Supporting of Waste Management in Indonesia Using Self Organizing Map for Clustering Analysis

Arum Handini Primandari*, Ayundyah Kesumawati, Tuti Purwaningsih
Department of Statistics, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia

*E-mail: primandari.arum@uii.ac.id

Abstract. The amount of waste in Indonesia is continuously increasing, along with the increasing population and welfare. The government already takes various efforts to control the amount of garbage; for example, they prohibit using disposable plastic in any form. Mapping waste generation and waste composition hopefully will help the government (or local government) make a more precise policy. Therefore, this research employed a modern clustering analysis that was Self Organizing Maps (SOM). SOM is an Artificial Neural Network (ANN) that uses unsupervised learning to reduce data dimensions into two dimensions. SOM produces low-dimensional visualizations of high-dimensional data. The data of waste generation and composition were obtained from the National Waste Management Information System in 2017-2018. The variables processed were the daily average of waste per square kilometer (ton/km²), the daily average of waste per person (kg/person), and the daily average of waste per sub-district area (tons/sub-district). The researchers built the 3x3 hexagonal topology. The districts of Brebes, Buleleng, Cilacap, and Jepara were grouped into areas with high waste generation compared to other districts. Meanwhile, based on the composition of garbage, Morowali, Sinjai, and Palangkaraya districts were the districts where the composition of food waste, plastic waste, and textile waste was high. Tableau was used to visualize the result into map.

1.  Introduction
The increasing number of the Indonesian population raises several problems, one of which is waste management. Even though the government escalates the amount and area of final waste disposal, the management and recycling systems cannot keep up with the growth of waste from either industry or households. Moreover, the government also issues a waste management policy.

To implement the provisions of Article 23 paragraph (2) of Law Number 18 of 2008 concerning Waste Management, President Joko Widodo established Government Regulation Number 27 of 2020 concerning Specific Waste Management. Specific Waste regulated in this Government Regulation includes Waste Containing B3; Garbage Containing B3 Waste; disaster waste; Building Demolition Debris; Technically Unprocessed Waste; and Waste That Arises Not Periodically [1].

The implementation of Specific Waste Management is carried out through reducing or recycling. Reducing waste activities includes restrictions on specific waste generation, specific waste recycling, and reuse of specific waste. Meanwhile, the recycling activities process includes sorting, collection, transportation, processing, and final waste processing.

There are various types of waste that are common in Indonesia, including food scraps, plastic waste, paper waste, textile waste, trash, wood, leaves, glass waste, and others. Another research, in Residential Areas of Nigerian Tertiary Institutions, also sorted the type of waste into some of composition, such as
paper, food, plastic, glass, can/tin, textile, nylon, sand, e-waste, hair, sanitary waste, and other [2]. Almost fifty percent of waste was food.

From this research, it is hoped that it can contribute to support better waste policy making by knowing how the distribution of waste in Indonesia using a waste map and knowing the proportion of what types of waste are most abundant in certain areas. This will make it easier for policymakers to be able to manage waste better.

This study attempts to analyze waste data using a statistical method, namely, clustering analysis. The method used was Self Organizing Maps (SOM). The Kohonen Neural Network Algorithm or Self Organizing Map (SOM) is an artificial neural network method introduced by Professor Teuvo Kohonen in 1982. SOM is one of the topologies of unsupervised Neural Network, where the training process does not require supervision [3].

There are two crucial characteristics in SOM which explain that SOM can visualize and analyze high-dimensional data. Furthermore, the network can be used for clustering, dimensional reduction, classification, vector quantization, data mining, and image recognition [4][5]. SOM is a grouping method that provides topology-based class arrangements. SOM practiced iteratively through several epochs.

Tableau is a popular tool that is used to easily and attractively visualize data. In this study, Tableau was employed to visualize waste data to assist solid waste management in Indonesia. The visualization is expected to make it easier for the general public to control waste and encourage awareness to reduce and manage waste wisely. Visualization results were published on tableau public servers.

