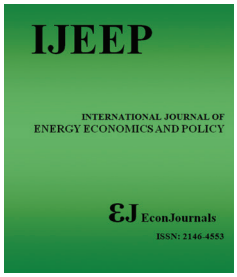


International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2021, 11(3), 215-220.

Data Optimization on the Accuracy of Forecasting Electricity Energy Sales Using Principal Component Analysis Based on Spatial

Iswan Iswan*, Iwa Garniwa, Isti Surjandari

Universitas Indonesia, Depok, Jawa Barat, Indonesia. *Email: iswan@ui.ac.id

Received: 23 November 2020 **Accepted:** 10 February 2021 **DOI:** https://doi.org/10.32479/ijeep.11010

ABSTRACT

It is very important to make forecasts to support future planning. In electricity field, for estimating the demand for electrical energy, there are several influential factors to be considered, e.g. economic growth, increased demand for electricity, and the capacity of power and electrical energy providers. The limited availability of data and variables causes the predictions made to be inaccurate. This paper focuses on the accuracy of forecasting with various numbers of variables to optimize the data held. The initial stage of this research is the division of clusters using the hierarchical clustering method to divide 24 administrative regions into 6 clusters, and to increase the accuracy of forecasting using principal component regression. Based on the results obtained, it can be seen that the MAPE values vary in each cluster. The use of 7 variables in forecasting, in general, shows better accuracy than the use of 6 or 5 variables. However, the difference between the number of these variables is narrow. In cluster 6, the MAPE value in 7 variables is 0.88% while in 5 variables the MAPE value is 0.91%. In cluster 1 and cluster 4, the use of 5 variables has a better value than the use of other variables. Thus, this model can be used and developed to do forecasting even though it uses limited data and variables.

Keyword: Spatial Forecasting, Clustering, Principal Component Analysis
JEL Clafissifications: C210, C25, C380

1. INTRODUCTION

It is very important to make forecasts in order to plan for the future. In the electricity filed, such estimate is the initial part of a long series of planning, both in the distribution system and in the power transmission system. Load estimation in each area is needed to increase the efficiency of the system. In addition to load estimation, efficiency is also influenced by technical characteristics and different economic calculations in each region (Grigoraş et al., 2012).

In estimating the demand for electrical energy, many factors can influence, including economic growth, increased use of electricity in household appliances and business facilities, and growth in generating capacity in an electricity system. On the consumer side, some of the conditions that influence needs (Seppälä, 1996) are consumer behavior (including type of consumer, size of house or building), time, weather, and past load requirements along with the shape of the curve. All of these factors and conditions in a region will be different from other regions.

The existence of different types of consumers makes demand also varies in terms of the period of use. For household consumers there is an increase in demand in the afternoon and evening, while for business consumers the increase in demand occurs mostly in the early morning to late evening. Thus, the supply of electrical energy must be continuously implemented to serve the needs of every consumer. Thus, future energy use patterns can be determined based on existing historical data.

Many studies link the level of energy use with economic growth patterns within a region or in a broad scope, for example a country. There is a causal relationship caused by energy consumption and economic growth (Mohamed and Bodger, 2005; Parajuli et al., 2014). Therefore, it is necessary to conduct an assessment of future energy consumption which is closely related to economic growth in a process of planning and development policies in the right direction. In addition, it is necessary to consider the diversification of the use of electrical energy in a certain shape and size in a system.

In several countries that consist of islands with low economic levels, a method is needed which can assist the implementation of development effectively. The capabilities of each region vary, of course, as well as the unequal distribution and population. Thus, the need for electrical energy will also differ according to progress.

The purpose of this paper is to predict the amount of energy sales in South Sulawesi Province, Indonesia by taking a spatial approach through the creation of clusters in this region. Limited data and variables used cause inaccurate predictions. To overcome this, predictions are made using a number of different variables.

