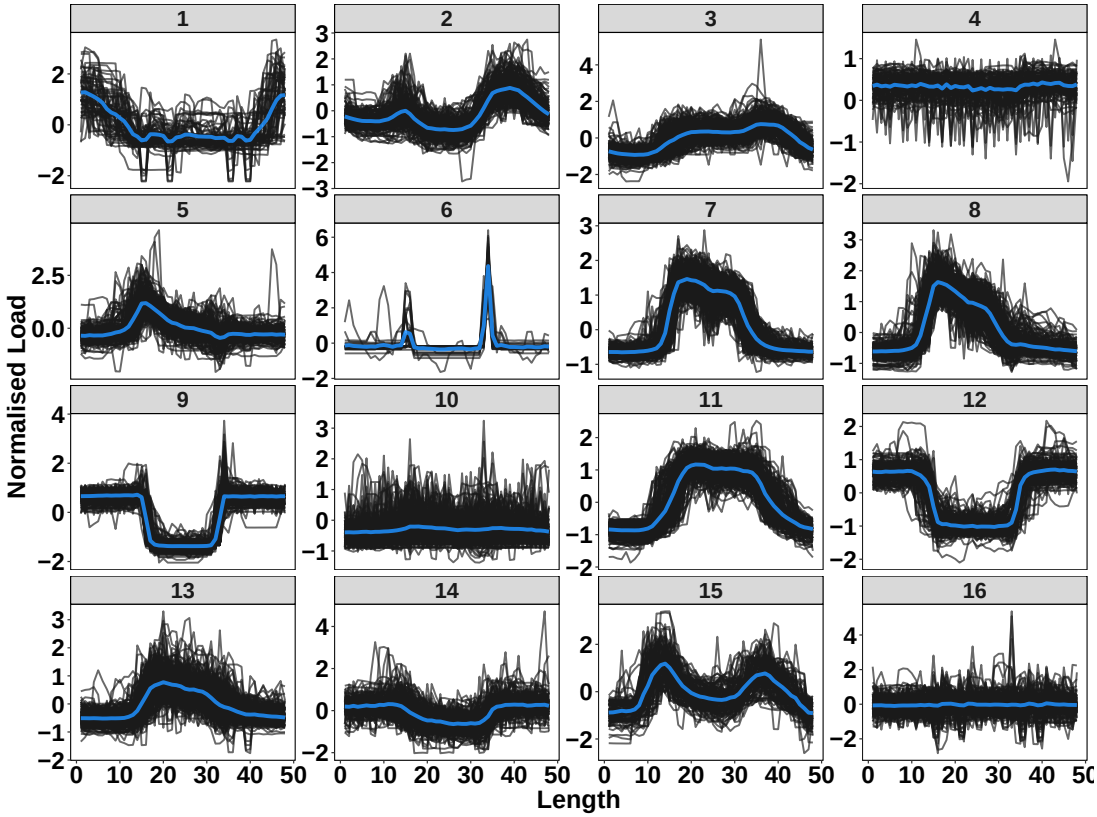

Figure 2: 16 clusters of median daily profile representations of consumers time series created by K-means. Centroids are drawn by a blue line.

