Segmentation of Natural Gas Customers in Industrial Sector Using Self-Organizing Map (SOM) Method

To cite this article: A M Masbar Rus et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 316 012046

View the article online for updates and enhancements.

You may also like

- Highly selective CD44-specific gold nanorods for photothermal ablation of tumorigenic subpopulations generated in MCF7 mammospheres
  Eugene Lee, Yoochan Hong, Jihye Choi et al.

- Dynamic behavior of dual cross-linked nanoparticle networks under oscillatory shear
  Balaji V S Iyer, Victor V Yashin and Anna C Balazs

- Monte Carlo simulation of an upgraded PGNAA shield at TRR–1/M1
  S Sangaroon, W Ratanatongchai, R Picha et al.
Segmentation of Natural Gas Customers in Industrial Sector Using Self-Organizing Map (SOM) Method

A M Masbar Rus 1, R Pramudita 1, I Surjandari 1

1 Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, 16424, Depok, Indonesia.

annisamarlin@ui.ac.id, redianipramudita@gmail.com, isti@ie.ui.ac.id

Abstract. The usage of the natural gas which is non-renewable energy, needs to be more efficient. Therefore, customer segmentation becomes necessary to set up a marketing strategy to be right on target or to determine an appropriate fee. This research was conducted at PT PGN using one of data mining method, i.e. Self-Organizing Map (SOM). The clustering process is based on the characteristic of its customers as a reference to create the customer segmentation of natural gas customers. The input variables of this research are variable of area, type of customer, the industrial sector, the average usage, standard deviation of the usage, and the total deviation. As a result, 37 cluster and 9 segment from 838 customer data are formed. These 9 segments then employed to illustrate the general characteristic of the natural gas customer of PT PGN.

1. Introduction

Nowadays the limitations of Indonesia’s petroleum could not meet the demands of the country’s needs, thus Indonesia need to find other applicable alternatives to substitute the petroleum. One of the alternatives which is considered to be the most suitable to bolster the country's energy needs is natural gas. In fact, Indonesia has abundant gas reserves to shift the energy needs from petroleum which the reserve is predicted to fulfil the demand until 41.2 years ahead [1].

Along with the transition of energy from oil to natural gas, gas consumption in Indonesia becomes increasing. This phenomenon could be seen in Figure 1., which depicts the significant rise of natural gas production and consumption in Indonesia [1].

The increase in the consumption of natural gas in Indonesia is influenced by some sector of gas customers. Based on Ministerial Regulation of Energy and Mineral Resources (ESDM) No. 03 of 2010 article 6, there is priority in gas usage for sector; a) oil gas lifting & b) fertilizer industry, c) the provision of electric power and another industry. These 4 sectors are natural gas customers that have been classified by the Government.
Natural gas which is used in Indonesia, commonly referred to as domestic gas, has already allocated towards each sector [2]. Based on Indonesia’s natural gas balance 2014, Indonesia have allocated 59.45% of the total production of natural gas for the domestic needs. The detail allocation is shown in figure 2.

Figure 1. Graphic of Natural Gas Consumption and Production in Indonesia [1]

Figure 2. Domestic Gas Allocation [3]

Natural Gas is considered as more affordable in terms of price and have a lower level of pollution rather than other alternatives [4]. For this reason, industrial sector consumes more gas than other sectors as it is depicted in figure 2, thus industrial sector play an important role to the development of natural gas in Indonesia. Therefore, focus need to be made to the development of gas usage in industrial sector. Since natural gas is a non-renewable energy, thus the development is suggested to be supported with a robust and efficient plan. In order to do this, observation is made to the condition of the gas usage in Indonesia.
Figure 3. shows a graphic of one of customer’s natural gas usage in industrial sector of PT PGN. The graphic shows the customer usage per hour and its maximum and minimum limits which determined according to the customer’s contract. Based on the graphic, the customer’s usage which exceed the maximum and minimum limit occurs very often. This condition happened on almost 80% of gas usage patterns sample in the industrial sector which influence the normal gas distribution. However, regardless of these deviation behavior, instead of penalizing the deviate customer, the producer still applying a same rate for all customers even though some of them exceed the maximum and minimum limit. This irregularity shows urgency in streamline natural gas usage for the future benefit.

