Analyzing Infectious Disease in Multiple District in East Nusa Tenggara (ENT) using K-Means Clustering and Correspondence Analysis

M F Adhari[1], J J Jakson[2], G L Sulistyoreni[2], A S Larissa[2], Y S Afrianti[3,*], and F H Sulaiman[4]

[1] Mathematics master study program, faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Ganesa street No 10 Bandung, 40132, Indonesia
[2] Mathematics bachelor study program, faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Ganesa street No 10 Bandung, 40132, Indonesia
[3] Statistics research division, faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Ganesa street No 10 Bandung, 40132, Indonesia
[4] Computational science master study program, faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Ganesa street No 10 Bandung, 40132, Indonesia

*Corresponding author's email: yuli.afrianti@itb.ac.id

Abstract. Infectious diseases remain a major public health concern in Indonesia, particularly in East Nusa Tenggara (ENT), where tuberculosis (TBC), dengue haemorrhagic fever (DHF), and HIV/AIDS are obtaining high cases. These diseases are not only influenced by individual and environmental factors but also by spatial characteristics such as population distribution and regional infrastructure. Therefore, analyzing spatial factors is crucial to better understand and manage the spread of infectious diseases in ENT. This study uses data from 2023 to 2024 across 22 districts in ENT, focusing on the prevalence of TBC, DHF, and HIV/AIDS. K-means clustering is first applied to classify the districts into three groups based on area size and population, aiming to identify spatial patterns of disease severity. The clustering process yields a silhouette coefficient of 0.48, indicating moderately valid group separation. Subsequently, correspondence analysis is used to examine the relationship between the resulting clusters and the three diseases. The result reveals that Cluster A, which has the highest population density, shows a strong association with all three infectious diseases. These findings suggest that population density plays a significant role in the transmission of infectious diseases and should be considered in future health intervention strategies.
Keyword: Correspondence analysis, East Nusa Tenggara, infectious diseases, K-means clustering

1. Introduction

In Indonesia, infectious diseases remain a health challenge because there is a lack of infrastructure on public health. According to recent research, many infectious diseases (e.g tuberculosis, dengue haemorrhagic fever, and HIV/AIDS) have significant impact on economic stagnation [1]. This problem is spreading also to East Nusa Tenggara (ENT) which has high incidences of TB, DHF, and AIDS rather than other cities due to systemic healthcare gaps [2]. Multiple factors actually contribute to the spreadof these diseases, especially population density, environmental conditions, and health center facilities [3]. Therefore, spatial analysis has an important role in public health research.

To that aim, this study applies K-means clustering, an unsupervised machine learning algorithm that partitions data into groups based on intra-cluster similarity [4]. This method is used for detecting patterns of disease severity by the location of each region [5]. Other than that, we also use correspondence analysis that explores relationships between categorical variables. It enables visualization of the association between disease types and regional clusters, facilitating a deeper understanding of how disease profiles align with regional classifications [6].

In summary, this paper analyzes infectious disease patterns across 21 districts in East Nusa Tenggara using data from 2023 and 2024. By integrating K-means clustering and correspondence analysis, we categorize districts based on disease severity and investigate the association between clusters and disease types. This integrated analytical approach aims to inform regional public health strategies through data-driven insights.

This study offers several novel contributions to the existing body of research on infectious disease analysis in Indonesia. First, this research integrates Correspondence Analysis (CA) and K-Means Clustering to explore and group districts based on their similarities in infectious disease profiles. Second, by examining multiple infectious diseases simultaneously, this study discovered the patterns and shared regional characteristics. Finally, the findings offer insights for local governments and health agencies by identifying district clusters that require similar types of interventions, thereby supporting more targeted and data-driven public health strategies in East Nusa Tenggara.

2. Research Method

This study focuses on the analysis of infectious diseases across districts and municipalities in the East Nusa Tenggara (ENT) Province using the K-Means clustering method and Correspondence Analysis. ENT was selected as the study area due to its consistently higher incidence of infectious diseases compared to other regions in Indonesia, as indicated by national statistical data. The data utilized in this study are secondary data obtained from the BPS Statistics of ENT Province, comprising the number of infectious disease cases reported in each district and municipality for the years 2023 and 2024. The K-Means clustering method is employed to group administrative regions based on similarities in disease incidence, while Correspondence Analysis is used to explore associations between specific regions and types of diseases, thereby providing a comprehensive overview of the spatial and categorical dynamics of infectious diseases in ENT.

