Analysis of Data Cleaning Techniques for Electrical Energy Consumption of a Public Building

Dacian I. Jurj
Electrical Engineering Department
T.U.C.N.
Cluj-Napoca
dacian.jurj@ethm.utcluj.ro

Dan D. Micu
Electrical Engineering Department
T.U.C.N.
Cluj-Napoca
Dan.Micu@ethm.utcluj.ro

Levente Czumbil
Electrical Engineering Department
T.U.C.N.
Cluj-Napoca
Levente.CZUMBIL@ethm.utcluj.ro

Alexandru G. Berciu
Electrical Engineering Department
T.U.C.N.
Cluj-Napoca
alexandru.george.berciu@gmail.com

Mircea Lancrajan
Electrical Engineering Department
T.U.C.N.
Cluj-Napoca
lancranjanmircea98@gmail.com

Denisa M. Bărar
Applied Computational Intelligence
U.B.B.
Cluj-Napoca
denisabarar@gmail.com

***Abstract*—Statistical Techniques and Artificial Intelligence are becoming much more a necessity in a fastened world rather than just a theoretical use case. In order to satisfy this need, the optimization process starts with data collecting and cleaning. The aim of this paper is to provide a short overview of the outlier detection methods and to explain the need for data cleaning in the field of energy consumption by analyzing the energetic profile data from the Technical University of Cluj-Napoca's swimming complex. In the first and second parts of the article, a short overview of cleaning methods are presented. The third part compares the efficiency of the proposed methods. Finally, but not least the fourth part of the article is dedicated to conclusions and future work.**

Keywords—Data Cleaning, Electricity Consumption, Outliers, LOF, IQR, Clusters, Median, Public Buildings, Cleaning Methods, Artificial Intelligence, Machine Learning, Database

I. INTRODUCTION

Before developing high-end hardware and software machines, data analysis was often just an exercise of applied theoretical algorithms on various dummy data sets for validating an isolated perspective. Nowadays, the theory stands, yet the dummy data has been replaced by real-time data that needs to prove its efficiency in future analysis. In the process of getting the real-time data cleaning has become an impetuous step in order to avoid a "Garbage In, Garbage Out" scenario. Having data gaps, outliers, missing values or outranged values can be a consequence of data entry, measurement, distillation, or data integration errors [1, 2]. With the help of the newly rapid development of cloud computing technologies [3], storage has become a wildly used application that facilitates most of the companies to collect and to store large packs of data. With a large volume of data, the probability of error occurrence is increased, and dirty data can lead to wrong decisions, and questionable analysis which makes data quality to be a major concern. Other types of common errors are typos, mixed formats, replicated entries and violations of business rules which analysts need to take into consideration as the key point when exploring the research side of the databases [4, 5]. In the process of forecasting the Technical University of Cluj-Napoca's swimming complex energetic profile we encountered gaps and abnormal values in the data sets given by our data feed provider. Because the given issue we decided that an outlier detection analysis will be requested and implemented before doing any forecast. The contribution of this paper is to identify efficient outlier techniques over an energy data set through an inside built intelligent scoring algorithm.

II. OUTLIER DETECTION TECHNIQUES

From the vast literature focused on outlier detection some definitions could be summarized; as a general one by Barnett and Lewis where they define an outlier as an observation or a set of observations which appear to be inconsistent with that set of data [6] or in short lines "Outliers do not equal errors. They should be detected, but not necessarily removed. Their inclusion in the analysis is a statistical decision" [7]. From a more consistent understanding of data we should take into consideration that outliers are indeed not necessarily to be removed or replaced in some cases, they can be a consistent observation in the long term, as a response for this, many different outlier detection methods were developed in the literature. [8, 9]. A detailed overview of methods used in outlier detection was presented in and distributed as: probabilistic models with parametric and nonparametric approaches, statistical models, and Machine Learning algorithms with clustering-based and classification-based techniques [10].