2. LITERATURE REVIEW

The development of a spatial method in forecasting was first carried out by (Willis and Northcote-Green, 1983) by developing predictions in the distribution area of the electricity system. The basis of this spatial method is geographic area so that it will be clearly seen where the need will occur. Although this method is more widely used in distribution systems in the electricity area, it turns out that this method can also be developed at a larger level in the transmission system (Sasmono et al., 2013; 2015) for industrial area development.

The use of fuzzy logic method in this spatial method is carried out to combine information in spatial load forecasting (Chow and Tram, 1996). In addition, fuzzy inference is also used to capture factors that influence growth patterns and mapping (Miranda and Monteiro, 2000).

The development of spatial forecasting in multi-agent systems places independent agents in a system. Each agent will be placed in a sub-zone, so that the inter-subzone is considered as the probability of data in the simulation (Melo et al., 2011; 2012). The application of multi-agents to the dynamics of urban areas resolves the different relationships in each zone. New loads in certain zones can change growth in other zones (Melo et al., 2012).

In the study (Bai et al., 2008), authors use a feeder system to get spatial forecasts so that this is easier and more practical. However, with this system, the forecast period is short term. The grid-based model is used to help improve the inference of the results of entering data into the cluster. This model provides high homogeneity in one cell (cluster) and high heterogeneity among others (Melo et al., 2014). Then the clustering system developed to obtain spatial forecasting. The clustering algorithm is one of the tools in data analysis, including using a partition model, a hierarchical model, a density model and a grid model. The use of clustering has developed in electric power systems, for example in power plants, energy mapping, and substations (Miranda et al., 2016; Shang and Wang, 2015; Tyaglov et al., 2019).

Economic variables in the forecasting process have long been used, as was done (Fullerton et al., 2015; Mohamed and Bodger, 2005) in their research. The level of availability of economic and demographic data is easier to find and is available in large quantities.

There are various methods currently being developed in forecasting, one of which is using principal component analysis (PCA). This method is then combined with other methods in order to obtain better results.

3. METHODS

3.1. Hierarchical Clustering

Cluster is a grouping of something into groups that have similarity. High similarity in a group can be said to be a good cluster. In general, the clustering method is divided into non-hierarchical clustering and hierarchical clustering. Clustering hierarchical itself is a sequence of division of members in a cluster. In each of these divisions there will be a merger in each stage in a certain order. This method is known as the agglomerative algorithm, which at the end of this division will contain all previously existing objects.

In this agglomerative algorithm, there are several linkages, including single linkage, complete linkage, average linkage, and ward linkage (Everitt et al., 2011). This study uses ward linkage which is a combination of the three previous linkages.

3.2. Principal Component Regression

One of the problems that arise in multiple linear regression is the multicollinearity. One way to overcome this is to use principal component regression (PCR). This PCR will replace the predicted variables with variables that have no correlation so that multicollinearity will not occur. In addition, the model will be simpler. If all PCs are included in the regression, then the model is the same as the least squares, the multicollinearity has not been lost because of the large variance. The use of Principal Component Regression will produce a more stable estimate than using direct calculations (Jun-long et al., 2015).

Based on the simple regression model, in equation (1) that is

$$Y=\alpha+\beta x+\epsilon \quad (1)$$

where Y is the vector of many observations on the dependent variable, x is the matrix $(n\times p)$ in the th elements (i, j), β is the vector of the regression coefficient p and ε is the error.

PCR assumes the standardized value of the prediction variable, namely $X'X$, which is proportional to the correlation matrix on the predictor variable. A similar reduction is possible if the predictor variables are in nonstandard form. The PC value for each observation in equation (2),

$$Z=X\,A \tag{2}$$

where Z is the value of the k-th PC in the i-th observation, A is the matrix (p x p) in the th column, k is the $X'X$ eigenvector. Since A is orthogonal, then $X\,\beta$ can be written $X\,A\,A\,'\beta = Z\gamma$, where $\gamma = A\,\beta$. So that the equation becomes

$$Y=Z\,\gamma+\epsilon \tag{3}$$

By only replacing predictor variable with PC the regression model of the PC become a reduced model

$$Y=Z_m\,\gamma_m+\epsilon_m \tag{4}$$

Where γ_m is vector of element m which is part of γ. Z_m is a matrix (n × m) from that according to part Z, and ε_m is error.