One of the preliminary measure which could be done to manage natural gas usage effectively is by improving current segmentation process of gas customers in industrial sectors. This segmentation improvement could help the company to design an appropriate marketing mix for segment, thus the marketing strategies could be directed to the desired target [6].

In a nutshell, to address the challenge of enhancing natural gas usage efficiency, segmentation of gas users is proposed as preliminary measure. This research will employ Self-Organizing Map (SOM) for improving the segmentation process and employ gas usage data from customer of PT. PGN to represent the condition of gas usage in Indonesia.

2. Methodology

In this research, Self-Organizing Map (SOM) method will be used to conduct the segmentation by considering the homogeneity of the input data which has been determined for every sample. The sample which will be observed are the gas customers of a stated-owned natural gas enterprise, PT. Perusahaan Gas Negara (PGN) (Persero) Tbk, who is the biggest gas distributor in Indonesia.

The data collected is secondary data, i.e. real-time data of gas usage obtained from the Gas Management System (GMS) of PT PGN. Moreover, the information covered in this data is the identity of the customer and the results of temperature, pressure, and volume of usage on each of customer’s gauge measurement. Furthermore, real-time data used in this research is the usage data per hour of 1172 gas customers in the industrial sector in Sub-Business Unit (SBU) I, encompass the southern part of Sumatra area and the western part of Java, within January to December 2014.

2.1. Data Preprocessing

In data pre-processing process, the previous real-time data will be pre-processed to produce variable data set as shown in Table 1. Later, this data will be used as attributes in the process of data processing.

| Table 1. Variables and Values |
| Variable               | Values                                                                 |
|------------------------|------------------------------------------------------------------------|
| Area                   | Jakarta, Bekasi, Bogor, Cilegon, Cirebon, Karawang, Tangerang, Palembang |
| Type of Customers      | IMP 1, IMP 2, IMP 3, IJK 1, IJK 2, IJK 3                                |
| Industrial Sector      | Glass, Wood, Ceramic, Paper, Chemical, Basic Metal, Food, Fabricated Metal, Cement, Textile, Textile, Power Plant, Others |
| Average Usage          | Gas usage standard volume per time unit (Nm$^3$)                        |
| Standard Deviation of Usage | The standard deviation of the gas usage standard volume                  |
| Total Deviations of Usage | The total volume of deviate usage from maximum and minimum limits (Nm$^3$) |

After that, data preprocessing is conducted by checking the data completion and the calculation of the average value, standard deviation, as well as the total deviation of gas usage as shown in table 1. Checking the completeness of the data is conducted by looking at whether each object is included in the whole database from January-December 2014. As a result, from the initial data set of 1172, there are 838 data set which is concluded as complete. Later, the average, standard deviation, and the total usage deviation of 838 data set is calculated using Microsoft Excel.
2.2. Data processing
In this process, the variable inputs which is used encompass variable area, type of customer, the industrial sector, the total deviation of use, average (per hour / day / week / month) and standard deviation (units of hours / days / weeks / months). The STATISTICA software is employed for the calculation process.

At the beginning of the process, initialization of number of clusters, the coefficient learning rate ($\eta$) and neighbourhood coefficient ($\sigma$) is determined. The initial determination of the coefficient does not have any specific rules, thus the coefficients which is used in this study is a default number from STATISTICA software, namely $\eta = 0.1$ and $\sigma = 3$. As for the determination of the number of clusters, experiments are conducted for different number of clusters until all certain number of clusters are occupied by at least 1 type of customer and has the lowest error validity.