A. Probabilistic models

Probability distribution functions were proposed to detect outliers as datapoints which have the highest probability to be outside the given threshold. In the two types of probabilistic approaches; parametric in which the data is analyzed with a predefine known distribution function and the nonparametric where the data is estimated based on a density or distance function , deviations are consider anomalies because they are not behaving like the majority of the tested population (data points) [11,12]. Gaussian distribution functions and median absolute deviation are usually applied in parametric probabilistic outlier detection methods [11]. Because most of the distributions are univariate and the primary distributions of the observation need to be known in advanced, probabilistic parametric models are failing to deliver when the data is not known [13].

978-1-7281-1078-3/20/$31.00 ©2020 IEEE

The method of detecting outliers using the median, although similar to the mean, is very insensitive to the presence of outliers by calculating the central tendency. The median method together with Median Absolute Deviation (MAD) represents also the statistical dispersion of a data set, being much more robust than the mean and standard deviation methods. To determine the insensitivity of the median method, one of the indicators used is the "breakdown point" (see, e.g., Donoho & Huber, 1983) [14]. This indicator represents the maximum number of contaminated data that can be in a dataset, without affecting the final result. For example, as a comparison, if a record in a data set is of infinite value, the method of the mean gives an infinite result, while the method of the median has an unchanged result. The only problem for the median method arises only when more than half of the values are of infinite values [15, 16].

Box-and-whiskers plots Fig.1 could be used for outlier detection as a numeric and graphical approach for tedious large sets of data and require human analysts in order to get accurate results.

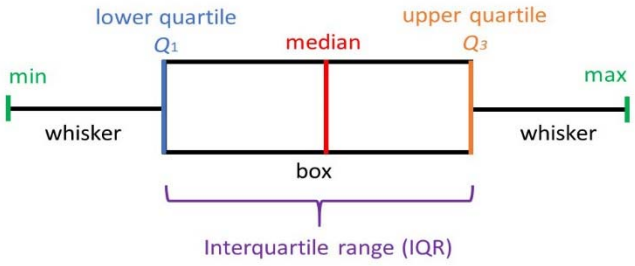

Fig. 1. Box-and-whiskers plot

The output can generally cut through a high density of outliers based on the studied time horizon still it's a good graphic indicator if it is tested for different time stamps. To have better accuracy in the process of data outlier detection automated techniques have been developed in the literature [9, 36].

The non-parametric methods are giving a more widely view being tested on multidimensional datasets [17,18] and can be combined with various clustering techniques as k-nearest neighbor [19], density based approaches (takes only global view of data sets) using kernel estimation and Parzen window [20-22] yet they are usually computationally expensive. Apart from the most common nonparametric methods: ranking or scoring data, based on differences and similarities [23]; Gaussian mixture models [24]; and probabilistic ensemble models using the density based local outlier factor (LOF) with distance based as k-nearest neighbor, are used [25].LOF itself is a calculation method that looks how close a certain point is to other points in its vicinity, in order to obtain the local data neighborhood density. This density is then compared to the density of the other points later on. This whole procedure is determined by a number *k*. This parameter dictates if the outlier detection will have a more local focus, in which case we use a smaller *k*, which is more erroneous the more noise is in the data. A large *k* however can miss local outliers. In this paper we treat multiple cases of *k* values to determine a suitable one for our dataset. In the context of data cleaning for electrical energy consumption this algorithm offers an interesting opportunity beyond merely identifying problematic, faulty or abnormal consumption readings as LOF does not treat the property of being an outlier a binary property. Thus, the result could not only be used to identify but also used to determine an adjustment factor [26].

B. Statistic approuaches

Residual analysis can be a good indicator for outlier detection when using statistical methods like auto regressive moving averages [27-30] even if it is hard to detect the polynomial function for the real time series data [31, 32]. Linear regression models were proposed in [33] where the dependent variables are the electric load consumption and the independent variables are the weather input. Because most of the times the parameters are calculated based on historical data, in general the statistical algorithms from the literature were developed only for offline anomaly detection, even if some of them were described as heavy online anomaly detection methods used in wireless networks [34, 35].