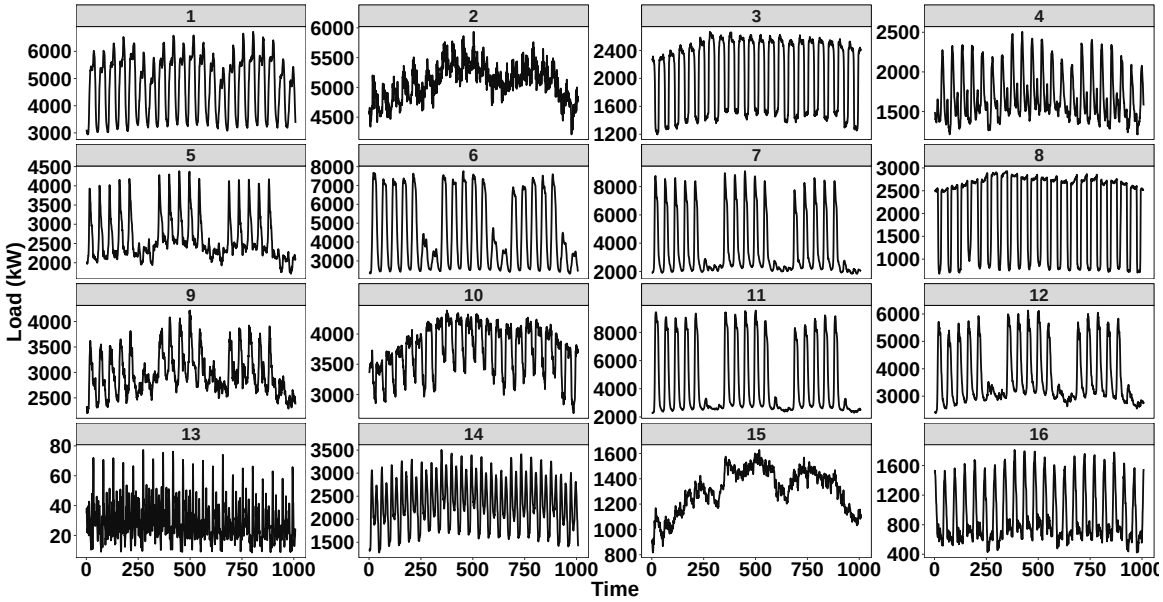

Figure 3: Final aggregated time series within clusters based on Fig. 2 which will give input to forecasting methods.

by clustering, which is newly generated in each data window. Reason for this claim is that each created clustered time series needs to apply different forecasting method.

We have evaluated our approaches again on datasets from Slovakia and Ireland. The forecasting accuracy of aggregated and also clustered electricity consumption was tested. On Irish dataset, our smoothed version of mbb had best and significant results. On Slovak dataset, our proposed K-means bootstrapping had best results, however, failed to overcome base (unbootstrapped) models. For this reason, we conclude that unsupervised ensemble learning is not suitable for every type of aggregated and also clustered time series forecasting, especially noisy and fluctuate ones. However, the clustering of consumers itself stably improved forecasting accuracy of all methods.

4. Multiple Data Streams Clustering

Data stream clustering is a rapidly evolving part of data mining [20]. However, most of the attention so far in the literature had been on object-based data stream clustering that focuses on the one data stream clustering. Since we are clustering multiple consumers, we are focused on attribute-based data stream clustering that clusters multiple streams. An attribute-based data stream clustering method should ideally take into account restrictions as similar to any data stream clustering algorithm as follows: a) performing fast and incremental processing of data objects; b) rapidly adapt to changing dynamics of the data (merging and removing of clusters); c) scale to the number of objects that are continuously arriving; d) rapidly detect the presence of outliers and act accordingly; e) detect changes (drifts) in streams [20].

All previously mentioned aspects have to be fulfilled, if we want to cluster and forecast electricity consumers data streams effectively - satisfies conditions of high-performance computing. These conditions were not satisfied in previous works [6, 2, 4, 17, 16]. Therefore, we proposed interpretable multiple data streams clustering method called *ClipStream* for improving forecasting accuracy and smart grid monitoring. It uses newly proposed feature extraction method from clipping representation called *FeaClip* for data stream representation. From *FeaClip* outlier consumers can be quickly and automatically detected. The K-medoids [11] is used for clustering non-outlier *FeaClip* representations. The k-sample Anderson-Darling test [18] was adapted for a change detection of aggregated time series streams. The source code of *ClipStream* method is available online[3]. The visualisation of the created *FeaClip* representation of a randomly picked consumer is shown in Figure 4. Short data stream windows resp. representations are bordered by grey dashed lines.

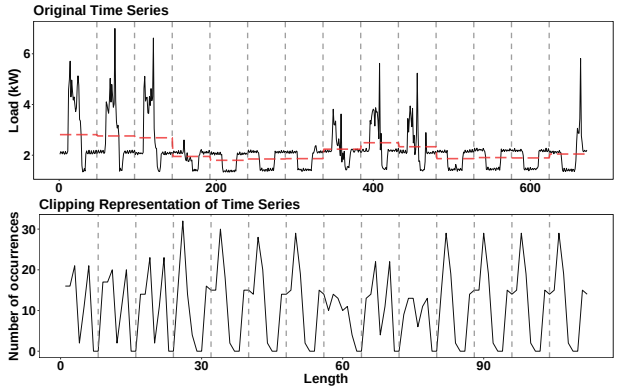

Figure 4: The original time series and its corresponding clipping representation of a randomly picked consumer. The red horizontal dashed line represents the average values of consumption during the day and the grey vertical dashed lines limit the windows of the length of one day.