3. Results and Discussion
In the previous section, Self-Organizing Map (SOM) has been employed to produce clusters based on their homogeneity. In this research, the result of the cluster is differentiated based on time unit as shown in table 2.

| Time unit | Training error | Test error | Validation error |
|-----------|----------------|------------|------------------|
| Hour      | 0.515248       | 0.566476   | 0.540686         |
| Day       | 0.486721       | 0.622279   | 0.537399         |
| Week      | 0.514967       | 0.615034   | 0.537586         |
| Month     | 0.729907       | 0.829682   | 0.762875         |

The table shows that grouping based on day unit demonstrating better results from other measure. This conclusion can be inferred from the error validation value that reaches 0.537399, which is lower than other group time. The error shows the distance between objects in the cluster and the distance between each cluster. The smaller the error then the closer the distance between objects in the cluster, and the further the distance between its clusters. The cluster result based on day-time unit is depicted in figure 4.

![Figure 4](image_url)

**Figure 4.** Frequency graphic of daily clustering results.

Objects of standard deviation on each cluster will interpret the usage distribution, which will be compared to its average. While, total deviation will interpret whether most of the usage pass the
maximum and minimum limit. Interpretation based on the obtained label will also be adapted with the type of customers on each cluster. Afterwards, the result of these labelling is used for the purpose of descriptive analysis.

**Table 3. Interpretations of Each Cluster**

| Cluster | Distribution of The Usage | Deviation of The Usage | Type of Customers |
|---------|---------------------------|------------------------|-------------------|
| 1       | very high                 | high                   | IMP3              |
| 2       | very high                 | very high              | IMP3              |
| 3       | very high                 | high                   | IMP3              |
| 4       | -                         | -                      | IMP3              |
| 5       | low                       | high                   | IJK3              |
| 6       | high                      | high                   | IJK1              |
| 7       | -                         | -                      | IMP3              |
| 8       | high                      | very high              | IMP3              |
| 9       | -                         | -                      | IMP3              |
| 10      | very high                 | high                   | IMP3              |
| 11      | very high                 | high                   | IMP3              |
| 12      | high                      | very high              | IMP3              |
| 13      | high                      | very high              | IJK2              |
| 14      | high                      | very high              | IMP3              |
| 15      | high                      | higher                 | IMP3              |
| 16      | very high                 | high                   | IMP3              |
| 17      | high                      | very high              | IMP3              |
| 18      | high                      | higher                 | IMP1              |
| 19      | high                      | higher                 | IMP3              |
| 20      | high                      | higher                 | IMP3              |
| 21      | high                      | very high              | IMP3              |
| 22      | high                      | very high              | IMP3              |
| 23      | very high                 | high                   | IMP2              |
| 24      | very high                 | high                   | IMP3              |
| 25      | high                      | higher                 | IMP3              |
| 26      | high                      | high                   | IMP2              |
| 27      | high                      | high                   | IMP2              |
| 28      | very high                 | high                   | IMP2              |
| 29      | high                      | high enough            | IMP1              |
| 30      | high                      | high enough            | IMP1              |
| 31      | high                      | high enough            | IMP1              |
| 32      | high                      | high                   | IMP1              |
| 33      | very high                 | high                   | IMP2              |
| 34      | very high                 | high                   | IMP2              |
| 35      | high                      | high                   | IMP2              |
The results of cluster analysis on the previous sub-section shows that each cluster has a characteristic that can be used as a reference in the creation of customer segmentation.

1) Segment 1 (consisting of cluster #2) is the 3rd type of customer in Tangerang area, and diverse sectors. This segment is dominated by chemical industry sector. Distribution and deviation characteristics of the usage of this segment is very high. Therefore, this paper recommends PT PGN to pay more attention to this segment because of the high fluctuation in the usage which could cause problem in gas distribution. One of the proposed solution to control the distribution is by managing higher rates for this segment.

