American Journal of Applied Mathematics and Statistics, 2022, Vol. 10, No. 1, 22-27
Available online at http://pubs.sciepub.com/ajams/10/1/4
Published by Science and Education Publishing
DOI:10.12691/ajams-10-1-4

Clustering Time Related Data: A Regression Tree Approach

K.A.D. Deshani[1,*], Liwan Liyanage-Hansen[2], Dilhari T. Attygalle[1]

[1]Department of Statistics, University of Colombo, Colombo 03, Sri Lanka
[2]School of Computing, Engineering and Mathematics, University of Western Sydney, Campbelltown, Australia
*Corresponding author: deshani@stat.cmb.ac.lk

Received February 12, 2022; Revised March 15, 2022; Accepted March 23, 2022

Abstract With the advancement of technology, vast time related databases are created from a plethora of processes. Analyzing such data can be very useful, but due to the large volumes and their relevance to time, extracting useful information and implementing models can be very complex and time consuming. However, using a comprehensive exploratory study to extract hidden features of the data can mitigate this complexity to a great extent. The clustering approach is one such way to extract features but can be demanding with time related data, especially with a trend in the data series. This paper proposes an algorithm, based on regression tree approach, to cluster a time series with a trend, along with other relevant variables. The importance of this algorithm is avoiding the misleading cluster allocations that can be created through clustering a differenced time series. Initially it identifies a suitable consistent time window with no trend, and implements separate regression trees for each window, to obtain the clusters. Through exploring the clusters generated from these trees, a general cluster formation is identified suitable for all windows. This is illustrated using hourly electricity demand in Sri Lanka for five consecutive years. Six meaningful clusters were identified based on the day of the week, specialty, and the time of the day. These cluster memberships provide useful additional information on the data structure, independent of the trend component, and can be used as an additional feature for improving model accuracies.

Keywords: *clustering, regression tree, load forecasting, trend, data exploration*

Cite This Article: K.A.D. Deshani, Liwan Liyanage-Hansen, and Dilhari T. Attygalle, “Clustering Time Related Data: A Regression Tree Approach.” *American Journal of Applied Mathematics and Statistics*, vol. 10, no. 1 (2022): 22-27. doi: 10.12691/ajams-10-1-4.

1. Introduction

This can be called as the era of data science where millions of data are recorded within a split of a second where most of such collected data are time related. Such datasets may have hidden structures related to time and thus demanding when extracting information and implementing models. A thorough understanding of the structure of the data is one strategy to meet this challenge. Literature reveals that clustering approach is an effective way to explore data especially with large volumes.

A ‘trend’, which is an increasing or decreasing slope, is commonly observed in a time series. Usually when dealing with such data, the series is detrended by differencing, prior to analysis. When clustering the detrended series, it will be based on the differenced values rather than the observed values of the series and hence may affect the cluster formations substantially [6,7]. On the other hand, if the series is not detrended prior to clustering, then the clusters will be formed mainly based on time, rather than the true structure of the data. Thus, careful attention must be paid when clustering such data.

Researchers have identified that including cluster membership details of the data when implementing models, improves the accuracy of forecasts [4,5]. However, when the time series data depicts a trend, there is evidence that the clustering process will not generate meaningful results [6,7]. This paper illustrates a clustering algorithm to address this issue, also accommodating other related variables both numerical and categorical in nature.

2. Literature Review

Literature related to time series clustering are broadly categorized into three types of research, namely, whole time series clustering, subsequence time series clustering and time point clustering [1]. This article is mainly aimed at time point clustering in a univariate time series and is based on clustering electricity consumptions or demands. Many studies have been done on this regard when forecasting time series data utilizing statistical approaches, artificial intelligence approaches and hybrid approaches.

To make the forecasting process more effective, some researchers have tried to categorize similar demands together. They have grouped days together based on the demand values as the day type directly influences the electricity consumption. Lee, Cha and Park have classified daily load curves based on weekday and weekend-day patterns into 5 groups considering the position of each week in a particular month. However, the technique used to classify curves was not presented, and therefore assumed to have taken place based on expert's knowledge [11].

Valero, Sanabre and Aparicio have trained five different Self Organizing Maps (SOM) for similar day types where the categories consisted of category 1: Sundays, category 2: Saturdays, category 3: Fridays, category 4: Mondays and days after a holiday and category 5: Tuesdays, Wednesdays and Thursdays [12].