The remainder of the paper is organized as follows. Section 2 describes the research methodology. Part 3 consists of a numerical result for waste generation and waste composition. Lastly, section 4 presents concluding remarks.

2. Research Methodology

2.1. Data and methodology

The data used in this study come from the National Waste Management Information System (Sistem Informasi Pengelolaan Sampah Nasional/ SIPSN) by the Ministry of Environment and Forestry on the website [6]. There were two types of waste data available in 2017-2018, namely: waste generation and waste composition. The following is the research flowchart.

![Figure 1. Research methodology](image-url)
1. First, the researchers collected the data from the SIPSN website. The process of data cleaning included matching locations (city, district, and province) and deleting blank data (NA). Afterward, the researchers calculated variables that were waste generation produced per square kilometer per day, waste generation produced by a single person per day, and waste generation produced in a district per day.

2. The descriptive analysis described the data using graphics and boxplots.

3. The outlier of waste generation data and waste composition was handled by deleting or keeping it in the calculation. Outliers that have too high values compared to others would be deleted. Meanwhile, outliers that were still close to the distribution of the data would be kept.

4. Because of using hexagonal topology, clustering analysis began with determining the size of the hexagonal.

5. The cluster profiling process identified the characteristics of each cluster.

6. Tableau software (education version) helped in making visualization, especially geovisualization.

2.2. Self Organizing Map (SOM)

The method the researchers used is known as a self-organizing map (SOM). This method is an unsupervised computational neural network that combines both data projection and quantization or clustering. The data projection reduces the number of attributes or data vector dimensions. In contrast, the clustering reduces the number of vector inputs without losing useful information and preserving topological relationships in output space [7][8].

The input space (also called the signal) is a collection of input data used to feed the algorithm; usually, the observations are multidimensional, and therefore expressed using a vector for each observation. In contrast, the output space (trained network or SOM) refers to the set of low-dimensional universes in which the algorithm represents the input data. It is usually two-dimensional and consists of a set of elements called neurons (or nodes) linked together. The algorithm represents the input space to the output space, stores all the relevant information, and orders the observations so that topological proximity in the output space implies statistical similarity in the input space.

The input space is composed of an n-dimensional vector that the researchers wanted to visualize/cluster in a low-dimensional environment. The researchers expressed the input vector \( t \) as [5][7][9]:

\[
X = [\varphi_1(t), \varphi_2(t), ..., \varphi_n(t)]^T \in \mathbb{R}^n
\]

where \( \varphi_i \) denotes the value of each dimension.

The output space is an array of x by y neurons (nodes), which is topologically connected following some geometric rule (the most common topologies are squares and hexagons). Each node is assigned a real, parametric vector of random initial values we call the model and expressed as:

\[
k_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{in}]^T \in \mathbb{R}^n
\]

The \( d(x, k_i) \) is Euclidean distance metric between two vectors \( x \) and \( k_i \).

The representation of the signal space is arranged topologically into the network. The SOM uses an iterative process called training, where each signal vector is sequentially presented to the output space. The best match unit (BMU) for \( x \) is defined as the neuron that minimizes the distance to \( x \). When this is found, BMU activated and an adaptive process begins wherein the neuron and its neighboring topology are modified according to the following scheme:

\[
k_i(t + 1) = k_i(t) + f_{ci}(t)[x(t) - k_i(t)]
\]

where \( t \) and \( t + 1 \) represent the initial and the final state after the signal has activated the neuron; \( f_{ci}(t) \) is called the neighborhood function and reveals how bmu and its neighbors are modified.

The procedure of SOM is explained bellow [10][11]:

1. Define \( m \) the number of input vector \( l = 1, 2, ..., m \)
2. The \( l \)-th input vector has \( n \) elements, \( X_l = (x_{l1}, x_{l2}, ..., x_{ln}) \)
3. Define the number of neuron 1, 2, ..., \( N \)
4. The $i$-th neuron has $n$ elements, $k_i = (m_{i1}, m_{i2}, ..., m_{iN})$. This neuron vectors represent as two-dimensional matrix. Also, assume all vector elements are real numbers, $\mathbb{R}$.