2) Segment 2 (consisting of cluster #8, #12, #13, #14, #15, #16, #22 and #23) is the 3rd type of customer located in Jakarta and Bogor. This segment is dominated by food industrial sectors. Characteristic deviation of this segment usage is very high. In other words, user in this segment often pass the limit even though its fluctuation is not as high as in Segment 1. Accordingly, appropriate supervision need to be assign to users in this segment as the previous segment. Moreover, if it is necessary, proper sanction could also be applied to reduce the occurrence of the deviation.

3) Segment 3 (consisting of cluster #1, #3, #10, #11, #16, #23 and #24) is a 3rd type of customers located in Tangerang and Karawangan area, which is one of industrial area in Indonesia. This segment has the most members and the usage distribution characteristics of this segment is very high or more likely to be unstable. Based on the information, this paper recommends PT PGN to determine the rate or the contract based on daily usage.

4) Segment 4 (consisting of cluster #18, #19, #20 and #25) is the 3rd type of customer located in Jakarta and Bekasi. Distributions and deviations of their usage are not too high. Accordingly, it is suggested to give standard supervision to maintain the distribution and the deviation.

5) Segment 5 (consisting of cluster #5) is the customer type of IJK3 or Service Industry with Big Capacity, which located in Jakarta area. Usage distribution level for this segment is relatively small but it has large deviations. This happened because the service industry operates continuously, thus its usage is considered high because it passes the maximum limit every day. Therefore, it is suggested that the contract is reviewed because it seems that the assign rates or contract is not suitable with the customer needs. Moreover, as the characteristic of this segment is different from other segments, different contract plan could help PT PGN to meet their customer needs.

6) Segment 6 (made up of #28, #33, #34 and #36) is IMP2 customer type which located in Tangerang and Jakarta. The usage distribution is high while its deviation is not too high. In other words, customer in this segment has big capacity with fluctuate usage. As a suggestion, implementation of daily rates could help the company to meet the customer needs.

7) Segment 7 (consisting of cluster #26, #27, #35 and #37) is the kind of customer in IMP2, in Bekasi and Bogor. Industrial sector in this segment is vary but mostly dominated by metal fabrication industry. Characteristics of the usage distribution and deviations are high. For this reason, it is suggested that higher rates policy for this segment is implemented as in segment 1.
8) Segment 8 (made up of cluster #6) consists of 2 IJK customers and a bit of IJK 1, in Jakarta. The spread of its usage is large but still below the limit. This segment could be considered as a good user because its usage is fluctuated but it does not interfere the distribution.

9) Segment 9 (consisting of cluster #29, #30, #31 and #32) is a type of IMP1 in Bogor area, with metal fabrication industry and textile. The usage distribution and deviations are high for customers with small capacity. Moreover, the usage is also fluctuated and often pass the minimum and maximum limit. Thus, review regarding rate, penalty policy and maximum and minimum limit could help the company to meet the customer needs.

4. Conclusion
By using the six variables (area, type of customer, industrial sector, daily average, standard deviation, and total deviation), 37 clusters are formed with lowest value of error validation. The clusters are based on similar standard deviation and its total deviation value. Moreover, each cluster is identified by the distribution and the deviation of its usage, which eventually produce nine customer segmentations. Furthermore, aside from the nine segments which are proposed, PT PGN is also expected to make the fares strategy that can focus on the criteria of each cluster. Thus, the distribution of natural gas could be managed effectively.

