Extended Analysis of Data Cleaning for Electrical Energy Consumption Data of Public Buildings

Dacian I. Jurj [1], Alexis Polycarpou [2], Levente Czumbil [1], Alexandru G. Berciu [1], Mircea Lancrajan [1], Denisa Barar [3], Dan D. Micu [1*]

[1] Electrical Engineering Department Technical University of Cluj-Napoca, Cluj-Napoca, Romania
[2] Department of Electrical and Computer Engineering and Informatics, Frederick University, Nicosia, Cyprus
[3] Applied Computational Intelligence, Babeș-Bolyai University, Cluj-Napoca, Romania
* Dan.Micu@ethm.utcluj.ro

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

Abstract

In the past years both companies and academic's communities pulled their efforts in generating input that consist in new abstractions, interfaces, approaches for scalability and crowdsourcing techniques. Quantitative and qualitative methods were created with the scope of error reduction and were covered in multiple surveys and overviews in order to cope with outlier detection. The aim of tis paper is to provide an outlier analysis over the consumption data of twelve public buildings from the Technical University of Cluj-Napoca, collected during an EU 2020 Horizon project.

1 Introduction

Based on the Global Energy Statistical Yearbook 2020 [1], the acceleration in energy consumption had a growth rate of 0.7% globally in 2019. The United States had a record high in 2018 of an additional 3.5% and China growth of 3.7% while the European Union had a decreased rate of 1% and 3.5% in Germany. The need for interconnectivity and balance between production and consumption of electrical energy is an "impetuous agreement" of cleaned data. Creating a large amount of energy not only causes unbalance in the network but also comes along with energy losses. On the other side, having an unexpected peak in demand can actually harm the network by creating undesirable spikes in electricity prices and transmission congestions [2]. In the chain of generation transition and distribution there are four types of consumers in the electrical industry [3]: residential customer, commercial customer, industrial customers, and transportation customer; at the level of each layer there is a chance to encounter anomalous data which can cause harm to the electrical net. For all the three types of customers excluding the transportation sector the most common misleading errors that occur and generates the most of unknown patterns are: missing data or missing components of aggregated data, main breaks, naïve disaggregation or a stuck meter, negative energy consumption, human error, mismatched meter factor or mismatched units of aggregated data and Outliers [4]. Outliers are data points which are different from the majority of data set, basically, if there is no correlation between the energy consumption and the factors that are diving the observation data set there is a chance that the studied element to be an outlier [5].

2 Data Cleaning

For the vast majority of the cases, the presented "most common" errors are identified using outlier detection techniques. Many studies have been performed on various domains like finance, health care, communication networks, and information technology in order to determine the anomalies [6]. The literature is showcasing two types of outliers in the time series data and this will be additive outliers and innovative outliers [7]. The difference between the two types of outliers is that the additive outliers are concentrating on a single observation while the innovative outlier is induced by an event that is harming the subsequent observations [8] and occurs as the result of feedback system which provides a lousy process. Usually, the innovative outliers do not need special correction of measurement because they are noise and most of the noise is reduced when the time series data is modelled [9]. In the process of detecting outliers there is only a slight chance to be able to detect multiple outliers and this is because of the masking effect which can prevail when the outlier cannot be detected because of the presence of the others [10], yet some studies showcased that by using sequentially correction anomalies can reduce the masking effect [11, 12]. For having better accuracy in the process of data outlier detection automated techniques have been developed. A detailed summary over the type of methods used in outlier detection were presented [13] and distributed as a probabilistic approach with parametric and nonparametric approaches, statistical approaches, and Machine Learning approaches with clustering-based approaches and classification-based approaches.

