Principal Component Analysis and Cluster Analysis in Profile of Electrical System

Iswan*, I Garniwa
Electrical Engineering Department, Universitas Indonesia

*is.one.wan@gmail.com

Abstract. This paper propose to present approach for profile of electrical system, presented approach is combination algorithm, namely principal component analysis (PCA) and cluster analysis. Based on relevant data of gross domestic regional product and electric power and energy use. This profile is set up to show the condition of electrical system of the region, that will be used as a policy in the electrical system of spatial development in the future. This paper consider 24 region in South Sulawesi province as profile center points and use principal component analysis (PCA) to asses the regional profile for development. Cluster analysis is used to group these region into few cluster according to the new variable be produced PCA. The general planning of electrical system of South Sulawesi province can provide support for policy making of electrical system development. The future research can be added several variable into existing variable.

1. Introduction
Economic growth will make increasing use of electricity so that the needed good operation and maintainance as well as the expansion of the system. Especially for expansion of the system, the results must consider the time scenario and maintain quality of service and system operating costs [1]. Planning system should be notice to the existing system, load and price forecasting as well as technical and economic considerations, including arrange the optimal size, place and season [2].

Electricity demand increased for all sectors, especially urban and industrial region can change all patterns of the electric load demand for a certain period. The use of electric greatly fluctuate depending on the season, weather, behavior and time. The other side, the electric load and the quality of the environment is also changing demand, temperature and generation which also dynamic with daily to seasonal timescales [3].

Method used to get the load profile is cluster analysis method. This profile can be use to categorize the region based on load and condition of the economy, so it can be used for spatial load forecasting [4]. To obtain more accurate results again, used principal component analysis (PCA). Both of these methods will create a profile with the region of every cluster a different.

This paper proposed methods of PCA and cluster analysis to see in general and classifies the level of development of the electric system in the South Sulawesi province. Later can be used as references for planning and development through a pattern of spatial load forecasting [5].
2. Methods

2.1. Principal Component Analysis

Follow the processes carried out [6], the process of calculating the Principal Component Analysis as follows,

- Collect Data. Assumption that the observed \( n \), while \( p \) is an indicator of the \( i \) observed object was \( x_{1i}, x_{2i}, ..., x_{pi} \). So the formulation can be seen in the form of a matrix as follows:

\[
X = \begin{bmatrix}
    x_{11} & x_{12} & ... & x_{1p} \\
    x_{21} & x_{22} & ... & x_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & ... & x_{np}
\end{bmatrix}
\] (1)

Variable is \( p \), while \( n \) is the number of the object.

- Normalize the data obtained by following equation

\[
z_{ij} = x_{ij} - \bar{x}_k/S_k; \quad i=1, 2, ..., n, \quad k=1, 2, ..., p
\] (2)

Where \( \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \) and \( S_k = \sqrt{1/n - \sum_{i=1}^{n} (x_{ij} - \bar{x}_k)^2} \)

The variance of variable was 1 and mean value was 0 after normalization.

- Determine the correlation coefficient matrix by following equation

\[
R = \begin{bmatrix}
    r_{11} & r_{12} & ... & r_{1p} \\
    r_{21} & r_{22} & ... & r_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{p1} & r_{p2} & ... & r_{pp}
\end{bmatrix}
\] (3)

Between

\[
r_{ii} = \frac{\sum_{i=1}^{n} x_{ij}^2 x_{ij}^2}{(n-1)}
\] (4)

- Seeked \( m \) non-negative eigenvalues \( \lambda_1, \lambda_2, ..., \lambda_m \) of the characteristic equation \(|R-\lambda I|=0\) and eigenvectors corresponded to the eigenvalue \( \lambda i \) according to correlation matrix \( R \):

\[
A_i = (a_{i1}, a_{i2}, ..., a_{im})^T \quad i=1, 2, ..., m
\]

- Principal components. \( m \) principal components composed of the eigenvectors were as the following:

\[
P_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \quad i=1, 2, ..., m
\]

The main components \( P_1, P_2, ..., P_m \) were unrelated, and their variances were decreasing.

- Select \( m(m<p) \) principal component. If the sum of the first \( m \) principal component \( P_1, P_2, ..., P_m \) variance of the total variance sealed 1. Then selected the first \( m \) principal component \( P_1, P_2, ..., P_m \). The sum of \( m \) principal component variances reached 85% of the total variances meant reserved original indicators or variable \( X_1, X_2, ..., X_p \)'s basic information, thus the number of the indicators or variable was reduced by \( p \) to \( m \), and played a role in screening index or variable.
2.2. Cluster Analysis
Cluster analysis is known as one of the statistical methods to be able to classify the data or variable according to their characteristics [7]. Firstly steps this method is determines the similarity between samples quantitatively, later classified into different groups according to the level or similarity. One part the cluster method is hierarchical method. This method has the main idea is to weigh the data as a separate category, and then combine them into a similar category according to the distance and level. Then combined again with a similar category according to the distance and level, so we will get only one category.

