www.ijaceeonline.com ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

Data Preprocessing in Electrical Energy Consumption Profile Clustering Studies

Mustafa Şen Yıldız
Department of Electrical and Electronics Engineering, Kırklareli University, Kırklareli, Turkey
Email: mustafasenyildiz@klu.edu.tr

Kadir Doğanşahin
Department of Electrical and Electronics Engineering, Artvin Çoruh University, Artvin, Turkey
Email: dogansahin@artvin.edu.tr

Bedri Kekezoğlu
Department of Electrical Engineering, Yıldız Technical University, İstanbul, Turkey
Clean Energy Technologies Institute, Yıldız Technical University, İstanbul, Turkey
Email: bkekez@yildiz.edu.tr

Abstract: *By the developing technology, the traditional power system has been modernized by advanced metering and communication infrastructures. Thus, a structure with high monitorability and interaction with users has been developed. In modern power systems, applications such as short-term demand forecasting, demand management or dynamic tariffs are actively used. The success of all these applications is closely related to mastering the opportunities provided by the modern power system and using these opportunities effectively. Electrical energy consumption profile clustering is a widely applied and highly functional approach that can be evaluated in this context. It stands out as a practical method that allows obtaining representative load profiles specific to each cluster by determining similar consumption behaviors from big consumption data. The success of cluster analysis is directly related to the quality of the data. In order to obtain consistent results, the analyzed data must have been successfully pre-processed and removed from missing and outliers. In this study, data processing steps before clustering analyses have been examined. A comprehensive review of the electrical energy consumption clustering literature has been performed. Explanatory tables have been created for the stages applied, the approaches adopted and the methods used in the data preprocessing.*

Keyword: *Data Preprocessing; Electrical Load Data; Preprocessing Techniques for Clustering*

1. INTRODUCTION

Electric is a type of energy that is easy to transmit and distribute. It is environmentally friendly, which does not cause waste or gas emissions during use. Thanks to its current infrastructure and operating mechanisms, its accessibility, reliability, and sustainability are quite high. These superior features make electrical energy highly preferable in final energy consumption. That is why, in many sectors today, there are transitions from the use of different energy sources to the use of electrical energy. The increase in electrical energy demand, which is currently shaped by the effects of developing countries and increasing welfare, is expected to accelerate with these transitions.

In order to meet the increasing demand, it is aimed to research new sources and to determine various strategies for increasing efficiency in existing systems. The rise of environmental and economic concerns has an impact on this quest. Within the last 10 years, the use of renewable energy sources such as wind and solar has been encouraged on the generation side, and the number of that facilities has increased rapidly. On the consumption side, necessary opportunities and incentives have been provided for practices aimed at increasing efficiency and meeting their own demand.

Considering the number and variety of consumers, the size of the service area, and the diversity of the components is noteworthy that it is a very large and complex structure. In such a system, the supply-demand balance must be maintained continuously and the system variables must be kept sensitively within certain limits.

The increasing penetration of intermittent renewable energy sources such as wind and solar, the diversification of the consumers and differentiation consumption behaviors make the operation of the

Cite this paper:

Mustafa Ş. Yıldız, Kadir Doğanşahin, Bedri Kekezoğlu, "Data Preprocessing in Electrical Energy Consumption Profile Clustering Studies", International Journal of Advances in Computer and Electronics Engineering, Vol. 8, No. 4, pp. 1-13, April 2023.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

modern power system more and more complicated. On the other hand, thanks to developments in measurement and communication systems, it has become possible to collect measurement data from more points and with higher resolution over the power system than in the past. Power system operators benefit from the data collected from the system in the planning and demand forecasting studies. Especially in the operation of distribution systems, these data and the findings from the analysis are of great importance.

Consumers in distribution systems are classified under three main categories, residential, commercial, and industrial. Individual measurements for each consumer are not feasible. It causes unnecessary data volume and processing effort. In practice, it is preferred to collect data from key points in the power system. However, consumers do not exhibit such a clearly distinguishable distribution within the power system. It is possible to have consumers from different consumer classes together under the same measurement point. Moreover, even consumers from the same consumer class are likely to exhibit different consumption behaviors. This makes each point special and causes simple classifications such as residential, commercial, or industrial to be inadequate. Therefore, it is a more functional approach to cluster the consumption data from the measurement points in proportion to their similarities and to create a single load profile that represents the consumption profiles within each cluster. All of the analyzes performed for this purpose are known as cluster analysis. It is a well-accepted and highly preferred instrumental in the studies of adjusting multi-time or dynamic tariffs [1]–[5], demand management [6]–[10], and demand forecasts [11]–[14].

It is not possible to use raw data directly in data analysis. Analyzing data with quality problems leads to misleading results. For this reason, preprocessing is very important in data analysis.

In clustering analysis, time-stamped electrical measurements taken from the power system constitute the main data set. Time series data may have more data quality problems than other types of data. Faults in the measuring system may cause loss and outlier values. Transients and blackouts within the power system may need to be extracted from the data set. Different datasets with different resolutions, sizes, and formats may need to be combined. Adopting the appropriate approach and using the right method in preprocessing is significant for the success of the analyses. [15].

In this study, four subsections specific to each component of preprocessing have been determined under a main heading for general definitions of data preprocessing. The methods used in the examined studies are appropriately distributed under these headings. Explanations of the methods commonly used in the literature have been given broader.

In this study, four subheadings specific to each component of data preprocessing have been designed. These components are data integration, cleaning, reduction, and scaling processes. The data preprocessing methods used in the literature have been classified under these four headings based on their functions. Explanations of the methods commonly used in the literature have been given broader. Explanatory tables have been created that summarize the processes applied in the data preprocessing in the reviewed studies and the methods used in these processes. In addition, a table in which the characteristics of the raw data sets subjected to data preprocessing and the data sets obtained afterward have been given together. The main motivation of this study is to provide a resource that presents an overview of data preprocessing and methods used in this process for the researchers studying on electrical energy consumption data. In the papers from relevant literature, details on the raw data and the methods used in the analysis are rarely shared. Most of the studies have been realized with the processed, namely ready-to-use data. Generally, clustering algorithms have been centered in the studies. In this respect, it is expected that the study will make an important contribution to the literature.

In the second section of the study, necessary definitions on data preprocessing have been given. In the third section, there are details of the literature review performed within the scope of the study and examinations on the methods used in data preprocessing. The last part of the study is the conclusion part, where summative information and important findings have been shared.