The complete steps for the Principal Component Regression are as follows:

1. Get the multiple regression equation according to the available data and also some of the variables that will be used. Check the value on the analysis of variance, although R^2 is high but sometimes only one variable is significant
2. Check the VIF value, if the value is more than 10, then it can be assumed that multicollinearity occurs in the equation obtained. For that, do PCA calculations.
3. Find the mean and standard deviation of the independent variables. Then do the standardization of data so that new variables will be obtained that represent the entire data

$$Z_n = \frac{X_n - \bar{X}_n}{St.\,Deviation} \tag{5}$$

4. Get the eigenvalue of the independent variable used. Eigenvalue which has a value of > 1, then this factor will be used. Remove all the factors in the equation by storing the PC values in the coefficients that will later be used in modeling
5. If there are two or more factors whose value is more than 1, then the number of values that is greater will be used
6. Create a regression equation by replacing all variables with W_n. Double check the VIF value and the values contained in the analysis of variance
7. Change the results of the factor modeling to the original using the PC values that have been previously obtained. $PC_n = a_1 z_1 + \ldots + a_n z_n$
8. Perform the transformation again using $W_n = PC_n$ so that a new equation will be obtained. Furthermore, the reverse standardization transformation uses standardization equations so that the equation will be obtained in accordance with the number of variables used previously.

To measure how much error is obtained between the actual value vs the predicted value, the Mean Absolute Percentage Error (MAPE) method is used, which shows how much error the forecast is compared to the actual value of the data.

$$MAPE = \frac{1}{N}\sum_{t=1}^{n}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t} \tag{6}$$

4. RESULTS AND DISCUSSION

The case study data in this study were taken from the province of South Sulawesi, with a total of 24 districts/cities. The existing electricity system currently exists is the South Sulawesi system.

The stages of cluster division in this study were taken from previous research (Iswan and Garniwa, 2017).

This section uses the hierarchical clustering method. The initial number of clusters is 24 clusters taken from the total areas in this province. The result is a division of regions into 6 clusters with details: Cluster 1 totaling 11 regions (Sinjai, Barru, Luwu, Enrekang, Bantaeng, Selayar, Soppeng, North Luwu, East Luwu, Tana Toraja, and North Toraja), Cluster 2 consisting of 8 regions (Jeneponto, Takalar, Pare-pare, Wajo, Sidrap, Palopo, Bulukumba, and Pinrang), Cluster 3 with 2 regions (Gowa and Maros), Cluster 4 consisting of 1 region (Pangkep), Cluster 5 consisting of 1 region (Bone), and Cluster 6 consists of 1 region (Makassar). Cluster 6 namely Makassar is the provincial capital. The cluster distribution map (different colors for each cluster) in South Sulawesi province can be seen in Figure 1.

After dividing the cluster into 6 clusters, the next step is to forecast for each cluster using a different number of variables. The dependent variable is Y while x is the independent variable. For independent variables, the number and types differ depending on the number of variables used, x_1 - x_7 for 7 variables (GRDP, industrial GRDP, commercial GRDP, construction GRDP, industrial energy, commercial energy, and household energy. At x_1 - x_6 for 6 variables There are 2 types, namely 6A (GRDP, agricultural GRDP, commercial GRDP, construction GRDP, commercial energy, household energy) and 6B (GRDP, industrial GRDP, commercial GRDP, construction GRDP, commercial energy, household energy). Then x_1 - x_5 for 5 variables there are 3 types, namely 5A (GRDP, agricultural GRDP, industrial GRDP, commercial GRDP, construction GRDP), 5B (GRDP, commercial GRDP, construction GRDP, commercial energy, household energy), and 5C (GRDP, industrial GRDP, GRDP construction, industrial energy, household energy).