3 Outliers Detection Techniques Analysis

3.1 Research overview

During DR-BoB "Demand Response in Blocks of Buildings" project funded by the EU Horizon 2020 innovation program under grant agreement No. 696114/2016 [14,15] data was collected from 12 buildings belonging to the Technical University of Cluj-Napoca to develop an energy monitoring tool and a targeting system with Demand Response curve control strategy. The selected buildings are situated in 4 different locations: Faculty of Electrical Engineering, Faculty of Building Services, Mărăști Students Campus and Swimming Pool Complex. Even if the social utility of each set of buildings is different, in the process of gathering the consumption data it has been observed that for each location there were abnormal values detected. Because one of the research purposes of collecting the data was to predict the energy consumption, an outlier detection analysis was requested before doing any forecast (Fig.1).

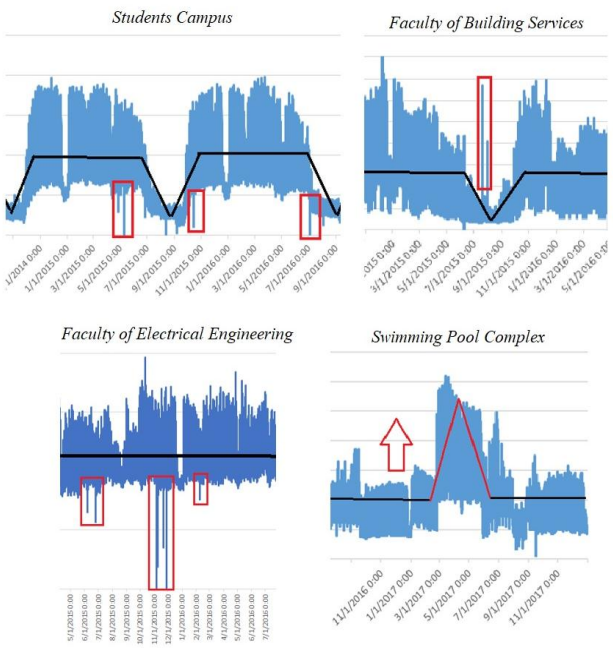

Fig. 1. Anomalous data in the consumption profile of the tested locations

In the study "Analysis of Data Cleaning Techniques for Electrical Energy Consumption of a Public Building" a detailed outlier detection empirical testing was conducted using probabilistic, statistical and machine learning techniques over the Technical University of Cluj-Napoca's swimming complex data [16]. It has been observed that the outlier detection techniques were not able to differentiate between the natural energy peaks and abnormal data without an additional support function. In order to confirm this hypothesis an extend study of the outlier detection methods was proposed for Faculty of Electrical Engineering, Faculty of Building Services and Mărăști Students Campus locations.

3.2 Proposed methodology

The proposed outlier detection techniques for the analysis are: Interquartile range (IQR), Median Absolute Deviation (MAD), Local Outlier Factor (LOF) and Density-based spatial clustering of applications with noise (*DBSCAN*). The research aim is to test some of the most common probabilistic, statistical and machine learning techniques in the context of energy consumption. To test the universal data approach of the methods we used as input parameters the most common threshold/parameters values from the literature. The IQR method helps not only in outlier detection but also in predicting the spread of energy consumption [17], yet it is tight to its mathematical limitation of identifying only the values which are between the tested threshold value. The same can be observed in the mathematical model of the MAD [17,18] where the limitations are given by the threshold values from the median value. The threshold value chosen for IQR method testing is 1.5 [19] and for median is 3. The advantage of using the IQR and MAD models is more related with "tracking" and maintaining a permeant control spread that will identify the extreme values for most of the cases. A confirmation of an outlier from both of the methods should always be take into consideration for future investigations. The LOF method [20] can achieve good results when the outliers are located in dense region of normal data which means that the accuracy of this method can be reduced when it is exposed to a high volatility data set. In our analysis we used the value of *k* equal with: 2, 3, 4, 5, 25, 50 and we determined that the more suitable k values for us would be the 2 and 3 which are the most common in the literature [21] and 25 which was empirically tested with a good accuracy compared with the other values. For the DBSCAN method [22,23] we used as input parameter 0.95 for epsilon (ε) as a default value, and the minimum number of points equal with 5. The support function or the "Intelligent scoring method" was designed to analyse the output from any outlier detection method and to decide in the context of electrical energy consumption if the detected anomalous value is a natural energy peak or abnormal data. The method can be classified as an intelligent clustering method that compares the input with the average energy consumption of four data clusters for the same interval of time and similar days (workdays or weekends). The firs cluster contains data for the same year and for the same 2-months period, the second one is composed by the data collected from the same year and same season (winter or summer), the third and the fourth clusters are composed from data from the all historical data set for the same 2-month period and respectively for the same season. After running the clustering process, the score from each value from the clusters can range between 0 and 5: 0 meaning that the detected data is a natural energy peak which makes the value an invalid outlier while 5 means that measured data is much higher/smaller than the 90% of the cluster data points. A final scaled score is determined by combining the score obtained for each cluster that was analysed.