2. DATA PREPROCESSING STEPS

By the developments on the measurement and communication infrastructures in power systems, it has become possible to collect data from more points and with higher resolutions compare to the past. Increasing data volume, on the one hand, increases the quality of the information possessed, on the other hand, it has made the processing of data more complicated [16]. With the increase in data volume, the size and variety of data quality problems has also increased. The success of data analysis is closely related to data quality. In order to obtain consistent results, missing or outlier data must be determined and removed from the data sets, and the data should be formatted in accordance with the study. All processes applied for this purpose are called data preprocessing. The data preprocessing is examined under four main headings.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

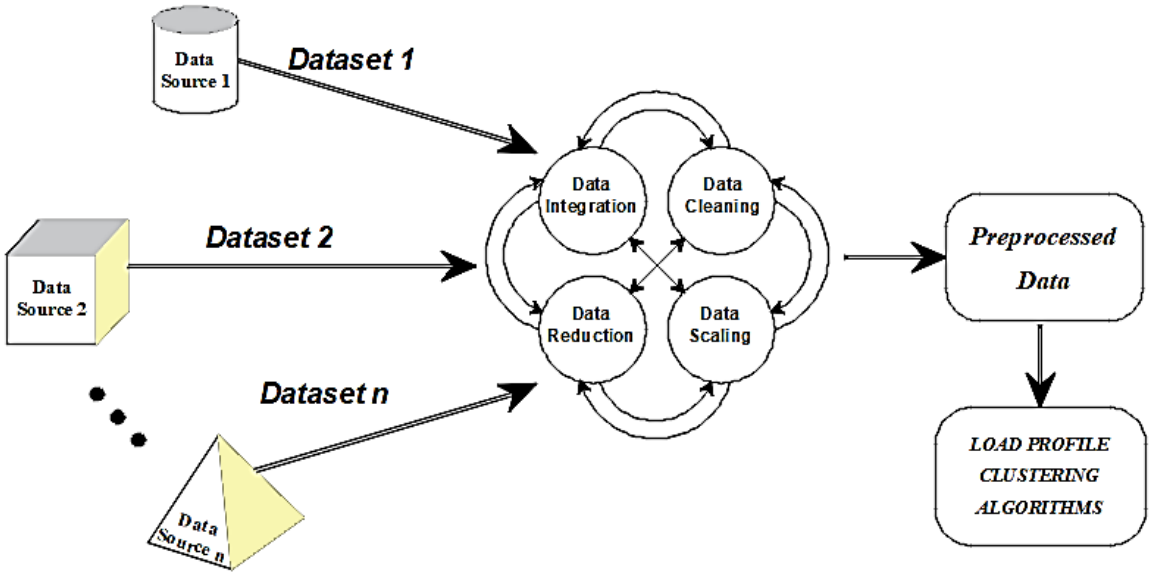


Figure 1 Data preprocessing steps and the place on load profile clustering work flow

A visualization of the functioning of these sub-processes specific to electrical energy consumption clustering analysis is presented in Figure -1.

2.1 Data integration

In general, electrical energy consumption clustering studies are based on consumption data only. However, in some studies, various data affecting electricity consumption can also be included in the analysis. In such multivariate studies, different data sets should be combined and analyzes should be performed on a single data set.

Different data storage mechanisms are used in data storage in order to use data space efficiently and to increase data speed. For this reason, data sets in different databases could be in different data architectures and encrypted. In addition, different formats can be preferred while storing data. In order to create a single usable data set from various databases created independently from each other, it is necessary to eliminate the differences between the data sets and bring them together under a standard structure. This process is called data integration [17], [18].

Apart from the structural differences, the content of the data sets to be combined may also differ. Electricity consumption data are time series type data. The difference in the time intervals and resolutions of such data sets may cause dimensional incompatibility. On the other hand, not all of the features in a dataset content may be needed. The new data set to be created by selecting only the required features may increase the speed and simplicity of the analysis [19].

In order to prevent such problems, the data sets can be treated with the help of other pre-processing steps to made ready for integration. It should be noted that there is no hierarchical order between the data preprocessing steps. In case of need, any of preprocessing step can be used over and over again.

2.2 Data Cleaning

In data analyses, it is not feasible to use raw data directly. Quality problems in the raw data may cause problems in the implementation of the analyses or in obtaining consistent results after the analysis. Some of the reasons leading to quality problems in electricity consumption data are listed below [16],

- Malfunctions that may occur in the measurement and data transfer processes of the data gathering system may produce erroneous data.
- As a result of situations such as transients or power outages that may occur in the system, any data may not be received through the measurement system or outlier values may be measured.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

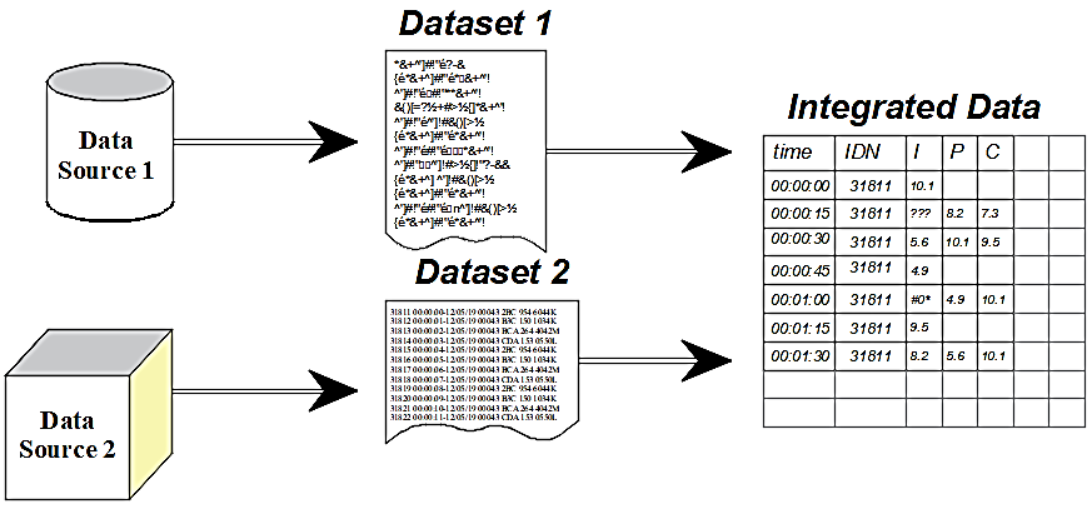


Figure 2 A representative illustration for data integration process

- Measuring devices in the system may have different accuracy or sampling rates. Measurement errors are usually caused by the limited capacities of the devices in various situations. In addition, the degree of resistance of the devices to external effects may vary. Low endurance devices can cause noisy data.

- The data gathering process is open to the intervention of the staff who follow up the process. During these interventions, data may be deleted or outliers may occur.

Data quality problems can be divided into three categories: outlier data, missing data and noisy data.