The results obtained by using linear regression in each cluster with different variables, then the equation has multicollinearity in the equation. This is indicated by the VIF value that exceeds the number 10. Therefore, the principal component regression technique is used to overcome the occurrence of multicollinearity.

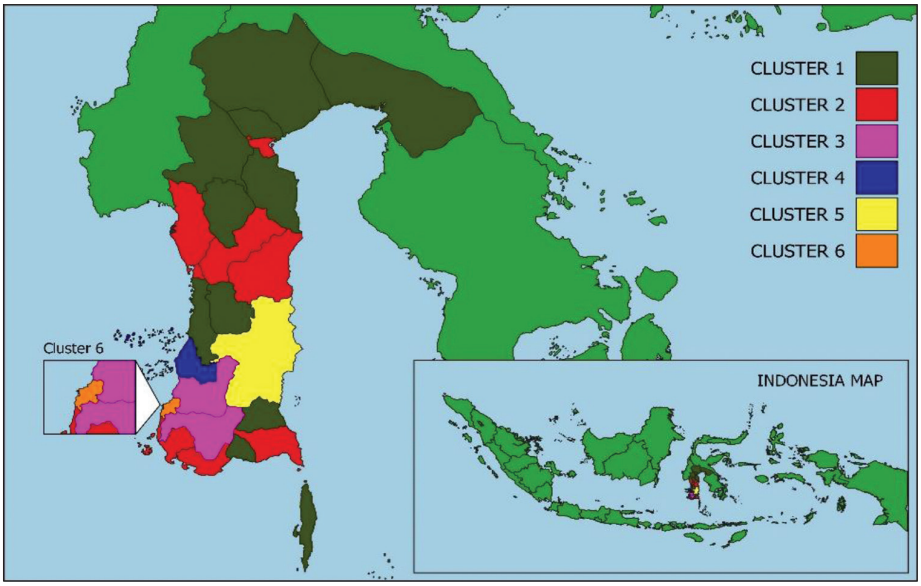
Based on the results of the calculation of the eigenvalue for each cluster, it is known that the eigenvalue that fulfills is only in the first principal component. Then the coefficients P1 are sufficient to meet the requirements to represent all available variables. Table 1 shows the eigenvalue on P1 in each cluster.

The initial regression equation by replacing the initial variable with a score of W1. The score of W1 was then transformed back using the P1 coefficients. Then the multiple regression equation will be retrieved that is in accordance with the initial number of variables. This multiple regression equation is then used to calculate forecasting in each cluster to obtain a predictive value.

The comparison of energy sales between the actual value and the forecasted value between 2006-2016 in each cluster can be seen in Figure 2.