The *ClipStream* behaviour and forecasting accuracy performance were evaluated on four various smart meter datasets. Four clustering benchmark approaches were also implemented for an extensive comparison. In the most cases, our *ClipStream* method had better results of forecasting than other benchmarks - in 16 from 24 cases were best. We also evaluated the *FeaClip* performance on clustering results on 85 time series datasets from the UCR repository [5]. We discussed and experimentally compared the computational complexity of *FeaClip* with other representations methods, in which *FeaClip* showed suitability for high-performance computing.

5. Individual Consumer Electricity Load Forecasting

The usage of all smart meter data and its clustering for forecasting individual end-consumer electricity load was not explored in the research papers before. We conducted to propose new clustering-based forecasting method for this task. The method uses all the data from the smart grid, cluster them, and then uses cluster prototypes as new training sets to forecasting methods. The motivation behind our approach is to overcome highly noisy and fluctuate (stochastic) character of individual consumers time series.

Before clustering, time series are normalised by z-score and normalisation parameters (mean and standard deviation) are saved for every time series (consumer). The normalised time series are then preprocessed by model-based time series presentations (e.g. L1, GAM, median daily profile). Final representations are clustered by K-means or K-medoids, and the optimal number of clusters are found by Davies-Bouldin index [7]. Centroids of original normalised time series are computed from clustering. These centroids are then used as training sets to forecasting methods. Forecasts from centroids are denormalised based on stored normalisation parameters to create final forecasts for every consumer. The source code of the proposed method is available online[4].

Our approach was compared with a fully disaggregated one, where an individual model was trained for every consumer separately. From the evaluation of experiments, we found that our approach is performing significantly better than benchmark on residential datasets. Our approach is also better scalable since it trains only K models, where K is the number of clusters (around 20-40), against a fully disaggregated approach that trains N models, where N is the number of consumers (around thousands).

6. Conclusions

The main goal of the thesis, the usage of time series data mining methods in order to improve the predictive performance of machine learning methods and its combinations applied on smart meter data, was achieved. The significant improvement of forecasting accuracy of aggregated or disaggregated consumption that was based on clustering and time series representations was proved.

The main contributions of the proposed thesis can be summarised as follows: a) new proposals of time series representations mainly for seasonal time series. In particular, new model-based and data dictated representation methods were proposed; b) the choice of the most suitable forecasting methods alongside clustering of time series; c) newly proposed ensemble learning methods; d) new interpretable multiple data streams clustering method; e) new

[3] `https://github.com/PetoLau/ClipStream`

[4] `https://github.com/PetoLau/ClusterForecast`

method for individual consumer forecasting based on clustering of all smart meter data; f) the open source software package for computing time series representations[5].

Acknowledgements. This work was partially supported by the Slovak Research and Development Agency, Grant No. APVV-16-0484 and No. APVV-16-0213, by the Scientific Grant Agency of The Slovak Republic, Grant No. VG 1/0752/14, and by the Research and Development Operational Programme for the project "International Centre of Excellence for Research of Intelligent and Secure Information-Communication Technologies and Systems", ITMS 26240120039, co-funded by the ERDF.

References

[1] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International Journal of Forecasting*, 32(2):303–312, 2016.

[2] J. Beringer and E. Hüllermeier. Fuzzy clustering of parallel data streams. *Advances in Fuzzy Clustering and Its Application*, pages 333–352, 2007.

[3] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] Y. Chen. Clustering parallel data streams. *Data Mining and Knowledge Discovery in Real Life . . .* , (February), 2009.

[5] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive. 2015.

[6] B.-R. Dai, J.-W. Huang, M.-Y. Yeh, and M.-S. Chen. Adaptive clustering for multiple evolving streams. *IEEE Transactions on Knowledge and Data Engineering*, 18(9):1166–1180, 2006.

[7] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[8] P. Esling and C. Agon. Time-series data mining. *ACM Computing Surveys*, 45(1):1–34, 2012.

[9] J. Gama and P. P. Rodrigues. Stream-Based Electricity Load Forecast. *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*, 4702:446–453, 2007.

[10] D. Ilić, P. G. da Silva, S. Karnouskos, and M. Jacobi. Impact assessment of smart meter grouping on the accuracy of forecasting algorithms. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13*, page 673, New York, New York, USA, 2013. ACM Press.