4 Results and Discussions

4.1 Faculty of Electrical Engineering

In order to confirm our observations from Fig.1 the data collected from Faculty of Electrical Engineering, Faculty of Business Services and Mărăşti Students Campus was analysed using the "Proposed Methodology". In the first iteration the data from Faculty of Electrical Engineering was first analysed using IQR and MAD algorithms on each year to extract yearly outliers for each year individually and then the same process was executed on all the data set. There are multiple ways a data can be analysed, because we wanted to highlight the most evident outliers, we did the intersection of the two processes, the result are represented by the IQR/comb and MAD/comb in Table1. and Table 2. We observed that the IQR method detected that 1.4% of all data is represented by outliers and for MAD 3.7%.

IQR	2014	2015	2016	2017	2018	2019	Total
IQR/Year	19	121	95	38	25	25	323
IQR/Total	19	70	174	226	115	118	722
IQR/Comb	19	70	95	34	25	25	268

Table 1. The absolute number of outliers detected using IQR at Faculty of Electrical Engineering Location

MAD	2014	2015	2016	2017	2018	2019	Total
MAD/Year	9	589	1102	109	356	149	2314
MAD/Total	12	214	467	595	356	272	1916
MAD/Comb	9	214	467	99	356	149	1294

Table 2. The absolute number of outliers detected using MAD at Faculty of Electrical Engineering Location

Having a closer look at the density of data we observed that using the LOF on each year we obtain a total of 5% of the data as outliers for k=2 and 1.5% for k=3 and 0.8%1 for k=2 in Table 3.

LOF	2014	2015	2016	2017	2018	2019	Total
k=2	485	450	418	426	450	349	2578
k=3	154	129	137	124	125	100	769
k=25	75	50	58	36	85	111	415

Table 3. The number of outliers detected using LOF at Faculty of Electrical Engineering Location

The process was also conducted on all the data and we obtain the same outlier percentage for k=2 and k=3 still, less than 1% for k=25.

DBSCAN	2014	2015	2016	2017	2018	2019	Total
DB/Year	78	202	170	181	128	216	975
DB/Total	52	23	20	29	9	16	149
DB/Comb	52	23	19	29	9	16	149

Table 4. The number of outliers detected using DBSCAN at Faculty of Electrical Engineering Location

For the DBSCAN process we used the intersection between two different density-based processes; one is based on the energy consumption value and hour and the second one is based on the energy consumption value and the day of the week. The testing was also conducted on yearly data and respectively on all the data and because between the two processes the yearly data contains all the data detected in total the intersection DB/Comb will be equal with IQR/Total. It can be observed in Table 4. that in total DBSCAN method detected 0.87% of the data to be an outlier.