5. For any given input $X_i$, find the closest neuron which has smallest Euclidean distance, to the given input and signified the neuron by the $k_c$:

$$k_c = \arg\min_i ||x - k_i||$$

(4)

6. For any given input $X_i$, after finding the $k_c$ neuron, update the neighborhood neuron set of $k_c$ using equation (3) for $t = 0, 1, 2, ..., T$ (T is the number of iteration).

7. Update the neurons around $k_c$:

$$f_{ci}(t) = \alpha(t) \exp\left(-\frac{d(r_c, r_i)^2}{2\sigma^2(t)}\right)$$

(5)

where $d(r_c, r_i)$ defines the distance between neurons two-dimensional matrix, $\alpha(t)$ is defined as learning rate ($0 < \alpha(t) < 1$) and $\sigma^2$ is the width of neighborhood radius. Both learning rate decrease monotonically using following equation:

$$\alpha(t) = \alpha(0) \left(\frac{\alpha(T)}{\alpha(0)}\right)^{t/T},$$

(6)

$$\sigma(t) = \sigma(0) \left(\frac{\sigma(T)}{\sigma(0)}\right)^{t/T},$$

(7)

where $T$ is the iteration or training length.

8. Do this process for every input data vector (the $m$ number of input). After $T$ iteration, the researchers will get fully learned neuron matrix that maps the input data value.

3. Numerical Result

3.1. Waste generation

The total waste generation of an area was obtained from the sum of the capital's daily waste and non-capital daily waste. The total waste generation was divided by the area, total population, and the number of sub-districts. Therefore, there were three variables for the daily average of waste, which included waste per square kilometer (ton/km2), waste per person (kg/person), and waste per sub-district (ton/sub-district). The boxplot helped to visualize the distribution of the three variables. Based on the figure, all variables have outlier values.

![Figure 2. Waste generations’ boxplot](image)

Outliers is an interesting discussion on another occasion because it is an initial identification of the high waste generation in an area. There are several possibilities for an area to be an outlier:
1. The possibility of the tourism area;
2. The area has landfills or waste management site (Reuse, Reduce, Recycle/3R) to accommodate waste from neighbor areas;
3. The area has a good waste management system so that they accurately measure and record the waste data.

The following are all districts and cities that have a high waste generation.

**Table 1. The Outliers**

| No | Variable                                                                 | District/City                                                                 | Total |
|----|--------------------------------------------------------------------------|------------------------------------------------------------------------------|-------|
| 1  | The daily average of waste generation per square kilometer (ton/km²)    | District of Bombana, Buton Tengah, Kepulauan Seribu, Kudus, Mamuju, Ogan Komering Ulu Timur, Pasaman, Purworejo, Sidoarjo, Sleman, Sukabumi, Sukoharjo, Sumbawa, Tangerang | 14    |
|    |                                                                          | City of Bandung, Bengkulu, Malang, Palangka Raya, Pontianak, Surabaya        | 6     |
| 2  | The daily average of waste generation per person (kg/person)             | District of Agam, Bangka, Buleleng, Ciamis, Deli Serdang, Gunung Kidul, Karawang, Kendal, Kepahiang, Kepulauan Seribu, Kolaka, Maros, Muara Enim, Ngawi, Pasaman, Pekalongan, Pelalawan, Pesisir Selatan, Purworejo, Sidenreng Rappang, Soppeng, Subang, Tanggamus, Tuban | 27    |

After the outliers in the X1 (daily waste generation ton/km²) and X2 (daily waste generation kg/person) variables were removed, below is a histogram of waste generation distribution. The Y-axis is the frequency showing the number of districts or cities, while the X-axis is the amount of waste generation. Based on the histogram, all the variables have left-skewed distribution.