C. Machine learning algorithms

On the more advanced artificial intelligence perspective machine learning with supervised (classification based) and unsupervised learning (clustering based) were used to detect outliers in fraud, health care, image processing and networks intrusions [36-38]. In the clustering approaches each data point is assigned with a degree of belonging to each data cluster. The anomalies are detected by comparisons to the given clusters threshold and understanding the associateship of the tested data. A good example of a clustering method is the *k*-means algorithm where selecting the top *n* points that are situated on a biggest distance from the nearest cluster as outliers [39].

The most feasible Machine Learning techniques for anomaly detection in an unsupervised environment are the clustering-based approaches represented by models like *k*-means [MacQueen] [40] or DBSCAN [Ester] [41]. DBSCAN (Density-based spatial clustering of applications with noise) was proposed in 1996 and proved excellent results in extracting the density information from data. DBSCAN presents some advantages over *k*-nearest neighbors' algorithm such as automatically adjusting the number of clusters to be computed and the ability to isolate the outliers in individual clusters.

DBSCAN classifies the data points in three groups: core points, border points, and outlier points. This model divides the samples into different classes based on the proximity of the samples and by considering the two input parameters: the ε (eps) parameter that represents the maximum distance between two samples and the minimum points that represents the number of samples in a neighborhood for a point to be considered as a core point. A sample is considered a border point if it is not a core point or an outlier point but it is a part of the cluster. As follows, the outlier points are the remaining points [40,41].

In the AI literature we can also find Machine learning classification-based approaches like neural networks and support vector machines to mimic classifiers that are for anomaly detection. Neural networks were used in various domains [42-44] with the advantage of clearly differentiating between different outliers' classes even if they need a rigorous definition for the cost function. In case of support vector machines, the algorithm is looking for the optimum

hyperplane that split two adjacent data classes [44] and finds the maximum margin necessary to separate them.
Hybrid methods were also proposed for finding outlier detection like Bayesian classifier which combine probabilistic and machine learning algorithms where Bayes" theorem is applied between the features and the given classes [45].

III. Applied Outliers Detection Techniques

A. Proposed outlier detection techniques

In the process of finalizing DR-BoB "Demand Response in Blocks of Buildings" project funded by the EU Horizon 2020 innovation program under grant agreement No. 696114/2016 [46,47] data was collected from Technical University of Cluj-Napoca's (TUCN) buildings in order to develop an energy monitoring tool and targeting system with a Demand Response curve control strategy. In the process of gathering the electrical consumption data from the swimming complex of the Technical University of Cluj-Napoca, there were detected inconsistent data (see Fig. 2). "You can observe a lot by just watching"-Yogi Berra:

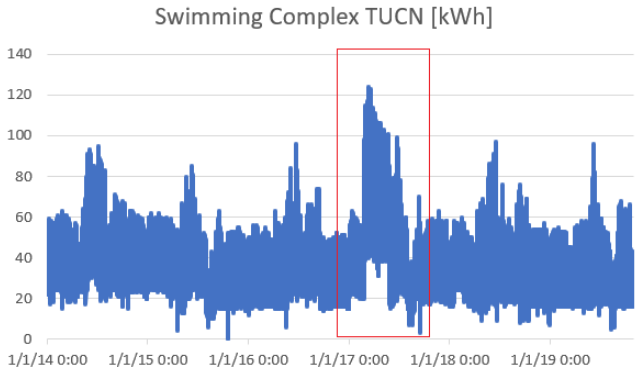

Fig. 2. Energy consumption of the Swimming Complex of the Technical University of Cluj-Napoca

In order to confirm our descriptive observations, we proposed the LOF method to understand at more than a binary level the distribution of the data. To avoid seasonality scenarios, the data was distributed for each year to be tested. In our analysis we used the value of *k* equal with 2,3,4,5,25,50 and we determine that the most suitable values for our scenario was 25 even if in the literature the most used parameters are usually 2 and 3 [26]. The reason for selecting a higher positive integer *k* was determined by the fact that we wanted to see the anomalous data from a more global perspective. It has been observed that on average 20% of our data is outlier. We consider that a point has a higher probability to be an outlier if it was selected by all the *k* values scenarios. After the first round of analysis there has been observed that from all outliers' techniques anomalous data was detected for more than one month in 2017.The interquartile range method and median were also applied on the same data set and a similar outlier detection was observed, on average more than 20% of our data was an outlier. Before adding the final results from detection algorithms, the data feeds were rechecked and the administration of the building was questioned in order to understand if a potential event could have been occurred in the tested period. During our investigation we determined that the anomaly from the data was determined by a mistaken collecting feed issue.