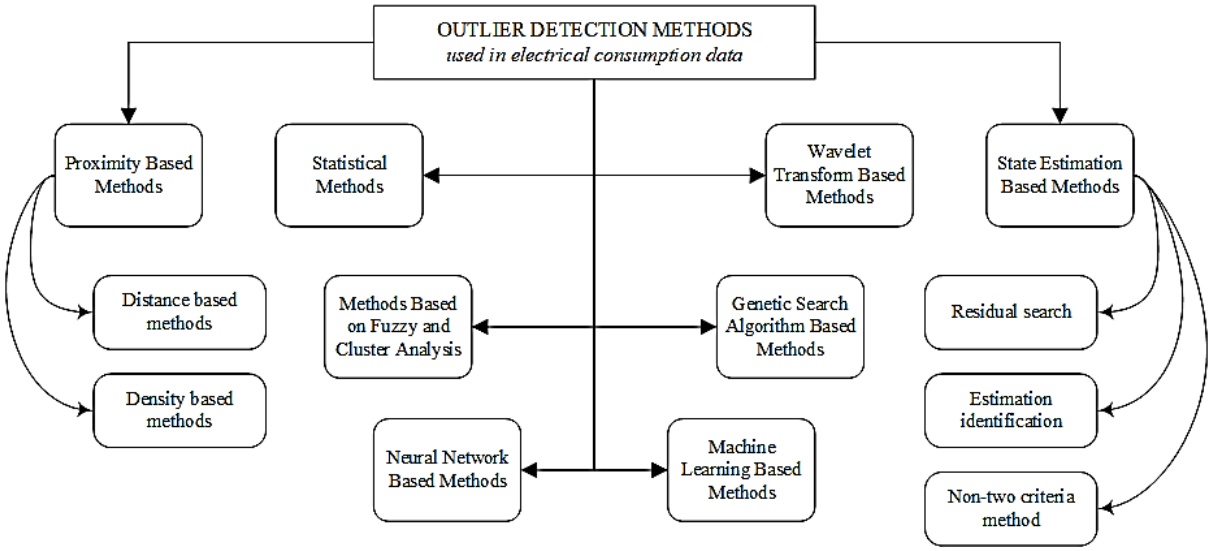


Figure 3 Classification of outlier data detection methods

2.2.1 Outlier Data

In its most general definition, it is the values that are far from the general data distribution and are statistically inconsistent with other data [20]. Power system transients or malfunctioning in measurement and communication infrastructure may cause outliers. For example, a value of 300kWh in the hourly energy consumption data of a facility with an installed power of 100kW is an outlier.

2.2.2 Noisy Data

It is low-quality data that is not possible to be used with the help of any software or device to make sense of the information it contains [16].

2.2.3 Missing Data

Missing data are empty or meaningless sections in the data set as the result of problems in the phase of measurement, transfer, or storage processes. [21].

The first step of data cleaning preprocessing is bad data (outlier, noisy data, or missing data) detection. Noisy and missing data can be detected simply but outlier detection is complicated. Unlike noisy and

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

incomplete data, outliers contain usable and complete information when considered externally. However, they exhibit anomalies when considered correlational with the other data in the dataset. There are many different methods for outlier detection [16], [22], [23]. The methods used in the outlier data detection are categorized as given in Figure 3.

Outlier detection methods differ in being subjective and objective. Some of the methods need user input to decide the variables taken as the basis in the process of detecting incorrect data. On the other hand, some methods are capable to determine the required variables and the steps, according to the data characteristics. Therefore, objective methods can provide an advantage in terms of being free from user error [24]. On the other hand, the adequacy of the methods against the volume of data is not the same. Some methods can be superior to others in terms of having less computational burden. However, the volume of data that can be processed with such methods may remain low compared to alternative methods, which are more complicated. Therefore, the choosing appropriate method according to the characteristics of the data enables to increase the success in detecting outliers and avoid unnecessary computational effort.

Consequently, detected outliers are excluded from the data set because they are values that do not serve the purpose. Therefore, data points defined as outliers are now missing data. Two approaches are adopted for handling missing data. The first option is to remove the missing data from the dataset with an appropriate deletion and so make the dataset free from data quality problems. However, in some cases, the volume of the bad data may be high in the dataset or the information of that data points may be important for the analysis. The new data set obtained may be insufficient to achieve the targeted results in data analysis [25]. In such cases, data imputation techniques are applied, which is the second approach. Data imputation is the process assigning values to missing data points with the values estimated from available data.

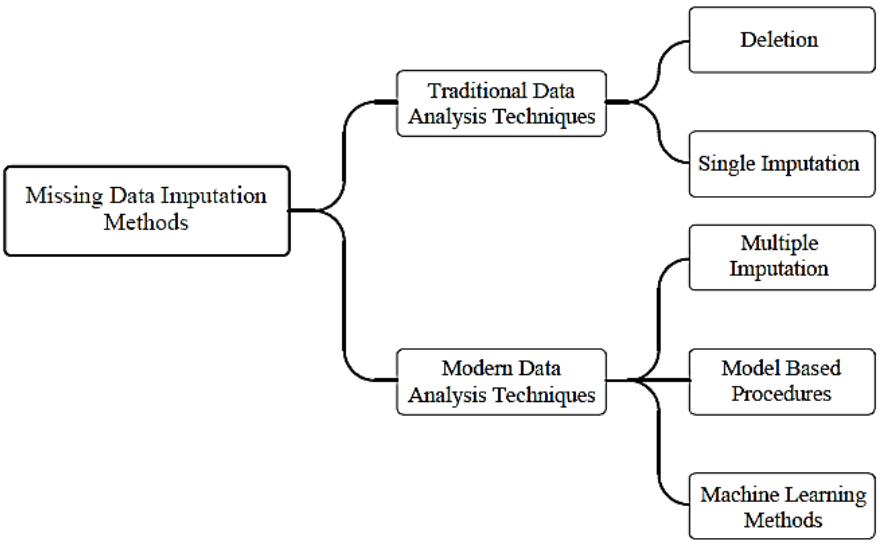


Figure 4 Classification of missing data imputation approaches

Data imputation techniques are classified as traditional and modern data analysis techniques, Figure 4. Analyzes involving single data imputation step are generally considered to be the traditional method. It stands out as a functional method for the data sets containing a small number of missing data. The techniques based on modern data analysis, on the other hand, are classified as multiple imputation, model-based procedures, and machine learning methods. In a data set with a high rate of missing data, imputing all missing values at once may not be a consistent approach. A cascaded imputation can be made instead. Imputations can be made in each layer by making use of the values assigned for the missing data in the previous layer [26].

2.3 Data Reduction

Datasets may have more features or instances than required. Working with an unnecessarily crowded data set increases the computational effort and time in the analysis. Instead, it is easier to perform the analysis with the new dataset formed by the selection of the necessary features and the instances from the raw data. The preprocessing performed for this purpose is called data reduction [27]. The methods used in the data reduction can be categorized as feature/instance selection, data discretization, and feature/instance extraction.