Chicco, Napoli & Piglione have compared different clustering techniques that can be used to cluster similar demands. In that article, the authors have used four unsupervised clustering algorithms: hierarchical clustering, k-means, fuzzy k-means and SOM to identify the cluster memberships [3]. Among the unsupervised clustering approaches, k-means clustering algorithm has been a common choice by many researchers to cluster electricity load related data [10,13,14]. Taking a different approach, a study has been based on the regression tree technique to explore the dependency between the factors that influence the electricity demand. By using this technique, a rule-based feature selection procedure has been obtained from the tree structure. The main advantage of using this technique is that it can be used for data with both numerical and categorical variables [8]. It is worth noting that some researchers have used the regression tree approach even to forecast the hourly electricity demand [9].

In the Sri Lankan context, several studies have investigated the possibilities of clustering electricity demands considering different time levels as half-hourly [6] and daily levels [7]. The main aim of these have been to increase the effectiveness of the short-term load forecasting models by incorporating the cluster identities derived from the analysis. Moreover, it was shown how clustering can be fruitfully utilized to increase accuracy levels of the forecasting models and to substantially reduce training times [4,5].

Literature reveals that less attention is given on the trend component, when addressing the issue of clustering time points in a time series with a trend. A common approach used to detrend a series is differencing, but there are situations where taking the difference will affect the inherent structure of the series [7].

3. Proposed Algorithm and Application

This algorithm proposes a remedial approach to cluster time series data with a trend in the presence of both numerical and categorical variables. The pseudocode for the algorithm and an application follows.

3.1. Pseudocode: Time Series Data Clustering Using Regression Trees

PSEUDOCODE: TIME SERIES DATA CLUSTERING USING REGRESSION TREES

```
# This pseudocode can be used to cluster functional data with a trend
INPUT:
Series = Time series to be clustered
OtherInputs = A set of other input variables needed for
              clustering
OUTPUT:
ClusterID = Cluster ID of the data point

START
# Explore the time series
# Decide on the maximum possible time period of the moving window which does not depict a trend = WindowSize (Start from the beginning).
# Replicates = Number of consecutive time segments with the length WindowSize
FOR (all the time series windows)
{
# Fit a regression tree model with suitable predictor variables
# Prune the tree if necessary
# Explore the cluster allocations
}
# Compare the pattern of cluster allocations of each time series segment and try to generalize the clustering pattern if possible.
END
```

3.2. An Application of the Algorithm for Clustering Electricity Demand Data

The data used for the illustration purpose consist of hourly total electricity demand in Sri Lanka from 1st January 2008 to 31st December 2012. This series is given in Figure 1 that displays the fluctuations of hourly electricity demand showing a stochastic trend. It can also be seen that there is a periodic repetitive pattern in the series where further explorations revealed that during mid-April each year the demand is comparatively low. The reason behind this is identified as the Sinhala-Tamil new year season which is holiday time for all Sri Lankans.

The other related variables and their descriptions are displayed in Table 1.

Table 1. Variables used for the analysis

Variable Name	Description
Time	Time of the day: AM1, AM2, AM3, … , AM12
Day	Day of the week: Sunday, Monday, …. , Saturday
Specialty	Specialty of the day: None, Poya day, PBM Holiday (PBM), PB Holiday, working day before a holiday (WDBH), working day after a holiday (WDAH), working day between a holiday and weekend (WDBHW), Saturday after a holiday (SAH), one day after New Year (ODA) and two days after New Year (TDA)
Month	Month of the year: January, February, … , December

Step 1:

When considering the algorithm, the first step is to decide on the size of the moving window that does not depict a trend. For this series, the ***WindowSize*** was identified as one year. When each year's demand data were tested for stationarity, the trend term in the ADF test was not significant in the model. This ensured that there is no trend within each considered year.

Step 2:

For each year, a regression tree was fitted with all the variables that were available in relation to the time series. Consequently, a combination of predictors was added to the model, where ***Month*** was not significant in any of the models. Thus, for each year, the regression trees for the model "***Demand ~ Time + Day + Specialty***" were generated.

Step 3:

Trees are grown to a maximal size without the use of a stopping rule; essentially the tree-growing process stops when no further splits are possible due to lack of data. In practical situations the tree should be stopped further growing by pruning the tree to obtain the optimal tree [15]. The stopping rule employed in this study is by observing the rate of decrement of the complexity parameter (CP) values with respect to the number of splits.