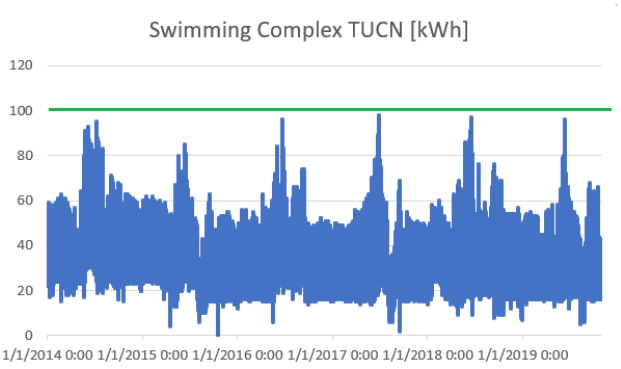

Fig. 3. Adjusted energy consumption of the Swimming Complex of the Technical University of Cluj-Napoca

After checking and adjusting the data feed (see Fig. 3 the correct data feed), Interquartile range IQR and Median methods were recompiled to identify the real anomalous data. The testing was conducted over each year from the beginning of 2014 and also for the whole data set. It was observed that on average 552 data points from a total of 51120 were detected to be outliers, which lead to a percentage of 0.9% of the analyzed data (see Table 1):

TABLE I. The Number of Outliers Detected Using IQR and Median Methods

Outliers IQR/Year						
2014	***2015***	***2016***	***2017***	***2018***	***2019***	***Total***
140	149	74	92	123	103	681
Outliers Median/Year						
2014	***2015***	***2016***	***2017***	***2018***	***2019***	***Total***
103	65	49	63	105	38	423

The LOF method was recompiled after the data adjusting process and it has been observed that on average 2576 outliers were detected Table 2. It can be observed that there is still a large gap difference of 2024 detected outliers between the interquartile range IQR/Median methods and the LOF method which conducted into a different type of analysis.

TABLE II. The Number of Outliers Detected Using LOF

Potential LOF	***2014***	***2015***	***2016***	***2017***	***2018***	***2019***	***Total***
k=2	561	538	578	596	703	541	3517
k=3	227	219	268	276	366	278	1634
k=25	10	10	2	6	11	4	43

To understand our data more, we also chose a clustering-based method, DBSCAN to find the anomalous data. As entry parameters 0.5 for epsilon (ε) as a default value and various minim number of points were used: 5, 10, 20 and 50 respectively. It was observed that the most relevant outliers were detected using 5 minim points. In order to cover more outliers, the process was conducted on different data variation

which included tuples based on registered consumption value and hour or registration day and consumption value.

Fig. 4. Estimated number of clusters by value and hour.

The results showcased an estimated number of 843 clusters with a silhouette score of 0.97 for consumption value and hour which detected 614 outliers (see Fig. 4). The same exercise was done for the consumption values and weekdays and the output showcased 359 clusters with a silhouette score of 0.98 and 285 outliers.

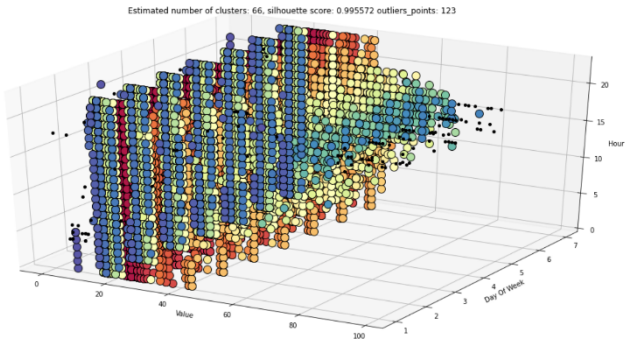

Fig. 5. Estimated number of clusters by value, hour and weekday.