4.2 Faculty of Building Services

For the collected data from Faculty of Building Services we managed to run the same test using IQR and MAD process. We observed that from all the data IQR detected 2%

IQR	2014	2015	2016	2017	2018	2019	Total
IQR/Year	227	43	481	70	78	165	1064
IQR/Total	180	109	6	207	263	68	833
IQR/Comb	180	43	6	70	78	68	445

Table 5. The number of outliers detected using IQR at Faculty of Building Services Location

MAD	2014	2015	2016	2017	2018	2019	Total
MAD/Year	180	16	1321	0	20	5	1542
MAD/Total	227	165	9	273	324	119	1117
MAD/Comb	180	16	9	0	20	5	230

Table 6. The number of outliers detected using MAD at Faculty of Building Services Location

of the data as outliers and 3% for MAD in Table 6. For the LOF process we have 2.2% of the data outlier for k=2 and for k=3 and k=25 we have 0.8% respectively 0.09% Table 7.

LOF	2014	2015	2016	2017	2018	2019	Total
k=2	174	168	136	234	258	186	1156
k=3	55	55	42	100	114	66	432
k=25	6	15	2	9	10	9	51

Table 7. The number of outliers detected using LOF at Faculty of Building Services Location

For all the Process were all the data was tested we obtain 6.8% of the data as outlier based on k=2 and 0.8% for the k=3 the algorithm found only one outlier.

DBSCAN	2014	2015	2016	2017	2018	2019	Total
DB/Year	33	37	22	33	22	28	175
DB/Total	7	3	3	10	11	1	35
DB/Comb	7	3	3	10	11	1	35

Table 8. The number of outliers detected using DBSCAN at Faculty of Building Services Location

Compared with other methods for the DBSCAN we obtain only 0.06% of the data as outlier. Even if the the total number of anomalous data is less the method indicated a silhouette score equal with 0.99 for both DBSCAN 1 and DBSCAN 2.

4.3 Mărăști Students Campus

For Mărăști Students Campus data we applied the same methodology and we observed abnormal behaviour in the IQR and MAD algorithms. For both of the methods and approaches we obtained for most of the tested year zero outliers detected even if our visual data interpretation suggested the opposite, Table 9, Table 10.

IQR	2014	2015	2016	2017	2018	2019	Total
IQR/Year	0	0	2	146	0	0	148
IQR/Total	97	0	4	0	0	0	101
IQR/Comb	0	0	2	0	0	0	2

Table 9. The number of outliers detected using IQR with threshold = 1.5 at Mărăști Students Campus Location

MAD	2014	2015	2016	2017	2018	2019	Total
MAD/Year	0	0	0	120	0	0	120
MAD/Total	21	0	0	0	0	0	21
MAD/Comb	0	0	0	0	0	0	0

Table 10. The number of outliers detected using MAD at Mărăști Students Campus Location

Continuing our analysis on the LOF process we observed that we have 5% detected outliers for k=2, 1.5% for k=3 and 0.8% for k=25.After running the same exercise for all the data we obtain the same proportions yet for k=25 the proportion was slightly close to 0% having only 0.02% of outliers detected Table 11.

LOF	2014	2015	2016	2017	2018	2019	Total
k=2	485	450	418	426	450	349	2578
k=3	154	129	137	124	125	100	769
k=25	75	50	58	36	85	111	415

Table 11. The number of outliers detected using LOF at Mărăști Students Campus Location

DBSCAN1	2014	2015	2016	2017	2018	2019	Total
DB1/Year	4884	4465	4512	3834	4267	4120	26082
DB1/Total	598	231	128	82	47	29	1115
DB1/Comb	-	-	-	-	-	-	-

Table 12. The number of outliers detected using DBSCAN (value+hour) at Mărăști Students Campus Location

For the DBSCAN we saw that for the energy consumption value and hour method (DBSCAN1) we obtain an average of 52% of the data as outlier and no intersection between the yearly data a total data Table 12. In the second part of the exercise the opposite was observed and for the method (DBSCAN2) where we take into consideration energy consumption value and the day of the week and we obtain an average of 2.2% of the data as an outlier Table13. It is to mention that for this exercise there was also no outlier intersection between the yearly data and total data approach, more than that the same could be noticed between DBSCAN1 and DBSCAN2.