**Figure 3. Waste generations’ histogram**

The 3x3 hexagonal topology was employed to map the data points into the two-dimensional lattice. Therefore, there were nine nodes in the network called neurons. The property of topology preserving
means that the mapping preserves the relative distance between the points. The result's visualization presented as follows.

**Figure 4.** Nodes plot of waste generation

The more intense the red color is, the less member the neuron has. The more intense the yellow color is, the more member the neuron has. The distribution of the nine neurons' members is uneven, which is between 5 and 25 members per neuron. The fan graph shows the variable values for each neuron. Based on the fan graph, neurons with a small number of members (below 6) have high scores on the average of daily waste. Meanwhile, neurons with many members have a high value on the average daily waste generated per person (in kg).

**Figure 5.** (a) The amount of members in every cluster; (b) The average of waste generation in every cluster

The values of x1 and x2 are comparable. Thus the high value of x1 is followed by a high value of x2. However, the fifth and seventh clusters have the opposite condition: the x2 has a high value, while x1 has a low value. The average daily waste per person in kg is much higher, but the average daily waste per square kilometer in ton is low. The fifth and seventh clusters have 19 and 17 members, respectively. The districts in the fifth and seventh cluster likely have a less dense population. Also, the average population in cluster 7 is the lowest, while cluster 9 is the highest.

The ninth cluster with four members has high x1 and x2 values. This cluster means that the average daily waste per person and per square kilometer has a high value. The average population in cluster 9 is the highest, which is around 5 million people. The four districts are Brebes Regency (Central Java Province), Buleleng Regency (Bali), Cilacap Regency (Central Java Province), and Jepara Regency (Central Java).
A complete visualization for waste generation in every province is published on http://bit.ly/wastegenerationmap. We can select the waste generation type to show the map. The color represents the amount of waste generation.

![Waste Generation Map in Indonesia](image)

**Figure 6. Waste generation map**

Through this dashboard, we are able to visualize the waste products of each province, which measurement is easier to understand: daily waste (kg/person), daily waste (ton/km²), and daily waste (ton/subdistrict). The more intense the red color is, the higher the amount of waste in the province. The more intense the green color is, the lower the amount of waste.

### 3.2. Waste Composition

Another waste data was the percentage of waste composition in several districts/cities in Indonesia. According to the boxplot, there are outliers in all types of waste, although not many.
The 3x3 hexagonal topology was employed to create a two-dimensional lattice so that there were a total of nine neurons. Out of the nine neurons, six were filled with members, while 3 of them were empty. Based on the figure, the neuron with intense red color has the fewest members. Cities who become a member of this neuron have evenly distributed waste composition values. Meanwhile, other neurons are less clear about the proportion of each type of waste.

To analyze each cluster's profile, the researchers took four variables types of waste to visualize the average of the percentage of leftovers, the percentage of plastic, the percentage of paper, and the percentage of the metal. The fourth cluster has significantly higher scores for the four percentages than the other clusters. Meanwhile, this cluster has the smallest members, namely the cities of Palangkaraya, the district of Morowali, and Sinjai.
Figure 9. The average of waste composition in every cluster

The following is a map of the percentage of food waste and the percentage of plastic waste. The bigger the circle’s size, the higher the percentage of plastic waste in a city. The more intense the blue color is, the higher the percentage of leftover food waste has.

Tableau allows us to create interactive visualizations. A complete visualization of waste composition can be accessed at [http://bit.ly/wastecompositionmap](http://bit.ly/wastecompositionmap). We can choose the type of composition that we want to display on this web, including the percentage of glass, fabric, rubber/leather, woody-leaves-twigs, paper, metal, plastic, leftover, and others. The dots represent the location of cities in Indonesia for which waste data is available.
Figure 10. Waste composition map

The bigger the circle is, the higher the presentation of the composition of certain categories of waste. The color of the circles indicates the cluster number—districts with the same circle color indicate that they are in the same cluster. For example, districts on the Java island are mostly turquoise—cluster number 6.