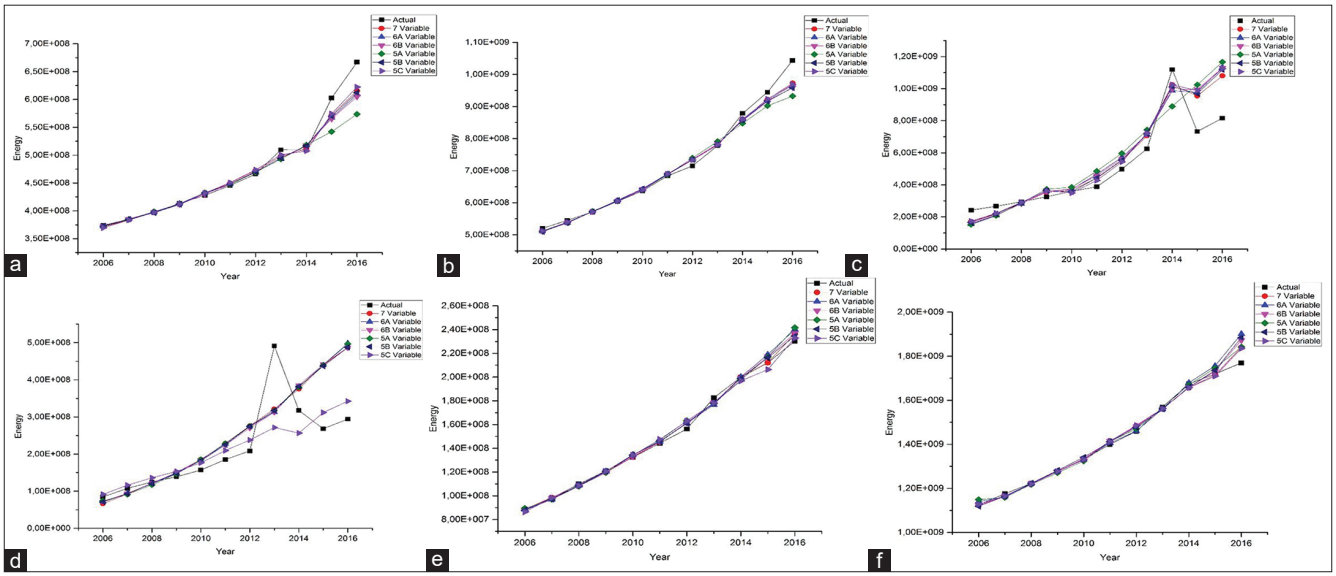
The MAPE value in each cluster is different, and this is due to the type and number of variables used (Table 2). For variables that use only the GRDP value, namely the 5C variable, some clusters

Figure 1: Map of clusters division



Source: Proceed author with QGIS

Figure 2: Comparison of actual versus prediction data on different number variable. (a) Cluster 1, (b) Cluster 2, (c) Cluster 3, (d) Cluster 4, (e) Cluster 5, (f) Cluster 6



Source: Author data proceed

Table 1: First Eigenvalue of the different variable

Cluster	7-Variable	6-Variable		5-Variable		
		6A	6B	5A	5B	5C
Cluster 1	6.8108	5.9412	5.9503	4.9516	4.9582	4.8453
Cluster 2	6.9331	5.9510	5.9498	4.9789	4.9553	4.9461
Cluster 3	6.4388	5.4828	5.5164	4.7148	4.7135	4.5862
Cluster 4	6.7936	5.9305	5.9222	4.9525	4.9377	4.8482
Cluster 5	6.8662	5.9019	5.8711	4.9237	4.8820	4.9342
Cluster 6	6.9390	5.8083	5.9414	4.8055	4.9591	4.9626