Linkage methods are good are good for clustering item and variable. Single linkage considers the average distance or nearest neighbour, average linkage considers the average distances, complete linkage considers the maximum distance or farthest neighbor [8]. In a hierarchical method, there are several methods including Ward. Ward uses a hierarchical procedure to minimize the loss of information by combining the two groups.

3. Study Case

3.1. Data
There are six variables for which data are taken from the PT. PLN and the BPS of the Republic of Indonesia, in 2014. Indicators in this paper should be build upon the current conditions, the data is easily available and nonoverlaping. Taking 24th region in South Sulawesi Province as research object, the regions evaluation based on principal component analysis and cluster analysis.

Statistical indicators in this paper are devided into demogragic index, economy index and technical index. Demogragic index including number of people (X1). Economy index includes the following three items: gross domestic product of regional (GDP R) (X2), industry GDP regional (X3), and commercial GDP regional (X4). Technical index includes the following 2 items: power connected of costumer (X5) and electricity sales (X6). This article selects the variable as shown in Table 1.

| Region   | X1   | X2   | X3   | X4   | X5   | X6   |
|----------|------|------|------|------|------|------|
| Selayar  | 128,744 | 2,530.65 | 80.52 | 221.13 | 17,034.40 | 22,625.73 |
| Bulukumba| 407,775 | 6,395.64 | 435.47 | 909.42 | 77,391.65 | 107,931.92 |
| Bantaeng | 182,283 | 3,805.21 | 182.07 | 564.14 | 25,398.12 | 38,409.86 |
| Jeneponto | 353,287 | 4,764.30 | 171.59 | 646.06 | 53,141.45 | 71,183.75 |
| Takalar  | 283,767 | 4,517.63 | 279.99 | 618.79 | 48,661.02 | 76,141.37 |
| Gowa    | 709,386 | 9,701.43 | 622.48 | 1,126.01 | 152,513.36 | 232,300.74 |
| Sinjai  | 236,497 | 5,035.78 | 132.06 | 641.72 | 33,464.20 | 47,641.88 |
| Maros   | 335,596 | 10,115.49 | 2,373.24 | 354.44 | 143,586.20 | 393,038.04 |
| Pangkep | 320,293 | 12,391.76 | 652.20 | 667.43 | 114,659.80 | 491,206.51 |
| Barru   | 170,316 | 3,453.22 | 190.83 | 325.13 | 94,293.25 | 46,968.32 |
| Bone    | 738,515 | 14,741.06 | 1,047.23 | 1,883.10 | 135,954.33 | 182,576.47 |
| Soppeng | 225,709 | 4,876.74 | 496.63 | 670.91 | 47,460.06 | 63,854.83 |
| Wajo    | 391,980 | 10,286.59 | 364.21 | 1,542.84 | 71,567.75 | 106,110.06 |
| Sidrap  | 286,610 | 6,104.74 | 867.01 | 656.48 | 68,730.67 | 105,562.16 |
| Pinrang | 364,087 | 8,941.22 | 599.43 | 1,234.55 | 80,938.20 | 115,136.69 |
| Enrekang| 198,194 | 3,385.82 | 257.81 | 345.52 | 37,672.15 | 39,946.70 |
| Luwu    | 347,096 | 6,299.56 | 316.15 | 790.69 | 30,279.70 | 47,470.07 |
| Tator   | 227,588 | 3,193.80 | 216.15 | 569.04 | 63,620.45 | 72,912.64 |
| Lutia   | 299,989 | 5,721.29 | 238.24 | 582.34 | 44,830.85 | 59,264.36 |
| Lutim   | 269,405 | 13,794.38 | 35.64 | 433.15 | 47,964.30 | 70,661.56 |
| Toraja U| 224,003 | 3,507.40 | 217.97 | 752.68 | - | - |
| Makassar| 1,429,242 | 82,592.00 | 16,985.53 | 15,658.92 | 807,669.00 | 1,569,669.4 |
| Pare-pare| 136,903 | 3,608.58 | 77.46 | 586.03 | 49,756.71 | 92,610.07 |
| Palopo  | 164,903 | 3,363.25 | 103.33 | 762.334 | 64,437.75 | 103,268.85 |
3.2. Principal Component
Firstly, calculated correlation coefficient matrix $R$ which exist in data Table 1. After each index data are transformed with the same tendency, the principal component analysis are obtained from the six variable by using the principal component analysis method. And then eigenvalue, variance contribution rate and cumulative variance contribution rate of each principal component are obtained, view show in Table 2.

$$R = \begin{bmatrix}
1 & 0.899 & 0.858 & 0.880 & 0.911 & 0.856 \\
0.899 & 1 & 0.980 & 0.983 & 0.983 & 0.959 \\
0.858 & 0.98 & 1 & 0.985 & 0.984 & 0.959 \\
0.880 & 0.983 & 0.985 & 1 & 0.976 & 0.933 \\
0.911 & 0.983 & 0.984 & 0.976 & 1 & 0.977 \\
0.856 & 0.959 & 0.959 & 0.933 & 0.977 & 1 
\end{bmatrix}$$

According to value in Table 2, the cumulative variance contribution rate of the first main component has been reach 95.193%, which describes 95.193% of authentic index information. In order to make the index of the original information contained in the principal component of a more comprehensive and to prevent the removal of important information. The first principal component can be extracted as a more comprehensive index.

