Biplot and cluster analyses to describe some infectious diseases in Southeast Sulawesi

Agusrawati¹*, Gusti Ngurah Adhi Wibawa¹, Irma Yahya¹ and Bahriddin Abapihi¹

¹Department of Statistics, Halu Oleo University, Kendari, Indonesia

hadiniaizizah@yahoo.com (corresponding author)

Abstract. Infectious diseases are caused by microorganisms that can be transmitted directly or indirectly to humans. Understanding the characteristics of these diseases helps us to prevent and to overcome them. This study aims to discuss the profile of districts and cities in Southeast Sulawesi according to the spread of several types of infectious diseases by employing biplot and cluster analyses. The results showed that when the prevalence of tuberculosis is fairly high in a district or city, the HIV’s also follows. However, if these two diseases occur barely then the prevalence of malaria tends to be high, and vice versa. HIV is more likely to occur in Kota Kendari (Kendari City) and then followed by Buton and Konawe. In contrast, District of Muna with the lowest incidence of HIV and tuberculosis indicates the high prevalence of Malaria. The other types of diseases, such as pneumonia, DHF, and filariasis, are commonly found in Konawe Selatan and Kolaka.

1. Introduction
One of the WHO standards that the status of a country to be categorized as healthy is the decreasing rate of mortality caused by infectious diseases. Infectious disease is a typical disease caused by microorganisms that can be transmitted to humans directly or indirectly. As a tropical country, Indonesia has a risk of the spread of infectious diseases, like ISPA, tuberculosis, DHF, and others.

Many efforts have been carried out by the Government of Indonesia, such as increasing the health services and various eradication programs of infectious diseases. These have succeeded in reducing morbidity and mortality, but until now infectious diseases are still a problem to public health, especially in Southeast Sulawesi. Based on the 2017 annual report of the Southeast Sulawesi Provincial Health Office, the top ten common diseases found in Southeast Sulawesi were infectious diseases. One way to overcome these health problems is to understand the spreading characteristics of the diseases throughout data exploration. Data exploration could help us to identify the areas where the diseases occurred, the factors that cause the diseases, and the appropriate methods to reduce the spread of the disease. In this paper, we aim to describe the characteristics of districts and cities in Southeast Sulawesi regarding to infectious diseases by utilizing biplot and cluster analyses. The results of the research are expected to be as supplementary information and as a recommendation to local governments in making policies to combat the spread of infectious diseases.
2. Literature Review

2.1. Biplot Analysis

Biplot analysis is one of the descriptive statistical techniques to display n objects and p variables in a two-dimensional graph in an overlapping mode. There are some advantages of performing biplot display as follows (Mattjik & Sumertajaya, 2011):

- Closeness (similarity) between observed objects. Two objects, that are close to each other, are said to have similar characteristics.
- Variability of variables. Since biplot provides vector display of all variables, the lengths of vectors show the variability of variables.
- Correlation between variables. Angle between two variable vectors indicates the correlation, in which zero angle means perfect correlation and perpendicular angle means no correlation.
- Value of variables on an object. The display of biplot also shows the level of value of an object compared to others in a variable.

To conduct a biplot, suppose we have a matrix of data, X, with each column representing a variable, and each row representing the object of research. Each column of X has been corrected to its means or has been standardized.

The matrix, $nX_p$, is then decomposed using singular value decomposition (SVD) into 3 matrices ($U$, $L$, and $A$) of rank $r \leq \min \{n, p\}$, which can be written as

$$X = ULA'$$ (1)

$U$ and $A$ are matrices with orthonormal columns and $L$ is a diagonal matrix of size $r \times r$ in size with diagonals as the root of the eigenvalues - $X'X$ eigenvalues (Mattjik and Sumertajaya, 2011).

After the SVD results are obtained, equation 1 is translated into equation:

$$X = UL^\alpha L'^{-\alpha} A'$$ (2)

According to Jollife (2002) in Johnson and Wichern (2007), suppose that $G = U L^\alpha$ dan $H' = L'^{-\alpha} A'$ with $\alpha$ of $0 \leq \alpha \leq 1$. Equation 2 becomes:

$$X = U L^\alpha L'^{-\alpha} A' = GH'$$ (3)