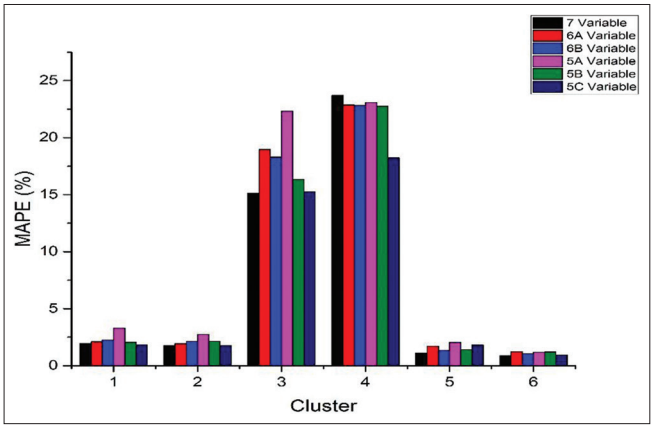
Source: Proceed author with Minitab

Table 2: Comparison of MAPE energy sold on various variable amounts

Cluster	7-Variable	6-Variable		5-Variable		
		6A	6B	5A	5B	5C
Cluster 1	1.9438144	2.0976666	2.2388992	3.3012454	2.0495275	1.8231719
Cluster 2	1.7544563	1.9303464	2.1207565	2.7332047	2.1386008	1.7759817
Cluster 3	15.1312687	18.9543458	18.2815007	22.3209236	16.3401194	15.2316420
Cluster 4	23.7010665	22.8558932	22.8116884	23.0656655	22.7580302	18.2190745
Cluster 5	1.1227773	1.7138582	1.3492147	2.0312074	1.3927709	1.8060542
Cluster 6	0.8859790	1.2083512	1.0373752	1.1573783	1.1871259	0.9145046

Source: Proceed author with Microsoft Excel

Figure 3: Comparison of MAPE for all cluster



Source: Author data proceed with OriginLab

have low MAPE values compared to clusters with the number of other variables, namely cluster 1 and cluster 4. Other clusters with the lowest MAPE value are Cluster 2, Cluster 3, Cluster 5, and cluster. 6. In Cluster 2, the difference in MAPE values between the 7 variables and the 5C variable values is not far apart, which is only 0.02%. In Cluster 6, the difference in MAPE values between 7 variables and 5C variables is 0.1%.

Especially for cluster 3 and cluster 4, the MAPE value obtained exceeds 10%. The MAPE in Cluster 3 ranged from 15.1% to 22.3%, while in cluster 4 the MAPE value was in the range of 18.2% to 23.7%. This is because the value of energy sales in the two clusters has increased sharply and then there has been a sharp decline thereafter. However, if traced annually, the difference between the actual value vs the predicted value does not differ much. This paper uses linear regression, so the results also tend to look for linearity in the predictions. This results in a large MAPE value in the accumulation over the 10 years.

By comparing the total number of variables to the resulting MAPE value, almost all clusters have a small MAPE value when using 7 variables (Figure 3). However, when compared to using fewer variables, the resulting difference is very small. Thus, the large number of variables is not a guarantee the predictive value will be better.

Limited data on an area can be optimized using this method. Merging in the form of clusters will make forecasting simpler and more efficient. Therefore, the government and stakeholders will find it easier to plan especially energy development, especially electrical energy in an administrative area.

5. CONCLUSION

The method developed is simpler and easier to use. If a region has many administrative areas, it is not necessary that all of these administrative areas have to be forecasted. Just need to form them in clusters. This cluster system is simpler and more effective because the cluster is a combination of administrative regions that have similarity. The principal component analysis method combined with the regression model provides a small difference between the predicted vs actual obtained. The error value for each forecast is better, even though in the two clusters the error is more than 10% due to a drastic spike in energy sales in a certain year. Data and variable limitations in implementing forecasting can still be done with very good results. This can be indicated by the very small error difference between those using 7 variables and 5 variables. Certain types of variables do not apply equally to every cluster.

Forecasting energy sales using the Principal Component Analysis method by combining areas in a cluster will make this forecasting simpler and easier to implement. Likewise, the level of forecast accuracy obtained by the Principal Component Regression in each cluster is also very good. Limited data and variables used can be optimized by this developed method. The few variables used can get the maximum forecast. Thus, this model can be used and can be further developed. In the territory of an archipelago, this model is very suitable because of the large number of small areas and is separated from the larger islands.

REFERENCES

Bai, X., Mu, G., Li, P. (2008), A Method of Spatial Load Forecasting Based on Feeder. Paper Presented at the 2008 3rd International Conference on Electric Utility Deregulation and Restructuring and Power Technologies.

Chow, M., Hahn, T. (1996), Application of Fuzzy Logic Technology for Spatial Load Forecasting. Paper Presented at the Proceedings of 1996 Transmission and Distribution Conference and Exposition.

Everitt, B.S., Sabine, L., Morven, L., Daniel, S. (2011), Cluster Analysis. 5th ed. Hoboken, New Jersey: John Wiley.