2.3.1 Feature Selection

Some of the features in a dataset may contain data

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

that is unnecessary for the intended work. Feature selection is the process to create a subset from the features in the raw data by examining their relevance, and validity in terms of the goal of the study [28]. For example, electrical energy consumption data and weather data are generally considered together in forecasting studies. Weather data are multidimensional time series. In this data set, feature selection methods are used to determine the ones related to electricity consumption among the features such as temperature, humidity, pressure, etc. The selected feature is directly included in the data to be analyzed without any transformation, as it is in the raw data [29]. Thus, instead of working with the complete set of weather data, a subset that includes only the features related to the analyses Feature selection methods can be categorized into four classes, filter, wrapped, embedded, and hybrid methods [30].

2.3.2 Instance Selection

Each row in the dataset is called an instance and each column is called a feature. A consumption dataset may contain a larger time period than is needed. Instance selection is performed in the process of separating data in a certain time interval or data points defined through a rule from raw data [31]. As in feature selection, the selected data is included in the data set without any transformation. An explanatory illustration of feature and instance selection is shared in Figure - 4.

2.3.3 Feature Extraction

Feature extraction methods work for deriving a more informative and functional new feature by applying required operations to multiple features selected from the raw data [30]. In consumption profile clustering, distinctive information for consumer behaviors (peak loading, daily average loading, etc.) is

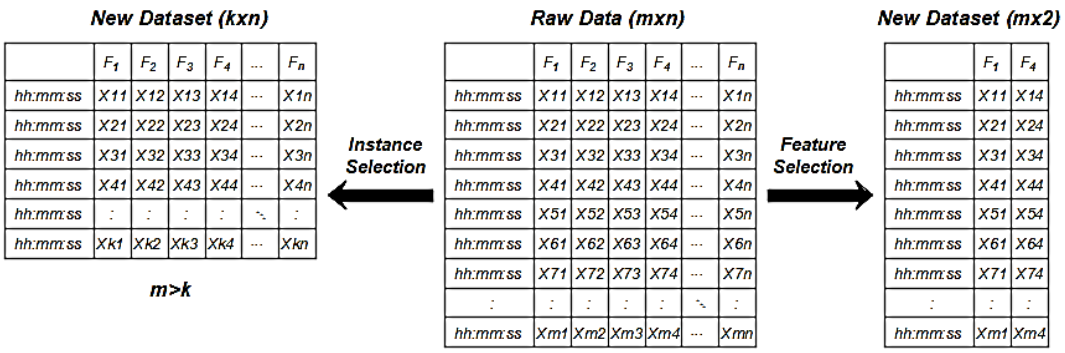


Figure 5 A representative illustration for Instance and Feature selection processes

useful for the determination of typical load patterns. These kinds of features can be extracted from the combinations of the original features in the raw data. The features selected from the raw data are integrated into the new dataset after processing with Feature extraction methods. Unlike feature selection, data is transformed here.

2.3.4 Instance Generation

Instance generation methods are the process of creating more functional and relevant new instances with transformations based on combinations of instances from the raw data. With the developing measuring technologies, it has become possible to take high resolution measurements from power systems. In electrical energy consumption cluster analysis, hourly data is generally used. Instance generation process is applied during the conversion of a data set with a resolution of 15 minutes to hourly data.

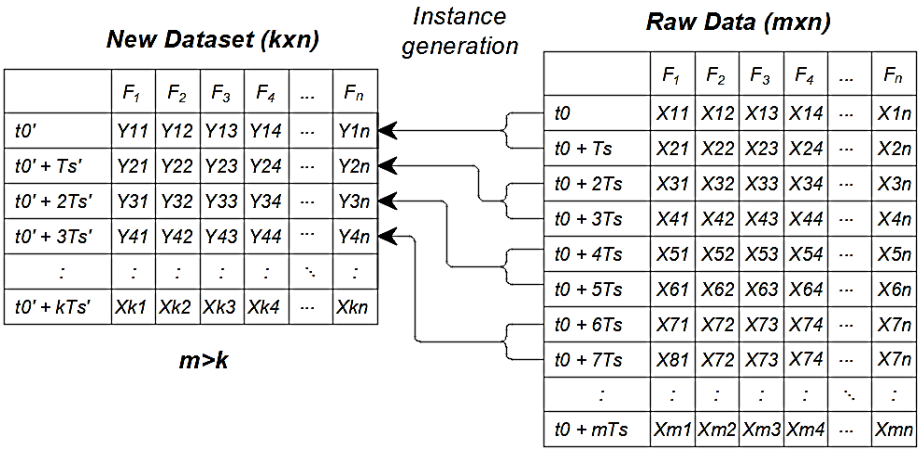
2.3.5 Discretization

Discretization is the process to identify and discretize data that are close to each other in various aspects.

2.4 Data Scaling

Consumption data of the different users may take values over a wide range from kW to MW. Clustering studies are based on consumption behaviors rather than consumption amount. In a clustering analysis to be performed on a data set with different consumers, the behavior of users with low consumption may not have an effect on the results. In order to achieve a standard in the analyzes, the consumption values of each consumer are normalized to be within a certain range. this process is accomplished by various data scaling methods use maximum, minimum or statistical measures of the relevant data. Thus, consumption profiles that vary at different intervals are reduced to same interval.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13



	F_1	F_2	F_3	F_4	...	F_n
t0'	Y11	Y12	Y13	Y14	···	Y1n
t0' + Ts'	Y21	Y22	Y23	Y24	···	Y2n
t0' + 2Ts'	Y31	Y32	Y33	Y34	···	Y3n
t0' + 3Ts'	Y41	Y42	Y43	Y44	···	Y4n
⋮	⋮	⋮	⋮	⋮	⋱	⋮
t0' + kTs'	Xk1	Xk2	Xk3	Xk4	···	Xkn

	F_1	F_2	F_3	F_4	...	F_n
t0	X11	X12	X13	X14	···	X1n
t0 + Ts	X21	X22	X23	X24	···	X2n
t0 + 2Ts	X31	X32	X33	X34	···	X3n
t0 + 3Ts	X41	X42	X43	X44	···	X4n
t0 + 4Ts	X51	X52	X53	X54	···	X5n
t0 + 5Ts	X61	X62	X63	X64	···	X6n
t0 + 6Ts	X71	X72	X73	X74	···	X7n
t0 + 7Ts	X81	X72	X73	X74	···	X7n
⋮	⋮	⋮	⋮	⋮	⋱	⋮
t0 + mTs	Xm1	Xm2	Xm3	Xm4	···	Xmn

Figure 6 A representative illustration for Instance generation process

The first is the scaling of the data, the other is the standardization of the data distribution. Various definitions are used for scaling purposes. Some of the most frequently used ones are given with the following equations.

In Equation 1, the min-max normalization is given. In this normalization, the minimum value of the feature is subtracted from all values in the relevant features and each obtained value is divided by the difference between the maximum and minimum values of that feature. The equation is given in Equation 1.

$$x' = \frac{x - min(x)}{\max(x) - \min(x)} \quad (1)$$

As another scaling method, mean value normalization is given in Equation 2. In this normalization, the average value of the feature is subtracted from all the values of the relevant feature, and each value obtained is divided by the difference between the maximum and minimum values as in the min-max normalization.

$$x' = \frac{x - avarage(x)}{\max(x) - \min(x)} \quad (2)$$

For the data distribution standardization, Z-score is the most commonly used one. The distribution around the average of the values in the data is calculated with the Z-score, Equation 3.

$$x' = \frac{x - \mu}{\sigma} \quad (3)$$

where μ is the mean and σ is the standard deviation of the values in the feature.