References
[1] British Petroleum. (2014). BP Statistical Review of World Energy. London: British Petroleum.
[2] Kementerian ESDM. (2014). Peta Jalan Kebijakan Gas Bumi Nasional. Jakarta: Kementerian Energi dan Sumber Daya Mineral
[3] Badan Pengkajian Penerapan Teknologi. (2014). Outlook Energi Indonesia 2014. Jakarta: Pusat Teknologi Pengembangan Sumberdaya Energi, BPPT
[4] Bakar, W. A., & Ali, R. (2010). Natural Gas. In P. Potocnik, Natural Gas (pp. 1-37). Johor: Sciyo.
[5] Perusahaan Gas Negara. (2012). Laporan Tahunan PGN. Jakarta: PGN.
[6] Kiang, M. Y., Hu, M. Y., & Fisher, D. M. (2006). The effect of sample size on the extended self-organizing map network—A market segmentation application. Computational Statistics & Data Analysis, 51, 5940 – 5948.
[7] Badan Pusat Statistik. (2014). Distribusi Persentase Produk Domestik Bruto Triwulanan Atas Dasar Harga Berlaku Menurut Lapangan Usaha, 2000-2014 (Persen). Retrieved 2015 Februari from Badan Pusat Statistik.
[8] Direktorat Jenderal Minyak dan Gas Bumi. (2010). Artikel Mengenal Jenis-jenis Gas Bumi. Retrieved 2015 Februari from Kementerian Energi dan Sumber Daya Mineral: http://mobile.migas.esdm.go.id/?op=Artikel&lid=4
[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 37-54.
[10] Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. Physics Procedia, 62, 18-22.
[11] Han, J., & Kamber, M. (2000). Data Mining: Concepts and Techniques. San Fransisco: Morgan Kaufmann.
[12] Han, J., & RC., D. (1988). Algorithms for Clustering Data. New Jersey: Prentice Hall Inc.
[13] Hanafizadeh, P., & Mirzazadeh, M. (2011). Visualizing market segmentation using self-organizing maps and Fuzzy Delphi method – ADSL market of a telecommunication company. Expert Systems with Applications, 198–205.
[14] Harsya, T. R. (2013). Analisis Keekonomian Pemanfaatan Kilang LNG Arun Sebagai Receiving
Gas Terminal di Provinsi Aceh. Universitas Indonesia, Departemen Teknik Kimia. Jakarta: Fakultas Teknik Kimia Program Pasca Sarjana Universitas Indonesia.

[15] Hung, C., & Tsai, C.-F. (2008). Market Segmentation Based on Hierarchical Self-Organizing Map. Expert Systems with Applications, 780-787.

[16] Kaski, S. (1997). Data Exploration Using Self-Organizing Maps. Finlandia: Laboratory of Computer and Information Science, Department of Computer Science and Engineering, Helsinki University of Technology.

[17] Kementerian ESDM, Dirjen Migas. (2014). Kebijakan Alokasi Gas Bumi Untuk Dalam Negeri. Jakarta: Kementerian Energi dan Sumber Daya Mineral.

[18] Kementerian Pendidikan dan Budaya. (2015). Kamus Besar Bahasa Indonesia Versi Online. Retrieved 2015 April from Kamus Besar Bahasa Indonesia Versi Online: http://kbbi.web.id/segmentasi

[19] Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 52-65.

[20] Kohonen, T. (2001). Self-Organizing Maps. New York: Springer.

[21] Larose, D. T. (2004). Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: John Wiley & Sons Inc.

[22] Pratte, J. M. (2004). Fossil Fuels: Natural Gas. Kennesaw: Environmental Science Activities for the 21st Century.

[23] Ra'sa'nen, T., Ruuskanen, J., & Kolehmainen, M. (2008). Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Applied Energy, 830-840.

[24] Rooijers, F., Welders, L., Scheipers, B., Croezen, H., & de Bruyn, S. (2009). Natural Gas as Transitional Fuel. AD Groningen: Castel International.

[25] Salazar, G., Veles, A., Parra, M., & Ortega. (2002). A Cluster Validity Index For Comparing Non-hierarchical Clustering Methods. Medellin: Engineering Faculty, University of Antioquia.

[26] Tamberi, F. (2007). Anomaly Detection Data Mining Techniques. Pisa: Department of Computer Science University of Pisa.

[27] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Boston: Addison-Wesley.

[28] Vondracek, J., Pelikan, E., Konar, O., Cermakova, J., Eben, K., Maly, M., et al. (2008). A Statistical Model for the Estimation of Natural Gas Consumption. Applied Energy, 362-370.

[29] Zaki, M. J., & Meira Jr., W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. New York.