Fullerton, T.M., George, N., David, T., Adam, G.W. (2015), Metropolitan econometric electric utility forecast accuracy. International Journal of Energy Economics and Policy, 5(3), R15.

Grigoraş, G., Florina, S., Gheorghe, C. (2012), Load Estimation for Distribution Systems Using Clustering Techniques. Paper Presented at the 2012 13th International Conference on Optimization of Electrical and Electronic Equipment.

Iswan, I., Garniwa, I. (2017), Principal Component Analysis and Cluster

Analysis for Development of Electrical System. Paper Presented at the 2017 15th International Conference on Quality in Research (QiR): International Symposium on Electrical and Computer Engineering.

Jun-long, F., Xing, Y., Fu, Y., Xu, Y., Liu, G.L. (2015), Rural power system load forecast based on principal component analysis. Journal of Northeast Agricultural University, 22(2), 67-72.

Melo, J.D., Carreno, E.M., Calvino, A., Antonio, P.F. (2014), Determining spatial resolution in spatial load forecasting using a grid-based model. Electric Power Systems Research, 111, 177-184.

Melo, J.D., Carreno, E.M., Padilha-Feltrin, A. (2011), Multi-agent Framework for Spatial Load Forecasting. Paper Presented at the 2011 IEEE Power and Energy Society General Meeting.

Melo, J.D., Carreno, E.M., Padilha-Feltrin, A. (2012), Considering Urban Dynamics in Spatial Electric Load Forecasting. Paper Presented at the 2012 IEEE Power and Energy Society General Meeting.

Melo, J.D., Carreno, E.M., Padilha-Feltrin, A. (2012), Multi-agent simulation of urban social dynamics for spatial load forecasting. IEEE Transactions on Power Systems, 27(4), 1870-1878.

Miranda, F.J., de Carvalho Filho, J.M., Paiva, A.P., de Souza, P.V.G., Samuel, T. (2016), A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context. Electric Power Systems Research, 136, 31-42.

Miranda, V., Cláudio, M. (2000), Fuzzy Inference in Spatial Load Forecasting. Paper Presented at the 2000 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 00CH37077).

Mohamed, Z., Bodger, P. (2005), Forecasting electricity consumption in New Zealand using economic and demographic variables. Energy, 30(10), 1833-1843.

Parajuli, R., Østergaard, P.A., Dalgaard, T., Pokharel, G.R. (2014), Energy consumption projection of Nepal: An econometric approach. Renewable Energy, 63, 432-444.

Sasmono, S., Sinisuka, N.I., Atmopawiro, M.W., Darwanto, D. (2013), Macro Demand Spatial Approach (MDSA) with Principal Component Analysis (PCA) on Spatial Demand Forecasting for Industrial Area in Transmission Planning. Paper Presented at the 2013 International Conference on Information Technology and Electrical Engineering.

Sasmono, S., Sinisuka, N.I., Atmopawiro, M.W., Darwanto, D. (2015), Macro demand spatial approach (MDSA) at spatial demand forecasting for transmission system planning. International Journal on Electrical Engineering and Informatics, 7(2), 193.

Seppälä, A. (1996), Load Research and Load Estimation in Electricity Distribution: Technical Research Centre of Finland. Finland: VTT Technical Research Centre of Finland.

Shang, L., Shoupeng, W. (2015), Application of the Principal Component Analysis and Cluster Analysis in Comprehensive Evaluation of Thermal Power Units. Paper Presented at the 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies.

Tyaglov, S.G., Sheveleva, A.V., Shurukhina, T.V., Guseva, T.B. (2019), Model for forming the interregional cluster of the alternative energy. International Journal of Energy Economics and Policy, 9(3), 373.

Willis, H.L., Northcote-Green, J.E.D. (1983), Spatial electric load forecasting: A tutorial review. Proceedings of the IEEE, 71(2), 232-253.