3. DATA PREPROCESSING IN ELECTRICAL CONSUMPTION CLUSTERING STUDIES

In the introduction, deregulation in modern power systems and the increasing uncertainties that come with this transition have been mentioned. The challenges in the operation of modern power systems have been addressed. Fortunately, developing system infrastructure has made power systems more monitorable. Different approaches and methods have been developed to overcome these difficulties by making use of the measurements obtained through the system. Consumption cluster analysis is one of these methods and is frequently used in many studies such as demand forecasting, demand management, and tariff planning. there are many studies on this method in the literature. In this section, among the studies from the relevant literature, the ones that have details about the data preprocessing process have been examined. Although the number of studies containing consumption clustering analysis is large, not much detail is shared about the data preprocessing applied. Considering this point, the literature gap filled by this study may be understood clearly.

Table I presents the raw data used in the studies and the data obtained after the data preprocessing, comparatively. The data preprocessing steps applied in the studies examined and the information on the approaches adopted in this context are given in Table II. In Table III, the methods used in the preprocessing stage and the classification of these methods according to the type of preprocessing are shared in Table III.

As can be seen from Table I, the raw data used in the studies differ in temporal resolutions, which are from 1 minute to 1 day. However, the representative load patterns (RLP) obtained after load profile clustering studies are generally daily profiles that have hourly resolutions. In some of the studies, there are differences

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

in temporal resolution between raw data and the data set used in clustering analyses. A lower resolution data set was created using various methods over the raw data with a higher temporal resolution than necessary. In all of the studies, the active component of the electrical energy consumption has been taken as consumption data. On the other hand, there are some studies that have been realized by taking into account additional data such as weather, climate, and the capacity of system components. Another point that should be considered is the difference between the number of loads in the raw data and the data set considered in cluster analysis. It is because the missing or bad data in the raw data cannot be used or the non-compatibility on some of the loads in the raw data with the criteria determined as a basis in the relevant study. In Table II, implemented data preprocessing steps and adopted approaches in the relevant preprocessing in the reviewed studies have been represented. In the preparation of the table, only the text of the referenced study has been taken as the origin. In load profile clustering studies, quality problems in the data used should be eliminated as much as possible in order to obtain consistent RLPs. Therefore, in most of the studies, a preprocessing for detecting outlier data which is highly misleading for cluster analysis have been applied. As the table shows, statistical approaches have been used in outlier data detection in general of the reviewed studies. The reason for this is that statistical methods are easy to apply, capable to process large amounts of data, and have less operational effort. Even though there are serious criticisms on the accuracy of statistical methods in outlier detection, it is quite suitable for electrical energy consumption data analyses. One of the reasons is that there is no need for quite precise sensitivity in detecting outliers in electrical energy consumption data. Another is that the electrical energy consumption data has a similar probability distribution, making it convenient to apply statistical methods [55].

After the detection of missing and bad data, there are two different options, deletion, and data imputation. A highly consistent and reliable RLP can only be obtained with a data set highly representative and comprehensive. At the same time, working with redundant data increase computational effort and time unnecessarily. In the studies, deletion has been used to trim unnecessary instances and redundant features from raw data. On the other hand, various data imputation methods have been used to handle the missing data in the selected data set. In the considered studies, single imputation techniques have been the most common solution preferred for missing data treatment. There only two studies that have been used other techniques. If a data set includes different values such as environmental conditions and spatial information in addition to electrical energy consumption values, it may be more appropriate to use multiple imputation, model-based procedures, or machine learning methods.

Feature selection and instance generation have been used extensively in the data reduction preprocessing. In feature selection, filtering methods are preferred. For consumption data, filtering methods are less complex than that of alternatives and the success rate is adequate. Instance generation has been generally implemented to decrease the temporal resolution of raw data. Another important approach adopted in the studies is discretization, which is used with the purposes of grouping loads in terms of the similarity in various manners, such as consumption level or activation hours.

It is possible to exist the users with different consumption levels together in a distribution area. In the analyses performed to obtain consistent RLPs, consumption behavior is more important than the amount of the consumed energy. Data scaling is used to ensure that the behavior of each consumer has an equal effect on RLPs.

Min-max normalization has been generally applied in reviewed studies. In studies with high data distribution in terms of consumer behavior, Z-score data distribution standardization has been used.

In general, the clustering studies in the literature do not provide details about the preprocessing steps used for creating the data set used in the analysis. The papers referenced in this study are the ones that have related details. Even in these studies, the process has not been expressed in a descriptive manner. In Table III, the methods used in the referenced studies have been given with the knowledge of the preprocessing step where it is applied.

4. CONCLUSION

Data science is a very young and rapidly developing field. At the same time, it has a very wide usage area consisting of different disciplines. For these reasons, there is a lot of confusion in both the terminology and the definition and classification of the components. In this study, electricity consumption profile clustering analysis, which is used as the basis of many applications such as short-term demand forecasting, demand management and dynamic tariff planning, has been focused on.

In the first part of the study, data pre-processing steps have been discussed and their applications on electrical energy consumption data has been examined. Studies in the literature have been reviewed and the ones that contain details about the data preprocessing have been taken into account. The data preprocessing steps applied to the electrical energy consumption data and the details of the methods used in the implementation of these steps are presented in tables.