Having a multi parameter approaches, a testing was conducted in outlier detection using consumption values with days of week and hours together, Fig. 5. The estimated number of clusters being equal with 66 having a silhouette score of 0.99 and a number of 123 outliers. It is important to specify that a higher silhouette scores indicate a high accuracy which qualifies the current testing as a relevant result.

B. *Intelligent Scoring Method*

After the first iteration we considered that the most relevant outliers are those who have the highest probability to occur in most of the tested scenarios. In order to create an automated system for anomalous data detection we decided to use an incidence factor for the values which are not in the pattern of the majority of data, in our case, for the tested data and methods we have the total number of common outliers equal to 123.

In order to avoid counting as outliers the natural energy peaks, an intelligent scoring method has been implemented. The method was designed to take the outputs from any outlier detection technique and to compare them with the average energy consumption over four different data clusters for the same time interval (hour) and similar working or weekend days. The first cluster contains data for days from the same year and the same 2-month period as the investigate outlier data point. The second cluster contains energy consumption data from the same year and the same season (winter or summer). The third and fourth clusters contain data from the entire historical data set for the same 2-month period and for the same season respectively. The aim of the exercise is to validate the outlier data points through a scoring process if they are unusual consumption and/or damaged data values. The score over each data cluster is distributed from 0 to 5: 0 meaning that the outlier detection output data point is a usual energy consumption data (an invalid outlier) while 5 means the outlier consumption data is much higher/smaller than 90% of the cluster data points. A final scaled score was evaluated from the individual score over each analyzed data cluster.

C. *Final Results and Resolutions*

Because the last iteration was not enough to cover all the outliers and to help us understand which should be the best method or combination of methods for errors detection, we decided to filter the initial results through the proposed intelligent scoring method. Having a small amount of data, we combined IQR and Median methods results into a single data base for the computation. It has been observed that from a common ground of 413 outliers only 322 were validated trough the system. The same process was conducted over the combined data of DBSCAN results and 95.2% of the observation were validated as real outliers. Because the LOF method had a larger number of detected issues we decided to run the K2 and K3 databases independently. For both of the LOF computations the results were lacking accuracy having only 755 valid outliers out of 3512 detected for K2 and only 292 outliers out of 1628 for K3 Table 3.

TABLE III. SCORED OUTLIER ACCURACY

Applied Method	Total Outliers	Valid		Invalid	
		Data Points	*[%]*	*Data Points*	*[%]*
IQR/Median	413	322	78	91	22
DBSCAN	728	693	95.2	35	4.8
LOF K2	3512	755	21.5	2757	78.5
LOF K3	1628	292	17.9	1336	82.1

TABLE IV. UNIQUE AND COMMON OUTLIERS

Method	Unique	Overlap	Common Outliers
IQR/Median	72	4 Methods	23
DB SCAN	393	3 Methods	63
LOF K2	458	2 Methods	416
LOF K3	43	Total	1468

After the scoring process the valid output was analyzed in one databased. It has been observed that from all the methods we have a total of 1468 unique valid outliers. Some of the methods validated the same data point as an anomaly: 23 common data points were detected by all the methods, 63 by three of them and 416 by any two methods that had a common

value Table 4. Having this common result between the methods we can be sure that there is a certain number of 502 anomalous data points in the data set.

IV. Conclusion

This paper presented various applied outlier detection methods for determining the sanity of the data collected from Technical University of Cluj-Napoca's swimming complex and to prepare it for a future forecast exercise. During the analysis we managed to understand the need for an intelligent scoring method as the presented outlier detection methods were unable to differentiate between the natural energy peaks and anomalous data. The exercise is enforcing the idea that the outlier methods are not giving a high accuracy in a universal usage approach. For the current test we obtain the highest accuracy for the BDSCAN method. The LOF method will be reviewed in our next uncases, if low accuracy persist it will be removed from future work. For future analysis we will extend our outlier detection adding up more methods and new data sets collected during the DR-BoB "Demand Response in Blocks of Buildings" project.