4. Conclusion
In the waste generation analysis, the outliers were 20 districts or cities with a high daily average of waste generation per square kilometer (ton). Meanwhile, 27 cities had become outliers because they had a high average daily waste generation per person (kg). All outliers were analyzed separately, both inference analysis and causal analysis. Fifty percent of outliers were located on the island of Java.

The 3x3 hexagonal topology formed 9 clusters for waste generation. The high daily average of waste generation (ton) per square kilometer (X1) was usually followed by the high average daily waste generation (kg) per person (X2). These two variables reversed their situation in areas with low population density, areas/cities in clusters 5 and 7. Cluster 9, which had a high value of X1 and X2, consisted Brebes Regency (Central Java Province), Buleleng Regency (Bali), Cilacap Regency (Central Java Province), and Jepara Regency (Central Java).

The 3x3 hexagonal topology was also employed to form clusters on the waste composition, consisting of 8 variables: the percentage of glass, fabric, rubber/leather, woody-leaves-twigs, paper, metal, plastic, leftover, and others. There are 6 clusters formed, with one cluster having a high percentage of leftovers, the percentage of plastic, the percentage of paper, and the metal percentage. This cluster had members from the city of Palangkaraya, Morowali Regency, and Sinjai Regency.

References
[1] Sekretariat Kabinet Republik Indonesia 2020 Inilah PP 27 / 2020 tentang Pengelolaan Sampah Spesifik. Pengelolaan Sampah Spesifik. [Online]. Available: https://setkab.go.id/inilah-pp-27-2020-tentang-pengelolaan-sampah-spesifik/.
[2] Amori A A, Fatile B O, Ihuoma S O and Omoregbee H 2013 Waste Generation and Management Practices in Residential Areas of Nigerian Tertiary Institutions J. Educ. Soc. Res. 3, July p. 45–52.

[3] Kesumawati A and Setianingsih D 2016 A segmentation group by Kohonen Self Organizing Maps (SOM) and K -means algorithms (case study: Malnutrition cases in Central Java of Indonesia) Int. J. Adv. Soft Comput. its Appl. 8(3) 110–115.

[4] Tewari D and Srivastava S K 2012 A Visual Recognition of Static Hand Gestures in Indian Sign Language based on Kohonen Self-Organizing Map Algorithm Int. J. Eng. Adv. Technol. 2 p. 165.

[5] Logeswari T and Karnan M 2010 An Improved Implementation of Brain Tumor Detection Using Segmentation Based on Hierarchical Self Organizing Map Int. J. Comput. Theory Eng. 2(4) 591–595.

[6] KLHK D P S 2018 Data Pengelolaan Sampah, Data Timbulan Sampah dan Data Komposisi Sampah. [Online]. Available: http://sipsn.menlhk.go.id/ . [Accessed: 06-Apr-2020].

[7] Arribas-Bel D Nijkamp P and Scholten H, 2011 Multidimensional urban sprawl in Europe: A self-organizing map approach Comput. Environ. Urban Syst. 35(4) 263–275.

[8] Kohonen T 2013 Essentials of the self-organizing map Neural Networks 37 p. 52–65.

[9] Bernard J A, Seidler R D, Hassevoort K M, Benson B L, Welsh R C and others 2012 Resting state cortico-cerebellar functional connectivity networks: A comparison of anatomical and self-organizing map approaches Front. Neuroanat. 6, AUG 2012 1–19.

[10] Hosaka K Goya T Umehara D and Kawai M, 2003 An efficient method for network topology identification based on an SOM algorithm Electr. Eng. Japan (English Transl. Denki Gakkai Ronbunshi) 142(4) 34–44.

[11] Chaudhary V Bhatia R S and Ahlawat A K, 2014 A novel Self-Organizing Map (SOM) learning algorithm with nearest and farthest neurons Alexandria Eng. J. 53(4) 827–831