ISSN: 2456 - 3935

International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

TABLO 1 THE RAW DATA USED IN THE STUDIES AND THE DATA OBTAINED AFTER THE PREPROCESSING STEPS

	Source Data					**Load Profile Clustering**				
	No of Load	**Type of Load**	**Voltage Level**	**Temporal Resolution**	**Period**	**No of Load**	**Considered Period**	**Temporal Resolution**	**Period**	**Analyzed Attributes**
[4]	?	University Campus Buildings	LV	hourly	2 years	?	1 year	hourly	Daily	P
[32]	165	Mix	LV	15 min	6 months	165	6 months	15 m	Daily	P, VL,CD
[33]	245	HV-LV Substations	LV	hourly	4+ years	245	4+ years	hourly	Daily	P
[34]	234	Non-residential	MV	15 min	-	234	-	15 min	Daily	P
[35]	229	-	MV	15 min	6 months	208	6 months	15 min	Daily	P, CD,
[36]	-	Mix	MV	30 min	1 month	155	1 month	30 min	Daily	P
[37]	18098	Residential and Commercial	LV	hourly	396 days	1824	366 days	hourly	Daily	P, W, DL
[38]	1100	Residential	LV	10 min	8 months	952	8 months	10 min	Daily	P,W, SEC
[39]	218090	Residential	LV	1 hour	3.5 year	123150	1 year	hourly	Daily	P, Climate data
[40]	103	Residential	LV	1 min	1 years	103	Seasonal	hourly	Daily	P
[41]	1022	Mix	MV	15 min	1 year	1022	1 year	15 min	Daily	P
[42]	>197	Residential	LV	7-8 s	1 year	197	1 year	0.5 – 240 min	Daily	P
[43]	824	Substation	HV/LV	10 min	1 year	730	1 year	10 min	Daily	P, SI, CD
[44]	1072	Residential	LV	1 min	18 months	1072	18 months	10 min	Daily	P
[45]	4232	Residential	LV	30 min	18 months	3440	Work days in 5 years	hourly	Daily	P, SEC
[46]	1200	Residential	LV	1 day	1 month	938	1 month	1 day	Monthly	P
[47]	1	City		15 min	5 years	1	5 years	daily	Seasonal	P
[48]	100	HVAC units	LV	5 min	1 day	89	1 day	15 min	Daily	P
[49]	203	Feeders	HV	hourly	1 year	183	1 year	hourly	Daily	P
[50]	10	Research Institute Buildings	LV	various	3 years	10	1 year	hourly	Daily	P, W
[51]	114	Residential	LV	15 min	1 year	114	1 year	hourly	Daily	P,W
[52]	3000	Residential	LV	15 min	1 year	171	4 months	15 min	Daily	P
[53]	370	-	-	15 min	4 years	317	1 year	15 min	Daily	P
[54]	10 1000	Transformer Transformer	-	1 hour 1 hour	33 months 33 months	10 1000	33 months 33 months	hourly hourly	Daily Daily	P, W, C P

P	Active power consumption data	SI	System Information (Component's capacity, number of feeders, etc.)
W	Weather data	SEC	Socio-economic data (obtained via surveys)
C	Calendar data	VL	Voltage level
CD	Commercial Data (contracted power, activity code etc.)	DL	Day Length

ISSN: 2456 - 3935

International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

TABLO 2 DATA PREPROCESSING STEPS USED IN STUDIES

		Data Cleaning													Data Reduction							Data Scaling	
		Outlier Detection								Missing Data Treatment					FS & IS				FE & IG				
											Data Imputation												
	Data Integration	PBM	SM	Fuzzy M.	NN M.	Genetic SA	Mac. L. M	Wavelet M.	S. Est. M.	Deletion	SI	MI	MBP	MLM	Filter	Wrapper	Hybrid	Discretization	PM	Transform.	No of New .F.	Linear Normalization	Z-Score
[4]					✓					✓	✓				✓					✓			
[32]			✓								✓				✓			✓				✓	
[33]			✓							✓					✓							✓	
[34]																					✓	✓	
[35]					✓					✓			✓							✓		✓	
[36]	✓		✓							✓													✓
[37]			✓							✓								✓		✓			✓
[38]			✓							✓		✓			✓								
[39]										✓								✓				✓	
[40]																					✓	✓	
[41]											✓											✓	
[42]										✓					✓							✓	
[43]	✓		✓																			✓	
[44]										✓													
[45]	✓									✓										✓		✓	
[46]					✓					✓													
[47]	✓		✓								✓									✓		✓	
[48]										✓	✓				✓			✓				✓	
[49]			✓							✓												✓	
[50]	✓										✓							✓		✓			
[51]	✓				✓					✓	✓				✓					✓			
[52]										✓	✓				✓						✓	✓	
[53]			✓							✓	✓							✓	✓	✓			✓
[54]	✓									✓	✓												

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

TABLO 3 METHODS USED IN THE DATA PREPROCESSING IN THE STUDIES

	Data Integration	Data Cleaning	Data Reduction	Data Transform
Listwise Deletion		[42],[33]		
Linear Regression		[33],[32],[37]		
Piecewise Aggregate Approximation			[42],[53]	
Linear normalization				[33],[32],[34] [35],[39],[41] [42],[43],[45] [47],[48],[49] [52],[54]
Principle Component Analysis			[33],[37],[34]	
Linear Interpolation		[4],[50]		
Cross-correlation Analysis			[50]	
Similar Day Approach		[50]		
Conditional permutation importance score			[50]	
Logistic Regression		[32]		
Piecewise Aggregate Approximation			[48],[53]	
Averaging		[48]		
Sense Checking Validity		[49],[43]		
Z-score				[37],[36],[53]
Nearest Neighbor interpolation		[38]		
Expectation Maximization		[38]		
Single Exponential Smoothing Technique		[52]		
Seasonal Auto Regressive Moving Average		[52]		
Savitzky–Golay Digital Filter		[52]		
Self Organizing Map		[52]		
Peak-Valley Attribute Analysis			[52]	
Sammon Map			[34]	
Curvilinear Component Analysis			[34]	
Symbolic Aggregation Approximation			[53]	
Extract, Transform and Load	[51]			
Recursive Feature Elimination			[51]	
Lasso Regularization			[51]	
Multilayer Perceptron Artificial Neural Network		[35]		
Forward Filling Method		[54]		
Sequential Backward Search			[54]	
Anderson-Darling test			[38]	
Durbin – Watson test			[38]	

It has been observed that, one of the main strategies adopted in the studies is avoiding from the computational effort. For this reason, the data with high temporal resolution are generally subjected to data reduction preprocessing step. As a matter of fact, high temporal resolution data is not required in load profile clustering [42]. The same strategy has been adopted in determining the methods used in data preprocessing. For example, in the outlier detections or data imputations, the methods that are easy to apply and have less computational burden have been generally preferred.

Although a large number of studies have been realized on electrical energy consumption profile clustering, very few of these studies have shared details about the preprocessing steps used within the scope of the study. These studies, which generally focus on proving the validity of the clustering approach or method they propose, do not provide details on how the data used in the case analysis were obtained. However, in practice, the first step for practitioners is to create the appropriate dataset from the raw data.

This study provides an overview of the methods used in data preprocessing. It aims to serve as a guide for researchers in order to determine the methods with functionality in accordance with the data used and intended outputs in their studies. Although the study focuses on electrical energy consumption data, it also provides details in the processing of data belonging to different disciplines.

Important findings about the performance of preprocessing methods may be reached by analyzing

different big data sets with identical systems. Defining the relationship between data-specific properties such as Variety, Volume, Velocity, Veracity, Value and the performance of preprocessing methods may provide significant contributions to the literature.

5. ACKNOWLEDGEMENTS

Provide acknowledgement if any.