V. ACKNOWLEDGEMENT

Renewable Cogeneration and Storage Technologies Integration for energy Autonomous Buildings, 815301-RE-COGNITION / H2020-LC-SC3-2018-RESTwoStages, 2019-2022.

References

[1] S.LakshmiMphil, Dr.S.v Prof, " An Overview Study on Data Cleaning, Itss Types and Its Methods for Data Mining," International Journal of Pure and Applied Mathematics Volume 119 No. 12 2018, 16837-16848J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Berkhin, P. (n.d.), "A Survey of Clustering Data in Mining Techniques. Grouping Multidimensional Data," August 2002, pp 25–71

[3] Su, L., Li, L., Zhang, L., & Nie, X, "Research and Design of Electric Power Private Cloud Data Storage Model," 2012 Fourth International Conference on Computational and Information Sciences, 2012

[4] Johnson, T., & Dasu, T, "Data quality and data cleaning," Proceedings of the 2003 ACM SIGMOD International

[5] ErhardRahm, HongHaiDo, "Data Cleaning Problems and Current Approaches," IEEE Data Engineering Bulletin, Volume 23, Decembre 2000

[6] V. Barnett and T. Lewis, "Outliers in statistical data," Wiley, New York, NY, 3rd edition, 1994

[7] Edwin de Jonge, Mark van der Loo, "An introduction to data cleaning with R," Statistics Netherlands, The Hague/Heerlen 2013

[8] Hawkins D.M, "General theoretical principles. In: Identification of Outliers. Monographs on Applied Probability and Statistics", Springer, Dordrecht 1980

[9] N. K. Jajo, "Graphical display in outlier diagnostics, adequacy and robustness," Statistics and Operations Research Transactions, SORT 29 (1) January-June 2005, pp 1-10.

[10] Akouemo Kengmo Kenfack, Hermine Nathalie, "Data Cleaning in the Energy Domain," Dissertation thesis, Marquette University, Spring 2015

[11] G. Buzzi-Ferraris and F. Manenti, "Outlier detection in large data sets," Journal of Computers and Chemical Engineering, 2010, pp. 388–390

[12] Rousseeuw, P. J., & Hubert, M, "Robust statistics for outlier detection," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, January 2011, 1, (1), pp 73–79

[13] M. Markou and S. Singh, "Novelty detection: A review - part 1: Statistical approaches," Journal of Signal Processing, 2003, pp 2481–2497

[14] Donoho, D. L. and Huber, P. J, "The notion of breakdown point," In A Festschrift for Erich L. Lehmann (P. J. Bickel, K. Doksum and J. L. Hodges, Jr., eds.), 1983, pp 157–184

[15] Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013), "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," Journal of Experimental Social Psychology, 49(4), 2013, pp 764–766.

[16] Tang G., Wu K., Lei J., Bi Z., & Tang J, "From Landscape to Portrait: A New Approach for Outlier Detection in Load Curve Data," IEEE Transactions on Smart Grid, 5(4), 2014, pp 1764–1773

[17] E. N. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," In Proceedings of the International Conference on Very Large DataBases, 1998

[18] E. N. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," The International Journal on Very Large DataBases, 8:237–253, 2000.

[19] S. Ramaswamy, R. Rastogi, and K. Shim, "E□cient algorithms for mining outliers from large data sets," In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data ACM Press volume 29, 2000, pp 427–438

[20] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation" Journal of Knowledge and Information Systems, 26(2):309–336, 2011.

[21] S. Liu, M. Yamada, N. Collier, and M. Sugiyama, "Change-point detection in time-series data by relative density-ratio estimation," Journal of Neural Networks, 2013, pp 72–83

[22] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," In Proceedings of the 4th IEEE International Conference on Artificial Neural Networks volume 4,1995, pp 442–447

[23] P. W. Wilson, "Detecting outliers in deterministic nonparametric frontier models with multiple outputs," Journal of Business & Economic Statistics, 1993, 11, (3), pp 319–323

[24] L. Tarassenko, P. Hayton, N. Cerneaz and M. Brady, "Novelty detection for the identification of masses in mammograms," 1995 Fourth International Conference on Artificial Neural Networks, Cambridge, UK, 1995, pp. 442-447

[25] M. Bouguessa, "A probabilistic combination approach to improve outlier detection," In 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, volume 1, 2012, pp 666–673

[26] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J, "LOF: identifying density-based local outliers", In ACM sigmod record, volume 29, May 2000, pp. 93–104.