The regression tree created for the year 2008 is presented to showcase the third step of the algorithm. The complexity parameter (CP) values for the year 2008 shown in Table 2 revealed that the decrement of the CP is very low with the increment of number of splits. Therefore, it was decided to stop splitting the tree further as the decrement of the overall lack of fit is lower than 0.02, which was selected by observing the decrement of cross-validation error with respect to number of splits. It should be noted that this specific criterion for pruning was considered as opposed to usual one Standard Error (SE) rule [2], due to the less meaningful splits of the tree.

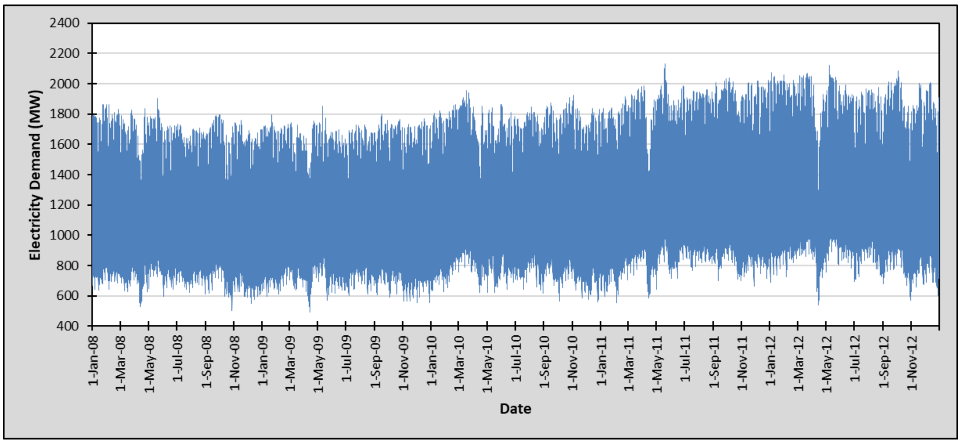

Figure 1. Hourly electricity demands over the years

Table 2. Complexity parameters and their associated errors for 2008

	CP	No. of splits	Relative error	Cross-vali error	Cross-vali error std dev
1	0.4930	0	1.0000	1.0004	0.0131
2	0.2376	1	0.5069	0.5071	0.0086
3	0.0655	2	0.2693	0.2695	0.0038
4	0.0432	3	0.2037	0.2062	0.0034
5	0.0343	4	0.1605	0.1629	0.0032
6	**0.0135**	**5**	**0.1262**	**0.1285**	**0.0022**
7	0.0100	6	0.1126	0.1131	0.0021

The final pruned regression tree is shown in Figure 2. The most impacting splitting variable on electricity demand is the ***Time*** of the day. Initially the tree was split into, low demand hours and high demand hours. Low demand hours were early morning and late-night hours whereas the high demand hours were mainly during daytime and late evenings. The high demanded hours were further impacted initially by ***Day*** followed by ***Specialty***, whereas the low demanded hours did not. The resulting tree was generated with 2 leaf nodes for the low demanded hours and 4 for the high demanded hours.

The 6 groups of demand for the year 2008 are displayed below.

Group 1: Demands at AM12, AM1, AM2, AM3, AM4

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON 12	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
🟦	🟦	🟦	🟦																				🟦

Group 2: Demands at AM5, AM7, AM8, PM11

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON 12	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
				🟩		🟩	🟩															🟩	

Group 3: Demands at AM6, AM9, AM10, AM11, NOON12, PM1, PM2, PM3, PM4, PM5, PM6, PM10 on Sundays

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON 12	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
					🟨			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				🟨		

Group 4: Demands at AM6, AM9, AM10, AM11, NOON12, PM1, PM2, PM3, PM4, PM5, PM6, PM10 on Monday to Saturday where Specialty is One day after New Year, Two days after New Year, PBM Holiday, Poya day, Saturday after a holiday

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON 12	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
					🟨			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				🟨		

Group 5: Demands at AM6, AM9, AM10, AM11, NOON12, PM1, PM2, PM3, PM4, PM5, PM6, PM10 on Monday to Saturday where Specialty is None, PB holiday, Working day after a holiday, working day after PB holiday, working day before a holiday, working day before a PB holiday, working day between a PB holiday and weekend, working day between a holiday and weekend

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON 12	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
					🟨			🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				🟨		

Group 6: Demands at PM7, PM8, PM9

AM1	AM2	AM3	AM4	AM5	AM6	AM7	AM8	AM9	AM10	AM11	NOON	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	AM12
																		🟥	🟥	🟥			

A similar procedure was carried out for the other four years, and the summary is in Table 3.

The structures of all the trees generated similar results of six groups of hourly demands, but with one exception.