According to Jollife (2002) in Johnson and Wichern (2007,) to describe the biplot it is necessary to take the $\alpha$ value in defining $G$ dan $H$. The selection of $\alpha$ values in $G = U L^\alpha$ dan $H = A L'^{-\alpha}$ is arbitrary with the terms $0 \leq \alpha \leq 1$, taking the values of $\alpha = 0$ dan $\alpha = 1$ is useful in biplot interpretation. To take $\alpha = 0$ can improve biplot interpretation much better, then take $\alpha = 0$, so the alleged matrix X can be arranged into :

$$\beta X_p = n U A'_p = n G_i H'_p$$ (4)

Which :

$G = U, H = AL$, and $\alpha = 0$, by taking the first two columns of the $G$ dan $H$ matrices, as follows:
The G matrix is the coordinate points of the object and the H matrix are the coordinates of the p variables to be plotted as coordinate points in the same dimension. Gabriel (1971) in Siswadi and Suharjo (1998) in presenting the size of the matrix X approach with biplots in the form of:

\[ G = \begin{bmatrix} g_{11} & g_{12} \\ \vdots & \vdots \\ g_{n1} & g_{n2} \end{bmatrix} \quad \text{dan} \quad H = \begin{bmatrix} h_{11} & h_{12} \\ \vdots & \vdots \\ h_{p1} & h_{p2} \end{bmatrix} \]

\[ \rho^2 = \frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{r} \lambda_k} \]  

(5)

With \( \lambda_1 \) is the 1st largest eigenvalue, \( \lambda_2 \) is the second largest eigenvalue, and \( \lambda_k, k = 1, 2, \ldots, r \) is eigenvalue to-\( k \). If \( \rho^2 \) approaches 1, the biplot provides a better representation of the actual data information.

2.2. Hierarchical Cluster Analysis

The procedure for group formation is divided into 2, namely hierarchy and non-hierarchy. The formation of hierarchical groups has the property of developing a tree-like hierarchy or system. The hierarchical cluster method is a grouping method in which the number of groups to be created is unknown. This technique is well processed through a combination of sequential (agglomerative) or sequential division (divisive).

2.3. Aggregation

In this method it is started by placing objects into different groups, then grouping objects gradually into larger groups. That is, first every object is treated as a bunch. Then the grouping is done gradually by checking one pair of objects most similar, then the two become one group. This process will continue, so that at the last stage, all objects become one group.

There are five agglomerative hierarchy methods in group formation, namely:

- Single Linkage
- Average Linkage
- Complete Linkage
- Ward Method (Ward’s Method)
- Centralized Method

3. Research Methods

3.1. Data Sources

The data analyzed is secondary data from the report of the Indonesia Health Profile 2017.

3.2. Research Objects and Variables

The objects used in this study were 17 districts / cities in SouthEastSulawesi and research variables are the types of infectious diseases that are commonly found, namely: TB, HIV, Leprosy, Malaria, Pneumonia in Toddlers, Filarisis, and DHF.

3.2.1. Steps of Analysis

- Prepare data that will be used in research (\( n \times p \) size data)
• Arrange data matrix that is corrected to the average, \( X \)
• Creating a matrix \( L, A \), and \( U \) matrix with the Singular Value Decomposition method
• Make a matrix \( G = UL \) and \( H' = L1-A' \)
• Take the first 2 columns of each matrix \( G \) and \( H \) so that it becomes \( G2 \) and \( H2 \) matrices
• Make a graph / plot the \( G2 \) matrix is the coordinates (x, y) for each object, in terms of this is the name of the regency / city of southeast Sulawesi, whereas each row of the \( H2 \) matrix is coordinate (x, y) for each variable, in this case the characteristic variable is the types of infectious diseases.
• Calculate the goodness of biplots in explaining the variety data.

4. Result and Discussion

4.1. Descriptive Statistics of Prevalence of Infectious Diseases
In 2017, the administrative area of the Southeast Sulawesi province consists of 15 regencies and 2 cities, namely: Buton, Muna, Konawe, Kolaka, South Konawe, Bombana, Wakatobi, North Kolaka, North Buton, North Konawe, East Kolaka, Konawe Islands, Muna Barat, Buton Tengah, Buton Selatan, Kendari City and Bau Bau City. The amount of illness due to infectious diseases is presented in Table 1