DBSCAN2	2014	2015	2016	2017	2018	2019	Total
DB2/Year	78	202	170	181	128	216	975
DB2/Total	58	23	20	29	9	16	149
DB2/Comb	-	-	-	-	-	-	-

Table 13. The number of outliers detected using DBSCAN at Mărăști Students Campus Location

Because we wanted to understand the huge consistency gap that occurred in the IQR, MAD and DBSCAN processes for this data set, we calculated the standard deviation of all the data sets and compared them. It has been observed that for: Faculty of Electrical Engineering we have a STD of 22.74, for Faculty of Building Services 6.44, for Swimming Complex of Technical University of Cluj-Napoca 10.35 and for Mărăști Students Campus 38.98. The value of the standard deviation from Mărăști Students Campus data set indicates a random behaviour in energy consumption, it is almost double compared with Faculty of Electrical Engineering data and way larger compared with other locations.

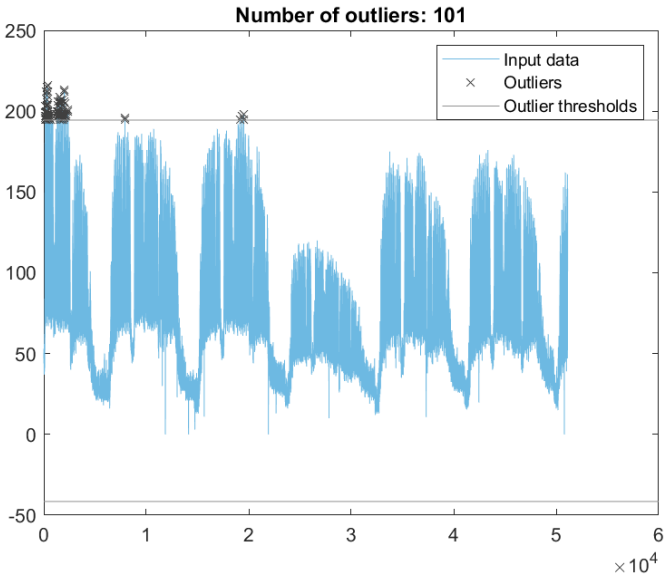

Fig. 2. The number of outliers detected using IQR with threshold 1.5 at Mărăști Students Campus Location

It is clear that, having increased volatility in the dataset can determine the used methods to generate large thresholds that are not capturing the extremes in case of IQR and MAD Fig.2. and random patterns that cannot be classified by DBSCAN method. In order to detect anomalous data with IQR method we restricted the threshold value to 1. After rerunning the process, the results showcased that for all the data we managed to obtain 2.1% of the data as anomalous data Table 14 but still we were unable to capture the outliers which are bottoming the "majority" of the data as presented

in Fig.3.It is also to mention the fact that stretching threshold will not guarantee that the method is capturing more damaged data. From the same chart we can understand that because of the narrowing part of the valid data is now to be considered as an outlier which highlights the inefficiency of an IQR method in the context of high volatility data scenario.

IQR	2014	2015	2016	2017	2018	2019	Total
IQR/Year	75	0	193	310	115	85	778
IQR/Total	544	270	202	6	41	13	1076
IQR/Comb	75	0	193	6	41	13	328

Table 14. The number of outliers detected using IQR with threshold = 1 at Mărăști Students Campus Location