The similarities are as follows:

- Early morning (12.00 a.m., 1.00 a.m., 2.00 a.m., 3.00 a.m., 4.00 a.m.), morning (5.00 a.m., 7.00 a.m., 8.00 a.m.) and peak time (7.00 p.m., 8.00 p.m., 9.00 p.m.) demands did not have any effect on the day of the week or the specialty of the day.
- The electricity demands during 6.00 a.m. to 6.00 p.m. highly depend on the day of the week and the specialty.
- Recall that there were 11 categories for the Specialty of the day as described in Table 1. These specialties were grouped into two where PBM, Poya day, SAH, ODA, TDA grouped together and the rest combining with the category None. The possible reason for this grouping may be that PBM, Poya day, SAH, ODA, TDA are holidays for all Sri Lankans whereas the other specialty days are enjoyed only by a portion of Sri Lankans.

The only exception is that in 2008 and 2012, the demand of working hours of Sundays were separately grouped, while in other years, Saturdays and Sundays were grouped into one. It should be noted that similar grouping pattern could also be found in previous research when clustering electricity demand in Sri Lanka at different levels [6,7].

These identified cluster memberships (six categories) can be used as an additional feature in the modeling process. A study using the same dataset to forecast hourly electricity demand using a dynamic iterative neural network has shown improved accuracies while also reducing training times, through incorporating cluster memberships in forecasting models [4]. When considering statistical techniques, [5] shows that the cluster membership details have been selected as one of the best regressor variables to forecast a functional principal component score series using an ARIMA model by using the same dataset.

4. General Discussion and Conclusion

This paper proposes an algorithm that can be used to identify cluster memberships of each time point of a time series with a trend, considering other numerical and categorical variables. A commonly used approach of clustering a time series is to first de-trend the series and then use clustering. Such an approach suffers from the differenced series being clustered rather than the observed series that can produce misleading clusters. The novelty of this research is to propose an alternative approach of clustering a time series with a trend that generates promising results. The algorithm was illustrated using a real life time series dataset related to electricity demand in Sri Lanka. For this data, six clusters were identified where clusters were mainly separated based on time of the day, day of the week and the specialty. The results clearly show that using the proposed clustering algorithm, a comprehensive structural information regarding time series data with a trend can be obtained. More importantly, the cluster membership can be incorporated as an additional feature in any predictive modelling process for improved accuracy and efficiency. However, if the clustering is less meaningful, the level of impact to the modelling procedure will be less.