| No | Regency/City     | TBC | HIV | Pneumonia (Toddler) | Leprosy | DHF | Filarisis | Malaria | Total |
|----|-----------------|-----|-----|---------------------|---------|-----|----------|---------|-------|
| 1  | Buton           | 200 | 11  | 74                  | 26      | 14  | 6        | 35      | 366   |
| 2  | Muna            | 13  | 0   | 224                 | 17      | 14  | 0        | 530     | 798   |
| 3  | Konawe          | 446 | 9   | 364                 | 9       | 107 | 3        | 11      | 949   |
| 4  | Kolaka          | 360 | 1   | 1204                | 68      | 243 | 17       | 15      | 1908  |
| 5  | Konawe Selatan  | 277 | 0   | 700                 | 17      | 120 | 8        | 31      | 1153  |
| 6  | Bombana         | 273 | 0   | 122                 | 30      | 26  | 0        | 24      | 475   |
| 7  | Wakatobi        | 97  | 0   | 2                   | 44      | 33  | 0        | 48      | 224   |
| 8  | Kolaka Utara    | 185 | 3   | 28                  | 48      | 9   | 0        | 16      | 289   |
| 9  | Buton Utara     | 70  | 0   | 19                  | 7       | 73  | 0        | 41      | 210   |
| 10 | Konawe Utara    | 78  | 0   | 1                   | 6       | 19  | 0        | 3       | 107   |
| 11 | KolakaTimur     | 143 | 0   | 34                  | 20      | 59  | 4        | 5       | 265   |
| 12 | KonaweKepulauan | 53  | 0   | 0                   | 3       | 0   | 0        | 6       | 62    |
| 13 | Muna Barat      | 116 | 0   | 5                   | 2       | 0   | 0        | 67      | 190   |
| 14 | Buton Tengah    | 137 | 0   | 26                  | 22      | 14  | 0        | 83      | 282   |
| 15 | Buton Selatan   | 113 | 0   | 0                   | 9       | 1   | 0        | 55      | 178   |
| 16 | Kendari City    | 705 | 14  | 164                 | 23      | 93  | 0        | 67      | 1066  |
| 17 | Bau-Bau City    | 324 | 0   | 84                  | 42      | 116 | 0        | 43      | 609   |
|    | **Total of Sufferers** | **3590** | **38** | **3051** | **393** | **941** | **38** | **1080** | **9131** |

Source: R.I Health Profile, 2017

Rates of morbidity due to infectious diseases are evenly found in each district / city in Southeast Sulawesi, but in amount varies. The prevalence of TB, DHF, Pneumonia, Malaria and Leprosy are
classified as high and evenly found in all districts / cities. While the prevalence of HIV and Filarisis prevalence is uneven, it occurs in certain districts. The prevalence of pneumonia and DHF is very high in Kolaka Regency, TBC in Kendari City and Malaria in Muna Regency.

4.2. Biplot Analysis of Main Components
The result of biplot analysis is shown in Figure 1 and Figure 2 through the coordinates of object G and the coordinates of the variable H together plotted in two dimensional axes. With the help of R software applications, object coordinates and variables are presented in Table 2 below.

| Coordinate 17 Regency/City (G) | Coordinate 7 Infected Diseases Variable (H) |
|-------------------------------|-------------------------------------------|
| [.1]  | [.2]                  | [.1]                         | [.2]                         |
| [1.] -0.10089699 | -0.16584904  | [1.] -903.456670 | -652.923726     |
| [2.] -0.14060971 | 0.10890040  | [2.] -9.282951  | -13.012712      |
| [3.] -0.32357872 | -0.22633427 | [3.] -1413.063345 | 426.681273     |
| [4.] -0.71661295 | 0.44643667  | [4.] -90.922916  | -26.078987      |
| [5.] -0.43545803 | 0.18731748  | [5.] -325.294375 | -15.259673      |
| [6.] -0.14704757 | -0.20997242 | [6.] -151.890754 | -37.554097      |
| [7.] -0.03812452 | -0.10792797 | [7.] -0.07336152 | -0.18148910     |
| [8.] -0.04090814 | -0.06618681 | [8.] -0.02679672 | -0.08358145     |
| [9.] -0.06472232 | -0.13167821 | [9.] -0.01163001 | -0.05714218     |
| [10.] -0.04142471 | -0.12469019 | [10.] -0.06089571 | -0.13465436     |
| [11.] -0.03781972 | -0.12456218 | [11.] -0.30886989 | -0.64666666     |
| [12.] -0.04090814 | -0.06618681 | [12.] -0.02679672 | -0.08358145     |
| [13.] -0.04142471 | -0.12469019 | [13.] -0.06089571 | -0.13465436     |
| [14.] -0.03781972 | -0.12456218 | [14.] -0.30886989 | -0.64666666     |
| [15.] -0.04090814 | -0.06618681 | [15.] -0.02679672 | -0.08358145     |
| [16.] -0.04142471 | -0.12469019 | [16.] -0.06089571 | -0.13465436     |
| [17.] -0.04090814 | -0.06618681 | [17.] -0.02679672 | -0.08358145     |