[11] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 2009.

[12] G. Kosková, P. Laurinec, V. Rozinajová, A. B. Ezzeddine, M. Lucká, P. Lacko, P. Vrablecová, and P. Návrat. Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica*, 13(2):97–117, 2015.

[13] P. Laurinec and M. Lucká. Comparison of representations of time series for clustering smart meter data. In *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016*, pages 458–463, 2016.

[14] P. Laurinec and M. Lucká. Usefulness of unsupervised ensemble learning methods for time series forecasting of aggregated or clustered load. In A. Appice, C. Loglisci, G. Manco, E. Masciari, and Z. W. Ras, editors, *New Frontiers in Mining Complex Patterns*, pages 122–137, Cham, 2018. Springer International Publishing.

[15] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[16] C. M. M. Pereira and R. F. de Mello. TS-stream: clustering time series on data streams. *Journal of Intelligent Information Systems*, 42(3):531–566, 2014.

[17] P. P. Rodrigues, J. Gama, and J. Pedroso. Hierarchical clustering of time-series data streams. *IEEE transactions on knowledge and data engineering*, 20(5):615–627, 2008.

[18] F. W. Scholz and M. A. Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[19] A. Shahzadeh, A. Khosravi, and S. Nahavandi. Improving load forecast accuracy by clustering consumers using smart meter data. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2015.

[20] J. a. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. D. Carvalho, and J. Gama. Data stream clustering: A survey. *ACM Computing Surveys*, 46(1):1–31, 2013.

[21] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 879–887. IEEE, 2015.

Selected Papers by the Author

P. Laurinec, M. Lucká. Interpretable Multiple Data Streams Clustering with Clipping Streams Representation for the Improvement of Electricity Load Forecasting. *ECML-PKDD 2018 Journal Track, Journal of Data Mining and Knowledge Discovery. IF: 3.16.* Submitted revised manuscript.

P. Laurinec, M. Lucká. Usefulness of unsupervised ensemble learning methods for time series forecasting of aggregated or clustered load. In A. Appice, C. Loglisci, G. Manco, E. Masciari, and Z. W. Ras, editors, *In New Frontiers in Mining Complex Paterns*, pages 122âĂŞ137, Cham: Springer International Publishing, 2018.

P. Laurinec. TSrepr R package: Time Series Representations. *Journal of Open Source Software*, volume 3, number 23, page 577, 2018.

T. JarÃąbek, P. Laurinec, M. Lucká. Energy Load Forecast Using S2S Deep Neural Networks with k-Shape Clustering. *In INFORMATICS 2017. Proceedings of IEEE 14th International Scientific Conference on Informatics*, pages 140–145, 2017.

P. Laurinec, M. Lucká. New Clustering-based Forecasting Method for Disaggregated End-consumer Electricity Load Using Smart Grid Data. *In INFORMATICS 2017. Proceedings of IEEE 14th International Scientific Conference on Informatics*, pages 210–215, 2017.

P. Laurinec, M. Lóderer, P. Vrablecová, M. Lucká, V. Rozinajová, and A. B. Ezzeddine. Adaptive time series forecasting of energy consumption using optimized cluster analysis. *in Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, IEEE, pages 398–405, 2016.

P. Laurinec, M. Lucká. Comparison of representations of time series for clustering smart meter data. *in Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2016*, pages 458–463, 2016.

G. Grmanová, V. Rozinajová, A. B. Ezzedine, M. Lucká, P. Lacko, M. Lóderer, P. Vrablecová, and P. Laurinec. Application of biologically inspired methods to improve adaptive ensemble learning. *in Advances in Nature and Biologically Inspired Computing*, Springer, pages 235–246, 2016.

A. B. Ezzeddine, M. Lóderer, P. Laurinec, P. Vrablecová, V. Rozinajová, M. Lucká, P. Lacko, and G. Grmanová. Using biologically inspired computing to effectively improve prediction models. *International Journal of Hybrid Intelligent Systems*, volume 13, number 2, pages 99–112, 2016.

G. Kosková, P. Laurinec, V. Rozinajová, A. B. Ezzeddine, M. Lucká, P. Lacko, P. Vrablecová, and P. Návrat. Incremental ensemble learning for electricity load forecasting. *IActa Polytechnica Hungarica*, volume 13, number 2, pages 99–117, 2015.

[5] `https://cran.r-project.org/package=TSrepr`