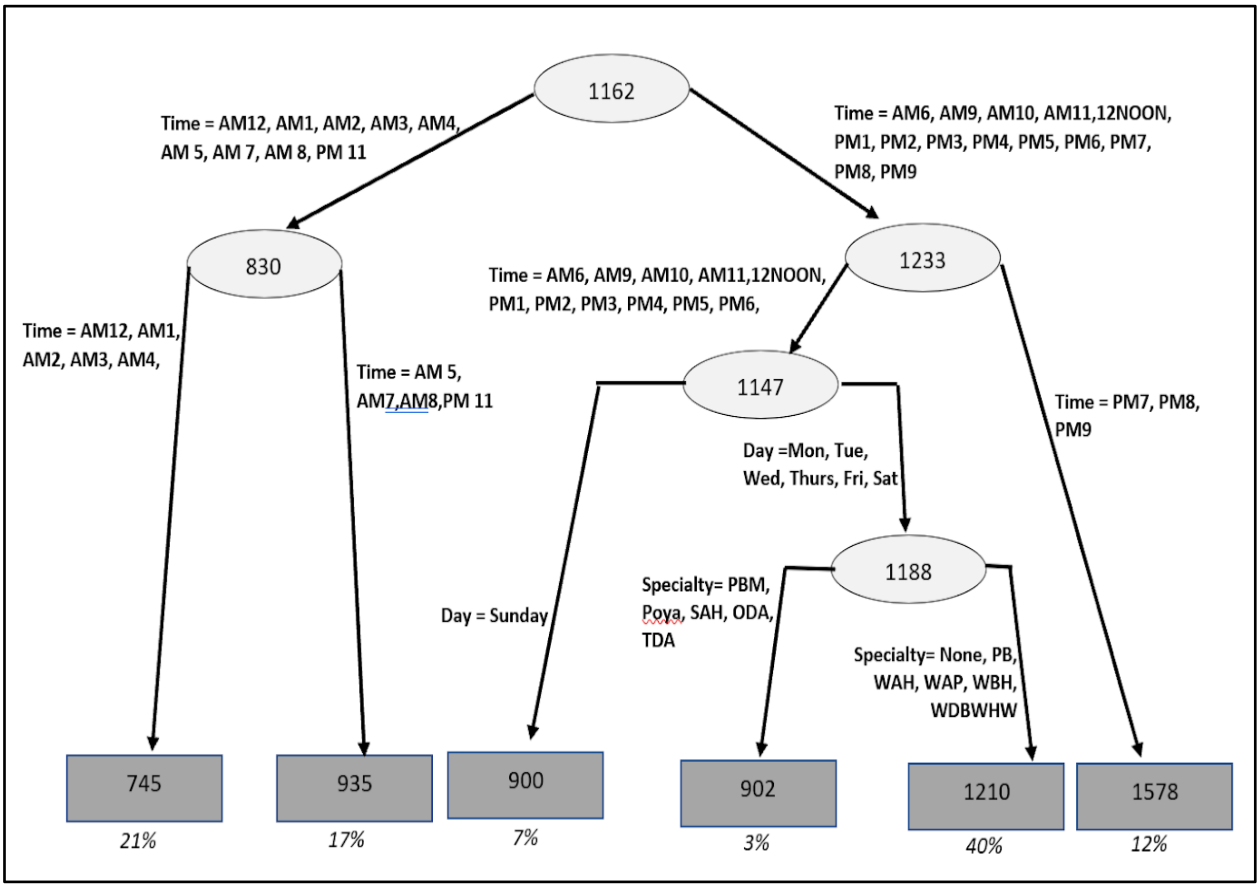
1162
Time = AM12, AM1, AM2, AM3, AM4, AM 5, AM 7, AM 8, PM 11
Time = AM6, AM9, AM10, AM11,12NOON, PM1, PM2, PM3, PM4, PM5, PM6, PM7, PM8, PM9
830
1233
Time = AM6, AM9, AM10, AM11,12NOON, PM1, PM2, PM3, PM4, PM5, PM6,
Time = AM12, AM1, AM2, AM3, AM4,
Time = AM 5, AM7,AM8,PM 11
1147
Time = PM7, PM8, PM9
Day =Mon, Tue, Wed, Thurs, Fri, Sat
Day = Sunday
1188
Specialty= PBM, Poya, SAH, ODA, TDA
Specialty= None, PB, WAH, WAP, WBH, WDBWHW
745
935
900
902
1210
1578
21%
17%
7%
3%
40%
12%

Figure 2. Pruned Regression tree for 2008 demand (Demand ~ Time + Day + Specialty)

Table 3. Summary of the regression trees for years 2008 – 2012

	2008	**2009**	**2010**	**2011**	**2012**
	12.00 a.m.	12.00 a.m.	12.00 a.m.	12.00 a.m.	12.00 a.m.
	1.00 a.m.	1.00 a.m.	1.00 a.m.	1.00 a.m.	1.00 a.m.
	2.00 a.m.	2.00 a.m.	2.00 a.m.	2.00 a.m.	2.00 a.m.
	3.00 a.m.	3.00 a.m.	3.00 a.m.	3.00 a.m.	3.00 a.m.
	4.00 a.m.	4.00 a.m.	4.00 a.m.	4.00 a.m.	4.00 a.m.
	5.00 a.m.	5.00 a.m.	5.00 a.m.	5.00 a.m.	5.00 a.m.
	7.00 a.m.	7.00 a.m.	7.00 a.m.	7.00 a.m.	7.00 a.m.
	8.00 a.m.	8.00 a.m.	8.00 a.m.	8.00 a.m.	8.00 a.m.
	11.00 p.m.	11.00 p.m.	11.00 p.m.	11.00 p.m.	11.00 p.m.
6.00 a.m. 9.00 a.m. 10.00 a.m. 11.00 a.m. 12.00 noon 1.00 p.m. 2.00 p.m. 3.00 p.m. 4.00 p.m. 5.00 p.m. 6.00 p.m. 10.00 p.m.	**Sunday**	**Sundays and Saturdays**	**Sundays and Saturdays**	**Sundays and Saturdays**	**Sunday**
	(Monday-Saturday) Specialty: OneDayAfterNY, TwoDaysAfterNY, PBM, Poyaday, SatAfterHoli	**(Monday-Friday)** Specialty: OneDayAfterNY, TwoDaysAfterNY, PBM, Poyaday	**(Monday-Friday)** Specialty: OneDayAfterNY, TwoDaysAfterNY, PBM, Poyaday	**(Monday-Friday)** Specialty: OneDayAfterNY, TwoDaysAfterNY, PBM, Poyaday	**(Monday-Saturday)** Specialty: OneDayAfterNY, TwoDaysAfterNY, PBM, Poyaday, SatAfterHoli
	(Monday-Saturday) Specialty: None, PB, WorkAfterHoli, WorkAfterPB, WorkB4Holiday, WorkB4PB, WorkBW_PBHoli&WEnd, WorkBWHoli&WEnd	**(Monday-Friday)** Specialty: None, PB, WorkAfterHoli, WorkAfterPB, WorkB4Holiday, WorkB4PB, WorkBWHoli&Wend	**(Monday-Friday)** Specialty: None, PB, WorkAfterHoli, WorkAfterPB, WorkB4Holiday WorkB4PB, WorkBWHoli&Wend	**(Monday-Friday)** Specialty: None, PB, WorkAfterHoli, WorkAfterPB, WorkB4Holiday WorkB4PB, WorkBWHoli&Wend	**(Monday-Saturday)** Specialty: None, PB, WorkAfterHoli, WorkAfterPB, WorkB4Holiday, WorkB4PB, WorkBW_PBHoli&WEnd, WorkBWHoli&WEnd
	7.00 p.m.	7.00 p.m.	7.00 p.m.	7.00 p.m.	7.00 p.m.
	8.00 p.m.	8.00 p.m.	8.00 p.m.	8.00 p.m.	8.00 p.m.
	9.00 p.m.	9.00 p.m.	9.00 p.m.	9.00 p.m.	9.00 p.m.