### Table 2. The Eigenvalue and Variance Contribution Rate

| Principle Component | Total Eigenvalue | Variance Contribution Rate % | Cumulative Contribution Rate % |
|---------------------|-----------------|-----------------------------|-------------------------------|
| P1                  | 5.712           | 95.193                      | 95.193                        |
| P2                  | 0.185           | 3.087                       | 98.280                        |
| P3                  | 0.073           | 1.216                       | 99.496                        |
| P4                  | 0.018           | 0.301                       | 99.797                        |
| P5                  | 0.008           | 0.127                       | 99.924                        |
| P6                  | 0.005           | 0.076                       | 100.000                       |

Therefore, principal component analysis was instrument in dimensional reduction and simplification of the problem. Based on the calculations of the factors that big influence on this profile is a gross domestic product of regional (GDP R) (X2), industry GDP regional (X3), and comercial GDP regional (X4) and power connected of costumer (X5). Followed an electricity sales (X6), but number of people (X1) factor did not affect entire calculation.

3.3. Cluster Analysis
The first principal component that are obtained by using the principal component analysis methods are used as the cluster variable. And the first principal component score instead of the original data were used to do system cluster analysis. Then dendritic diagram is obtained by using the ward methods, as shown in Figure 1 the classification of each sample can be seen visually from Figure 1.

From Figure 1, the classification of each region can be visually. In this paper, using 24 region are devided into 4 cluster, as follow (1) Sinjai, Barru, Luwu, Enrekang, Bantaeng, Selayar, Soppeng, Luwu Utara, Jeneponto, Takalar, Luwu Timur, Tana Toraja, Pare-par, Wajo, Sidrap, Palopo, Bulukumba, Pinrang, Toraja Utara. (2) Gowa, Bone. (3) Maros, Pangkep. (4) Makassar
3.4. Result
Based on the classification using cluster analysis, there are four clusters grouped by similarity data owned by each region. Cluster 4 is the cluster that have no resemblance with other regions, namely Makassar as its province capital. Maros and Pangkep is an area with heavy industries to be around Makassar. Gowa is the buffer zone, whereas the Bone are transportation lines connecting other areas outside of South Sulawesi province (in west region). While other areas are regions that extends from north to south, this area commonly are farming area and mining area.

By grouping of these regions, it will note areas that needs to prioritized in the development of electrical systems, including those for spatial load forecasting.

![Cluster Dendrogram](image)

**Figure 1.** The Cluster Dendrogram

4. Conclusion
In this paper, the principal component analysis and cluster analysis were combined and applied in profile of electric system that can be used for the purposes development of network through a of spatial load forecasting. Using the principal component analysis to extract the main factor of the region, and the comprehensive evaluation index system was built. Then the principal component were extracted as new data matrix for clustering, which avoid selecting cluster variable subjectively. Later need an additional variables to be able to see the overall factors that influence the electrical system profile of a region's.
References
[1] Vieira D a. G, Cabral MM a., Menezes TV, Silva BE, Lisboa a. C. Measuring the Spatial Error in Load Forecasting for Electrical Distribution Planning as a Problem of Transporting the Surplus to the In-Deficit Locations. 2012 11th Int Conf Mach Learn Appl [Internet]. Ieee; 2012 Dec [cited 2014 Nov 11];475–80. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6406781
[2] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy [Internet]. Elsevier Ltd; 2014;135:461–71. Available from: http://dx.doi.org/10.1016/j.apenergy.2014.08.111
[3] Shu J, Wu L, Li Z, Shahidehpour M, Zhang L, Han B. A new method for spatial power network planning in complicated environments. IEEE Trans Power Syst. 2012;27(1):381–9.
[4] Akperi B, Matthews P. Analysis of Clustering Techniques on Load Profiles for Electrical Distribution. 2014;(Powercon):20–2.
[5] Bai X, Peng-wei G, Gang M, Gan-gui Y, Ping L, Hong-wei C, et al. A Spatial Load Forecasting Method Based on the Theory of Clustering Analysis. Phys Procedia [Internet]. Elsevier B.V.; 2012 Jan [cited 2014 Oct 5];24:176–83. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1875389212000703
[6] Jun-long F, Yu X, Yu F, Yang X, Guo-liang L. Rural Power System Load Forecast Based on Principal Component Analysis. J Northeast Agric Univ (English Ed [Internet]. 2015;22(2):67–72. Available from: http://www.sciencedirect.com/science/article/pii/S1006810415300349
[7] Shang, Liqun; Wang S. Application of the Principal Componen Analysis and Cluster Analysis in Comprehensive Evaluation of Thermal Power Unit. 2015;
[8] Filho JM, De Carvalho Filho JM, Paiva AP, De Souza PVG, Tomasin S. A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context. Electr Power Syst Res [Internet]. Elsevier B.V.; 2016;136:31–42. Available from: http://dx.doi.org/10.1016/j.epsr.2016.02.012