REFERENCES

[1] Mansour Charwand, and Mohsen Gitizadeh, (2017), "Optimal TOU tariff design using robust intuitionistic fuzzy divergence based thresholding", Energy, vol. 147, pp. 655–662.
[2] Hyun Cheol Jeong, Jaesung Jung, and Byung O. Kang, (2020), "Development of Operational Strategies of Energy Storage System Using Classification of Customer Load Profiles under Time-of-Use Tariffs in South Korea", Energies, Vol. 13, Issue. 7, pp. 1723.
[3] Derck Koolen, Navid Sadat-Razavi, and Wolfgang Ketter, (2017), "Machine Learning for Identifying Demand Patterns of Home Energy Management Systems with Dynamic Electricity Pricing", Applied Sciences, Vol. 7, Issue. 11, pp. 1160.
[4] Bishnu Nepal, Motoi Yamaha, Aya Yokoe, and Toshiya Yamaji, (2020), "Electricity load forecasting using clustering and ARIMA model for energy management in buildings", Japan Architectural Review, Vol. 3, Issue. 1, pp. 62–76.
[5] Ravindra R. Rathod, and Rahul Dev Garg, (2017), "Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data", International Journal of Energy Sector Management, Vol. 11, Issue. 2, pp. 295–310.
[6] Patricia R. S. Jota, Valéria. R. B. Silva, and Fábio G. Jota, (2011), "Building load management using cluster and statistical analyses", International Journal of Electrical Power & Energy Systems, Vol. 33, Issue. 8, pp. 1498–1505.
[7] Selin Yilmaz, Jonathan D. Chambers, and Martin K. Patel, (2019), "Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management", Energy, Vol. 180, pp. 665–677.
[8] David F. Rogers, and George G. Polak, (2013), "Optimal Clustering of Time Periods for Electricity Demand-Side Management", IEEE Transactions on Power Systems, Vol. 28, Issue. 4, pp. 3842–3851.
[9] Rita Pereira, A. Fagundes, et al., (2016), "A fuzzy clustering approach to a demand response model", International Journal of Electrical Power & Energy Systems, Vol. 81, pp. 184–192.
[10] Maria A. Z. Alvarez, Kodjo Agbossou, Alben Cardenas, Sousso Kelouwani, and Loıc Boulon, (2020), "Demand Response Strategy Applied to Residential Electric Water Heaters Using Dynamic Programming and K-Means Clustering", IEEE Transactions on Sustainable Energy, Vol. 11, Issue. 1, pp. 524–533.
[11] Zafar A. Khan and Dilan Jayaweera, (2020), "Smart Meter Data Based Load Forecasting and Demand Side Management in Distribution Networks with Embedded PV Systems", IEEE Access, Vol. 8, pp. 2631–2644.
[12] Haihong Bian, Yiqun Zhong, Jianshuo Sun, and Fangchu Shi, (2020), "Study on power consumption load forecast based on K-means clustering and FCM–BP model", Energy Reports, Vol. 6, pp. 693–700.
[13] Jamer Jiménez Mares, Loraine Navarro, Christian G. Quintero M., and Mauricio Pardo, (2020), "A Methodology for Energy Load Profile Forecasting Based on Intelligent Clustering and Smoothing Techniques", Energies, Vol. 13, Issue. 16, p. 4040.
[14] Mohamed Chaouch, (2014), "Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves", IEEE Transactions on Smart Grid, Vol. 5, Issue. 1, pp. 411–419.
[15] Salvador García, Julián Luengo, and Francisco Herrera, (2015), "Data Preprocessing in Data Mining", Springer.
[16] Wen Chen, Kaile Zhou, Shanlin Yang, and Cheng Wu, (2017), "Data quality of electricity consumption data in a smart grid environment", Renewable and Sustainable Energy Reviews, Vol. 75, pp. 98–105.
[17] Vidyasagar Potdar, Anulipt Chandan, Saima Batool, and Naimesh Patel, (2018), "Big Energy Data Management for Smart Grids---Issues, Challenges and Recent Developments", Smart Cities: Development and Governance Frameworks, Springer International Publishing, pp. 177–205.
[18] Manuel Pereira, Nuno Velosa, and Lucas Pereira, (2019), "dsCleaner: A Python Library to Clean, Preprocess and Convert Non-Instrusive Load Monitoring Datasets", Data, Vol. 4, Issue. 3, pp. 123.
[19] Yang Zhang, Tao Huang, and Ettore F. Bompard, (2018), "Big data analytics in smart grids: a review", Energy Informatics, Vol. 1, Issue. 1, pp. 8.
[20] Vic Barnett, and Tobby Lewis, (1994), "Outliers in Statistical Data", Wiley.
[21] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José M. Benítez, and Francisco Herrera, (2016), "Big data preprocessing: methods and prospects", Big Data Analytics, Vol. 1, Issue. 1.
[22] Li Sun, Kaile Zhou, Xiaoling Zhang, and Shanlin Yang, (2018), "Outlier Data Treatment Methods Toward Smart Grid Applications", IEEE Access, Vol. 6, pp. 39849–39859.
[23] Hongzhi Wang, Mohamed J. Bah, and Mohamed Hammad, (2019), "Progress in Outlier Detection Techniques: A Survey", IEEE Access, Vol. 7, pp. 107964–108000.
[24] Yassine Himeur, Khalida Ghanem, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira, (2021), "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives", Applied Energy, Vol. 287.
[25] Mehdi Pazhoohesh, Adib Allahham, Ronnie Das, and Sara Walker, (2021), "Investigating the impact of missing data imputation techniques on battery energy management system", IET Smart Grid, Vol. 4, Issue. 2, pp. 162–175.
[26] Muhammed S. Osman, Adnan M. Abu-Mahfouz, and Philip R. Page, (2018), "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case", IEEE Access, Vol. 6, pp. 63279–63291.
[27] Thamer Alquthami, Ahmed AlAmoudi, Abdullah M. Alsubaie, Abdulrahman Bin Jaber, Nassir Alshlwan, Murad Anwar, and Shafi Al Husaien, (2020), "Analytics framework for optimal smart meters data processing", Electrical Engineering, Vol. 102, Issue. 3, pp. 1241–1251.
[28] Irena Koprinska, Mashud Rana, and Vassilios G. Agelidis, (2015), "Correlation and instance based feature selection for electricity load forecasting", Knowledge-Based Systems, Vol. 82, pp. 29–40.
[29] Aida M. Pirbazari, Antorweep Chakravorty, and Chunming Rong, (2019), "Evaluating Feature Selection Methods for Short-Term Load Forecasting", 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 1–8.
[30] A. Jovic, K. Brkic, and N. Bogunovic, (2015), "A review of feature selection methods with applications", 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200–1205.
[31] Chih-Fong Tsai, William Eberle, and Chi-Yuan Chu, (2013), "Genetic algorithms in feature and instance selection", Knowledge-Based Systems, Vol. 39, pp. 240–247.
[32] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, (2005), "An electric energy consumer characterization framework based on data mining techniques", IEEE Transactions on Power Systems, Vol. 20, Issue. 2, pp. 596–602.
[33] M. Espinoza, C. Joye, R. Belmans, and B. DeMoor, (2005), "Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume: 8 Issue: 4, April 2023, pp. 1 – 13