References

[1] Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering – A decade review. Information Systems, 53, 16-38.

[2] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. (1984). Classification and Regression Trees. Taylor & Francis.

[3] Chicco, G., Napoli, R., & Piglione, F. (2006). Comparisons among Clustering Techniques for Electricity Customer Classification. IEEE Transactions on Power Systems, 933-940.

[4] Deshani K.A.D, Liyanage-Hansen L. and Attygalle D. (2019). Artificial Neural Network for Dynamic Iterative Forecasting: Forecasting Hourly Electricity Demand, American Journal of Applied Mathematics and Statistics, Vol. 7, No. 1, January 2019.

[5] Deshani K.A.D, Attygalle D., Liyanage-Hansen L. and Lakraj G.P., (2017). Dynamic Short Term Load Forecasting using Functional Principal Component Regression, International Conference on Machine Learning and Data Engineering Sydney, Australia.

[6] Deshani, K.A.D, Attygalle, M.D.T, Hansen, L. L., & Karunarathne, A. (2014). An Exploratory Analysis on Half-Hourly Electricity Load Patterns Leading to Higher Performances in Neural Network Predictions. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 5, No. 3, May 2014 (pp. 37-51)

[7] Deshani, K.A.D, Hansen, L. L., Attygalle, M.D.T, & Karunarathne, A. (2014). Improved Neural Network Prediction Performances of Electricity Demand: Modifying Inputs through Clustering. Second International Conference on Computational Science and Engineering (pp. 137-147). India: AIRCC.

[8] Gładysz, Barbara and Kuchta, Dorota, (2008). Application of regression trees in the analysis of electricity load, Operations Research and Decisions, 4, issue, p. 19-28.

[9] Hambali, M., Akinyemi, A., Oladunjoye, J., & Yusuf, N. (2016). Electric Power Load Forecast Using Decision Tree Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal, 29-42.

[10] Hernández, L., Baladrón, C., Aguiar, J., Carro, B., & Sánchez-Esguevillas, A. (2012). Classification and Clustering of Electricity Demand Patterns in Industrial Parks. Energies 2012, 5215-5228.

[11] Lee, K., Cha, Y., & Park, J. (1992). Shoert Term Load Forecasting using Artificial Neural Networks. Transactions on Power Systems, 124-132.

[12] López, M., Valero, S., Senabre, C., & Aparicio, J. (2011). A SOM Neural Network Approach to Load Forecasting. Meteorological and Time Frame Influence. Proceedings of the 2011 International Conference on Power Engineering, Energy and Electrical Drives. Torremolinos.

[13] Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., & Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy, 87, 3538-3545.

[14] Ranaweera, D., Hubele, N., & Papalexopoulos, A. (1995). Application of radial basis function neural network model for short-term load forecasting. IEE Proceedings-Generation, Transmission and Distribution, 142, 45-50.

[15] Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. In The top ten algorithms in data mining, (p. 179).