Based on the scatter plot of the two Main Components in Figure 1, Biplot Analysis in Figure 2, and Cluster analysis in Figure 3. Grouping the Regencies/cities in Southeast Sulawesi according to the prevalence level of various types of infectious diseases. There are 4 regencies, there are Kolaka, Muna, Konawe Selatan and Kendari City, each of them isolated by themselves while the other regencies are seen clustered.
In Figure 2, there is a fairly high correlation between the prevalence of tuberculosis and HIV because the vectors of these two variables form a sharp angle ($<90^\circ$), it means that if the spread of HIV disease increases in an area then followed by high spread of TB disease. Vice versa appliesTB disease and HIV negatively correlated with Malaria (opposite vector direction). DHF and pneumonia in toddlers also have a fairly high correlation.
Figure 3. Dendrogram of districts/cities in Southeast Sulawesi based on prevalence level of infectious disease

Based on Figure 2 and Figure 3, about 5 groups of districts / cities can be formed based on the prevalence rate of various infectious diseases. When viewed from the regional groups, group I is Kolaka Regency characterized by a very high prevalence of Pneumonia, DHF, Filarisis and Leprosy, or above average, while other diseases are relatively low or below average, because the vector's direction is opposite. Group II is Muna Regency which is characterized by high sufferers of Malaria disease, showed from the direction of the Malaria disease vector in the direction of the position of Muna Regency. Group III is Konawe Selatan Regency characterized by a high prevalence of DHF, Pneumonia and Leprosy. Group IV namely Kota Kendari and Konawe were characterized by very high TB and HIV sufferers, other districts as group V namely Buton, Bau-Bau, Wakatobi, Kolaka Utara, Bombana, Kolaka Timur, Buton Selatan, Buton Tengah, Muna Barat. Konawe Kepulauan, Konawe Utara and Buton Utara, the spread of all types of diseases is below average because there is no vector of disease types that support it.

Scater Plot 2 Main Components in Figure 1 and Biplot in Figure 2 explain the diversity of data 71.1% of the total information contained in the data. This means that the resulting map can represent information about the mapping of 17 Regencies/Cities of Southeast Sulawesi based on 7 types of infectious diseases by 71.1%. This result is quite good because it has a percentage of over 60%. Similar research with this research has been done by Bambang. H et al. (2013) case studies of several regions in Indonesia, but the Southeast Sulawesi region has never been studied.

5. Conclusions and Suggestions
The spread of seven types of infectious diseases namely, tuberculosis, HIV, DHF, Malaria, Leprosy, Pneumonia in Toddlers, and Filarisis evenly occurred in Regencies/ Cities in Southeast Sulawesi. Mapping of the Muna Regency, Kolaka, Konawe Selatan are far apart from the other districts. Muna Regency is characterized by a high number of Malaria sufferers. Kolaka Regency is characterized by Pneumonia, DHF, Filarisis and Leprosy which is very high, or above average, and follows South Konawe District. The prevalence of HIV and tuberculosis is very high in the cities of Kendari and Konawe, and other Regency groups namely Buton, Bau-Bau, Wakatobi, Kolaka Utara, Bombana, Kolaka Timur, Buton Selatan, Buton Tengah, Muna Barat. Konawe Kepulauan, Konawe Utara and Buton Utara, the spread of all types of diseases is below average or the number of sufferers is relatively low. The results of this analysis can hopefully be a source of information and recommendations to the local government in making policies to deal with the spread of infectious diseases, especially paying attention to areas which is convenient to hit by infectious diseases.
References

[1] Johnson & Winchern, 2007, Applied Multivariate Statistical Analysis, Sixth Edition, Prentice Hall International, Inc. Upper Saddle River, New Jersey

[2] Heryanto, B and Kinasi, R.R, 2013. Bipot Analysis of Disease Case Data in Several Areas of Indonesia in 2009. Health Research Bulletin. Vol. 41, No.2.

[3] Mattjik, A.A and Sumertajaya, I.M, 2011, Double Variable Investigation Using SAS, IPB. Bogor

[4] Siswadi and Suhrjo, B, 1998. Analysis of Exploration of Multiple Variable Data. IPB. Bogor

[5] Data and Information, Indonesian Health Profile, 2017. Ministry of Health of the Republic of Indonesia, Jakarta