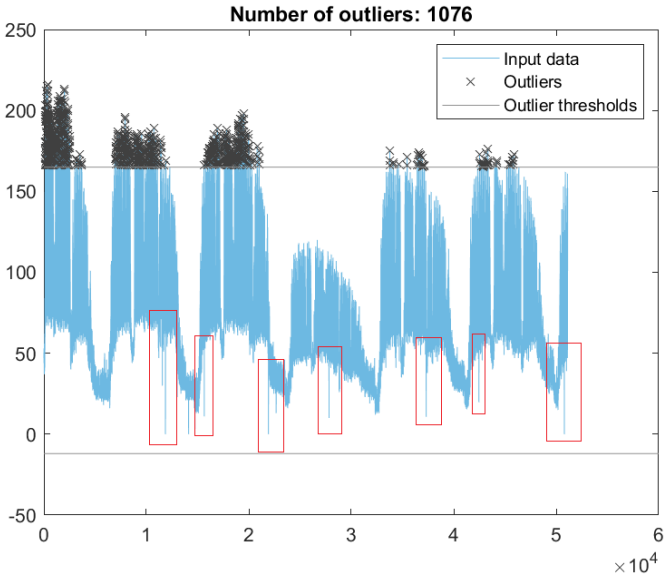

Fig. 3. The number of outliers detected using IQR with threshold 1 at Mărăști Students Campus Location

Having these results, determined us to find alternative threshold outlier detection techniques for this specific use case. Moving median method (MM) was proposed to be executed on this specific dataset with a threshold equal with 3 and the results suggested that for all the data we managed to identify 0.9% of the data as damaged data Table 15.

MM	2014	2015	2016	2017	2018	2019	Total
MM/Year	86	82	57	89	67	76	457
MM/Total	86	81	56	90	67	75	455
MM/Comb	86	81	56	89	67	75	454

Table 15. The number of outliers detected using MM at Mărăști Students Campus Location

5 Intelligent Scoring Method Results

5.1 Faculty of Electrical Engineering

With the aim of understanding the validity of the tested methods we compiled our intelligent scoring method over the outlier outputs. Because the IQR and MAD are both "spread control" based methods we will use as input for our clustering method the intersection of both.

Method	Total Outliers	Valid (%)	Invalid (%)
IQR/MAD	249	46.18	53.81
LOF K=2	2608	22.17	77.82
LOF K=3	1412	21.33	78.66
LOF K=25	18	0	100
DBSCAN	149	17.67	82.32

Table 16. The number of valid/invalid outliers detected using our intelligent scoring method for Faculty of Electrical Engineering location

After the first iteration of the intelligent scoring method we observed that none of the tested outlier detection techniques have detected more than half of the data as valid data Table 16 for Faculty of Electrical Engineering data set.

5.2 Faculty of Building Services

Method	Total Outliers	Valid (%)	Invalid (%)
IQR/MAD	227	50.2	49.7
LOF K=2	3513	20.67	79.3
LOF K=3	424	17.49	82.5
LOF K=25	1	0	100
DBSCAN	35	28.57	71.42

Table 17. The number of valid/invalid outliers detected using our intelligent scoring method for Faculty of Building Services location

The same was noticed for the Faculty of Building Services where only the IQR/MAD combination managed to detect more than 50% of the data as valid outlier Table 17. In case of the analysis for Mărăști Students Campus most of the methods failed having on average around 80% of the outliers as invalid outliers Table 18. More than that because of the increased volatility and random behaviour in energy consumption the DBSCAN was unable to run properly.

5.3 Mărăști Students Campus

Method	Total Outliers	Valid (%)	Invalid (%)
IQR/MM	2885	20.24	79.75
LOF K=2	2558	17.87	82.12
LOF K=3	743	20.62	79.38
LOF K=25	9	0	100
DBSCAN	-	-	-

Table 18. The number of valid/invalid outliers detected using our intelligent scoring method for Faculty of Mărăști Students Campus location

6 Conclusion

An empirical outlier detection testing was conducted over energy consumption data collected during the DR-BoB "Demand Response in Blocks of Buildings" project funded by the EU Horizon 2020 innovation program under grant agreement No. 696114/2016 in order to prepare it for future forecast exercise. As concluded in our previous article [16], the outlier detection techniques were unable to differentiate between natural energy peaks and abnormal data for most of the cases without our proposed intelligent scoring method. For future work the outliers will be adjusted and forecasting methods will used on the current data sets. LOF K=25 will be removed from future analysis due to its inability to capture damaged data.