Time Series" IEEE Transactions on Power Systems, Vol. 20, Issue. 3, pp. 1622–1630.
[34] G. Chicco, R. Napoli, and F. Piglione, (2006), "Comparisons Among Clustering Techniques for Electricity Customer Classification", IEEE Transactions on. Power Systems, Vol. 21, Issue. 2, pp. 933–940.
[35] Sergio Ramos, Zita Vale, Joao Santana, and Jorge Duarte, (2007), "Data Mining Contributions to Characterize MV Consumers and to Improve the Suppliers-Consumers Settlements", 2007 IEEE Power Engineering Society General Meeting, pp. 1–8.
[36] Tiefeng Zhang, Guangquan Zhang, Jie Lu, Xiaopu Feng, and Wanchun Yang, (2012), "A new index and classification approach for load pattern analysis of large electricity customers", IEEE Transactions on Power Systems, Vol. 27, Issue. 1, pp. 153–160.
[37] Matti Koivisto, Pirjo Heine, Ilkka Mellin, and Matti Lehtonen, (2013), "Clustering of connection points and load modeling in distribution systems", IEEE Transactions on Power Systems, Vol. 28, Issue. 2, pp. 1255–1265.
[38] Adrian Albert and Ram Rajagopal, (2013), "Smart Meter Driven Segmentation: What Your Consumption Says About You", IEEE Transactions on Power Systems, Vol. 28, Issue. 4, pp. 4019–4030.
[39] Jungsuk Kwac, June Flora, and Ram Rajagopal, (2014), "Household Energy Consumption Segmentation Using Hourly Data", IEEE Transactions on Smart Grid, Vol. 5, Issue. 1, pp. 420–430.
[40] Joshua D. Rhodes, Wesley J. Cole, Charles R. Upshaw, Thomas F. Edgar, and Michael E. Webber, (2014), "Clustering analysis of residential electricity demand profiles", Applied Energy, Vol. 135, pp. 461–471.
[41] Sérgio Ramos, João M. Duarte, F. Jorge Duarte, and Zita Vale, (2015), "A data-mining-based methodology to support MV electricity customers' characterization", Energy and Buildings, Vol. 91, pp. 16–25.
[42] Ramon Granell, Colin J. Axon, and David C. H. Wallom, (2015), "Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles", IEEE Transactions on Power Systems, Vol. 30, Issue. 6, pp. 3217–3224.
[43] Ran Li, Chenghong Gu, Furong Li, Gavin Shaddick, and Mark Dale, (2015), "Development of Low Voltage Network Templates—Part I: Substation Clustering and Classification", IEEE Transactions on Power Systems, Vol. 30, Issue. 6, pp. 3036–3044.
[44] Akito Ozawa, Ryota Furusato, and Yoshikuni Yoshida, (2016), "Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles", Energy and Buildings, Vol. 119, pp. 200–210.
[45] Joaquim L. Viegas, Susana M. Vieira, R. Melício, V. M. F. Mendes, and João M. C. Sousa, (2016), "Classification of new electricity customers based on surveys and smart metering data", Energy, Vol. 107, pp. 804–817.
[46] Kaile Zhou, Shanlin Yang, and Zhen Shao, (2017), "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study", Journal of Cleaner Production, Vol. 141, pp. 900–908.
[47] Jatin Bedi and Durga Toshniwal, (2018), "Empirical Mode Decomposition Based Deep Learning for Electricity Demand Forecasting", IEEE Access, Vol. 6, pp. 49144–49156.
[48] Shunfu Lin, Fangxing Li, Erwei Tian, Yang Fu, and Dongdong Li, (2019), "Clustering Load Profiles for Demand Response Applications", IEEE Transactions on Smart Grid, Vol. 10, Issue. 2, pp. 1599–1607.
[49] Mudassir Azizahmed Maniar, and Abhijit R. Abhyankar, (2019), "Two-Stage Load Profiling of HV Feeders of a Distribution System", IEEE Systems Journal, Vol. 13, Issue. 3, pp. 3102–3110.
[50] Kanggu Park, Seungwook Yoon, and Euiseok Hwang, (2019), "Hybrid Load Forecasting for Mixed-Use Complex Based on the Characteristic Load Decomposition by Pilot Signals", IEEE Access, Vol. 7, pp. 12297–12306.
[51] Simona-Vasilica Oprea, and Adela Bâra, (2019), "Machine Learning Algorithms for Short-Term Load Forecast in Residential Buildings Using Smart Meters, Sensors and Big Data Solutions", IEEE Access, Vol. 7, pp. 177874–177889.
[52] Kushan Ajay Choksi, Sonal Jain, and Naran M. Pindoriya, (2020), "Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset", Sustainable Energy, Grids and Networks, Vol. 22, pp. 100346.
[53] Ying Shi, Tao Yu, Qianjin Liu, Hanxin Zhu, Fusheng Li, and Yaxiong Wu, (2020), "An Approach of Electrical Load Profile Analysis Based on Time Series Data Mining", IEEE Access, Vol. 8, pp. 209915–209925.
[54] Dabeeruddin Syed, Haitham Abu-Rub, et al., (2021), "Deep Learning-Based Short-Term Load Forecasting Approach in Smart Grid With Clustering and Consumption Pattern Recognition", IEEE Access, Vol. 9, pp. 54992–55008.
[55] Desh Deepak Sharma, S. N. Singh, Jeremy Lin, and Elham Foruzan, (2017), "Identification and characterization of irregular consumptions of load data", Journal of Modern Power Systems and Clean Energy, Vol. 5, Issue. 3, pp. 465–477.

Authors Biography

Mustafa Şen Yıldız, is a research assistant in Kırklareli University. He completed BSc in Electrical and Electronics Engineering Department in Karaelmas University. He received MSc degree in Electrical Engineering Department at Yıldız Technical University. His research interests are smart grid, demand response, load forecasting, artificial intelligence, and deep learning.

Kadir Doğanşahin, is an assistant professor in Artvin Çoruh University. He received BSc in Electrical Engineering Department in Yıldız Technical University. His MSc degree is from College of Engineering at University of Texas at San Antonio. He received PhD degree from Electrical Engineering Department in Yıldız Technical University. His research interests are power distribution systems, power quality and photovoltaic systems.

Bedri Kekezoğlu, is an Associated Professor in Yıldız Technical University. He completed BSc, MSc and PhD degrees in Electrical Engineering Department from Yıldız Technical University. His research interests are power system analysis, electric power transmission, distribution and protection and renewable energy.

Cite this paper:

Mustafa Ş. Yıldız, Kadir Doğanşahin, Bedri Kekezoğlu, "Data Preprocessing in Electrical Energy Consumption Profile Clustering Studies", International Journal of Advances in Computer and Electronics Engineering, Vol. 8, No. 4, pp. 1-13, April 2023.