2.1. K-Means Clustering

K-Means Clustering is a non-hierarchical data clustering method that divides data into several clusters by grouping data with similar features together and separating data with different characteristics into different groups [7]. The data set is divided by the algorithm into k predetermined, unique, non-overlapping subgroups (clusters), with each data point belonging to a single group. In addition to maintaining the clusters as distinct (far) as possible, it attempts to make the intracluster data points as comparable as possible. It groups data points into clusters so that the total squared distance between the cluster's centroid and the data points is as small as possible. Data points within the same cluster are more homogeneous (similar) when there is less diversity within the clusters. The k-means algorithm is composed of the following steps:

1. Determine the number of desired clusters, denoted by k.
2. Initialize k centroids.

3. Calculate the distance from each data point to every centroid using the Euclidean Distance formula.
4. Assign each data point to the cluster with the nearest centroid based on the calculated distances.
5. Recalculate the positions of the centroids based on the current cluster members.
6. Once the final condition is met, the centroid values from the last iteration are parameters for data classification.

Steps 2 to 5 are repeated until the centroids no longer move. As a result, the patients are divided into homogeneous groups, maximizing group heterogeneity. The elbow approach was used to determine the ideal number of clusters (k) [8]. Elbow method provides a systematic, data-driven approach to select the best value of k by analyzing the Within Cluster Sum of Squares (WCSS). WCSS itself is a measure of how compact the clusters are. This consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of groups to use.

2.2. Correspondence Analysis

Correspondence Analysis (CA) is a multivariate graphical technique designed to explore relationships among categorical variables [9]. CA decomposes the chi-square statistics associated with the data table into two sets of orthogonal components that describe, respectively, the pattern of associations between the elements of the rows and between the elements of the columns of the data table [10]. The correspondence analysis algorithm is composed of the following steps:

1. Contingency Table

Contingency table of size $I \times J$ is a matrix that displays the joint frequencies of occurrence between two categorical variables. Let $N = [n_{ij}]$ be the contingency table, where n_{ij} indicates the number of observations that fall into the combination of the i-th category of the first variable and the j-th category of the second variable, and $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ is the grand total.

2. Relative Frequency Matrix

The matrix of relative frequencies is given by:

$$P = \frac{1}{n}N, \qquad p_{ij} = \frac{n_{ij}}{n} \tag{1}$$

3. Row and Column Masses

Row masses r_i and column masses c_j are computed as:

$$r_i = \sum_{j=1}^{J} p_{ij}, \qquad c_j = \sum_{i=1}^{I} p_{ij} \tag{2}$$

4. Distance Matrix

Correspondence analysis uses the Chi-squared distance measure to produce the distance matrix. The chi-squared distance between two rows i and k is defined as:

$$d^2(i,k) = \sum_{j=1}^{J} \frac{1}{c_j}\left(\frac{p_{ij}}{r_i} - \frac{p_{kj}}{r_k}\right)^2 \tag{3}$$

5. Matrix Centering and Standardization

The expected frequencies under independence are given by the outer product rc^T. The matrix of standardized residuals is:

$$S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}} \tag{4}$$

where D_r and D_c are diagonal matrices of row and column masses, respectively.

6. Singular Value Decomposition (SVD)

The matrix S is decomposed using SVD:

$$S = U\Sigma V^T \tag{5}$$

where U and V are matrices of left and right singular vectors, and Σ diagonal matrix that contains the singular values. The squared singular values correspond to the principal inertias (i.e., explained variance) of the dimensions.

7. Principal Coordinates
The coordinates used for visualization are:

$$F = D_r^{-\frac{1}{2}} U\Sigma, \qquad G = D_c^{-\frac{1}{2}} V\Sigma \tag{6}$$

These coordinates represent the positions of rows and columns in the reduced-dimensional space, typically two dimensions.

8. Inertia
Inertia is analogous to variance in PCA and quantifies the overall dispersion of profiles from their centroid:

$$Total\ Inertia = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{7}$$

This is also the Pearson chi-square statistic divided by n:

$$Inertia = \frac{1}{n}\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \qquad where\ e_{ij} = n r_i c_j \tag{8}$$

3. Result and Discussion

3.1. Arrangement of Clustering and Spatial Grouping

This study utilized area size and total population in the K-Means clustering algorithm to identify spatial patterns across 22 districts in East Nusa Tenggara. Since K-Means depends on Euclidean distance, both variables were standardized to ensure equal contributions in distance calculation. The Elbow method was used to determine the ideal number of clusters, and a distinct "elbow" was flattened at k = 3.