7 Acknoledgement

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

8 References

[1] 'Global Energy Statistical Yearbook 2019', https://yearbook.enerdata.net/totalenergy/worldconsumption-statistics.html, accessed 12 June 2020

[2] D. I. Jurj, D. D. Micu, Muresan A.: 'Overview of Electrical Energy Forecasting Methods and Models in Renewable Energy'. 2018 International Conference and Exposition on Electrical And Power Engineering (EPE), Iasi, 2018, pp. 0087-0090

[3] 'U.S. Energy Information Administration (EIA) '. https://www.eia.gov/energyexplained/use-of-energy, accessed 1 September 2020

[4] E.Q. McCallum.: 'Bad Data Handbook: Mapping the World of Data Problems', (O'Reilly Media, Inc. U.S. 2012)

[5] D. M. Hawkins.: 'Identification of Outliers',(Chapman and Hall, England, United Kingdom, 1980)

[6] Steinbach, and V. Kumar P.Tan, M.: 'Introduction to Data Mining', (Pearson Addison Wesley, Boston, MA, USA, 2006), pp. 651-683

[7] Choy K.: 'Outlier detection for stationary time series', Journal of Statistical Planning and Inference, 2001, pp 111-127

[8] I. Chang, G. C. Tiao, Chen. C.: ' Estimation of time series parameters in the presence of outliers', Journal of Technometrics, 1988, pp 193-204

[9] Salgado C.M., Azevedo C., Proença H., Vieira S.M.: ' Noise Versus Outliers'. Secondary Analysis of Electronic Health Records, Springer, 2016

[10] Kaya A.: ' Statistical modelling for outlier factors', Ozean Journal of Applied Sciences, 2010, pp 185-194

[11] C. Chen, L.-M. Liu.: 'Joint estimation of model parameters and outlier', Journal of American Statistical Association, 1993, pp 284-297

[12] A. Gran_e, Veiga. H.:'Wavelet-based detection of outliers in financial timeseries', Journal of Computational Statistics and Data Analysis,2010. pp. 2580-2593

[13] Akouemo Kengmo Kenfack, Hermine N.: 'Data Cleaning in the Energy Domain'. Dissertation thesis, Marquette University, 2015

[14] B. Bârgăuan, M. Crețu, O. Fati, A. Ceclan, L. Dărăbant, D.D. Micu, D. Șteț, Czumbil L.: 'Energy Management System for the Demand Response in TUCN Buildings'. 53rd International Universities Power Engineering Conference, September 2018

[15] B. Bârgăuan, O. Fati, A. Ceclan, D.D. Micu, D. Șteț, L. Czumbil, Mureșan P.: ' Demand Response on Blocks of Buildings – Romanian Pilot Site Innovation Project'. 7th International Conference on Modern Power Systems (MPS), June 2017

[16] Dacian I. Jurj, Dan D.Micu, L.Czumbil, Alexandru G. Berciu, Denisa M. Bărar, Mircea L.:'Analysis of Data Cleaning Techniques for Electrical Energy Consumption of a Public Building'.2020 55th International Universities Power Engineering Conference (UPEC), Torino, Italy, 2020, pp. 1-6

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

[18] 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

[19] Chao Chen, Diane C.:' Energy Outlier Detection in Smart Environments'. School of Electrical Engineering and Computer Washington State University, 2011

[20] Bouguessa M.: 'A probabilistic combination approach to improve outlier detection'. IEEE 24th International Conference on Tools with Artificial Intelligence, 2012, pp. 666–673

[21] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J.: 'LOF: identifying density-based local outliers'. In ACM sigmod record, May 2000, pp. 93–104

[22] 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

[23] 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