[27] C. Fauconnier and G. Haesbroeck, "Outliers detection with the minimum covariance determinant estimator in practice," Journal of Statistical Methodology, 6,(4), 2009, pp 363–379

[28] A. Gran´e and H. Veiga, "Wavelet-based detection of outliers in financial time series," Journal of Computational Statistics and Data Analysis, 2010, pp 2580–2593

[29] A. R. Weekley, R. K. Goodrich, and L. B. Cornman,"An algorithm for classification and outlier detection of time-series data," Journal of Atmospheric and Oceanic

[30] A. Zaharim, R. Rajali, R. M. Atok, I. Mohamed, and K. Jafar, "A simulation study of additive outlier in ARMA (1,1) model," International Journal of Mathematical Models and Methods in Applied Science, 2009, 3, (2), pp 162–169

[31] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, "Time Series Analysis: Forecasting and Control," John Wiley & Sons, Hoboken NJ USA 4th edition, 2008.

[32] G. E. Schwarz, "Estimating the dimension of a model," The Annals of Statistics, 6,(2),1978, pp 461-464

[33] Akouemo, Hermine N. & Povinelli, Richard J., "Probabilistic Anomaly Detection in Energy Time Series Data," January 2015, unpublished.

[34] H. Liu, S. Shah, and W. Jiang, "On-line outlier detection and data cleaning," Journal of Computer and Chemical Engineering," 28, (9), 2004, pp 1635–1647

[35] K. Yamanishi, J. Takeuchi, and G. Williams, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM Press, 2000, pp 320–324

[36] S. Velilla, ”A note on the behaviour of residual plots in regression,” Statistics &Probability letters, 1998, pp 37:269

[37] S. Hawkins, H. He, G. Williams, and R. Baxter, “Outlier detection using replicator neural networks,” Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science, 2002, pp 170–180

[38] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgensen, and J. Ucles “A hierarchical network intrusion detection system using statistical preprocessing and neural network classification,” In Proceedings of IEEE Workshop on Information Assurance and Security, 2001, pp 85–90

[39] S. Chawla and A. Gionis, “k-Means: A unified approach to clustering and outlier detection,” In The 13th SIAM International Conference on Data Mining, 2013 pp 189–197

[40] MacQueen, James B, “Some methods for classification and analysis of multivariate observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, July 1965, pp 281-298

[41] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density based algorithm for discovering clusters in large spatial databases with noise,” in KDD-96 Proceedings, 1996, pp 226-231

[42] S. Hawkins, H. He, G. Williams, and R. Baxter, “Outlier detection using replicator neural networks”., Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science, 2002, pp170–180

[43] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, “Novelty detection for the identification of masses in mammograms,” In Proceedings of the 4th IEEE International Conference on Artificial Neural Networks, volume 4, 1995 pp 442–447

[44] I. Rish, “An empirical study of the naive Bayes classifier,” In IJCAI 2001 124 workshop on empirical methods in artificial intelligence, volume 3, 2001, pp 41–46

[45] B. Bârgăuan, M. Creţu, O. Fati, A. Ceclan, L. Dărăbant, D.D. Micu, D. Şteţ & L. Czumbil, “Energy Management System for the Demand Response in TUCN Buildings,” 53rd International Universities Power Engineering Conference, September 2018

[46] B. Bârgăuan, O. Fati, A. Ceclan, D.D. Micu, D. Şteţ, L. Czumbil & P. Mureşan, ”Demand Response on Blocks of Buildings – Romanian Pilot Site Innovation Project,” 7th International Conference on Modern Power Systems (MPS), June 2017