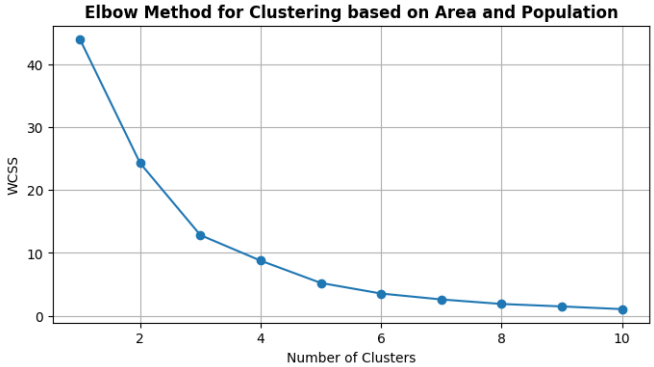

Figure 1. Elbow Method Curve for Optimal Number of Clusters (k = 3)

K-Means++ initialization was used to minimize randomness in the centroid selection process. **Figure 2** shows that the clustering converged in three iterations with three distinct groups, such as clusters A, B, and C.

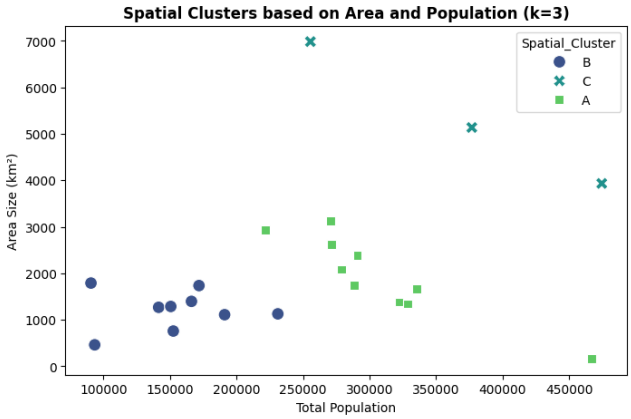

Figure 2. Scatter Plot of K-Means Spatial Clusters (Area vs Population)

Additionally, Euclidean distance calculations were used to determine each cluster's centroid. The details of cluster for each region can be seen in **Appendix A**.

3.2. Cluster Characteristics and Disease Indicators

These cluster assignments are spatially mapped throughout the province in **Figure 3.** with each district colored based on its cluster assignment.

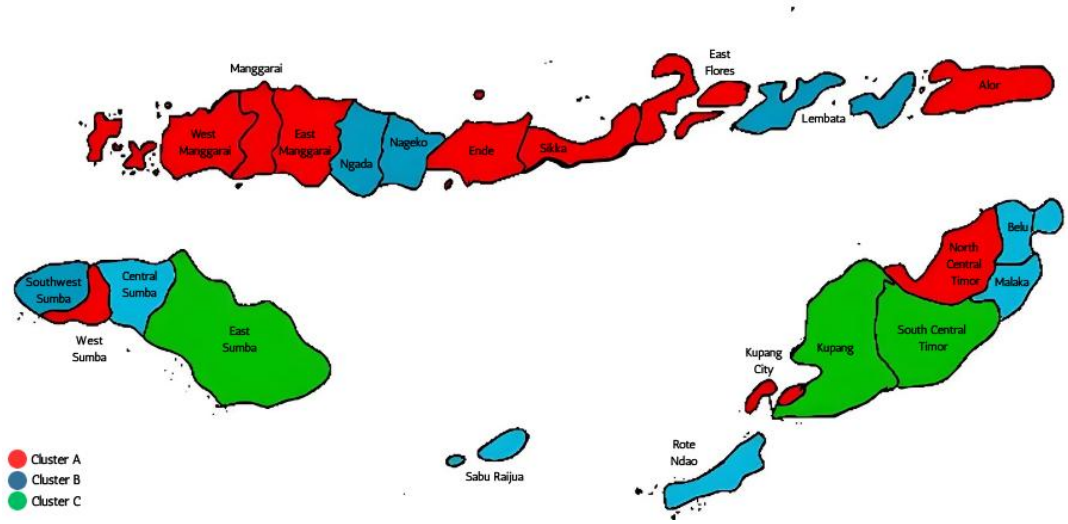

Figure 3. Cluster Assignments Visualized on the East Nusa Tenggara Map

Consequently, the following are the specific hypotheses about each cluster and its disease cluster:

- Regions with a moderate area and a high population density (e.g. Kupang City and Southwest Sumba) are included in Cluster A. These areas have high cases per km^2, indicating concentrated disease zones, but moderate total cases.
- Small, sparsely populated areas with the lowest values in both disease indicators (e.g. Belu and Sabu Raijua) belong to Cluster B. This implies diffused and low-intensity risk.
- Large, populated areas (e.g. East Sumba and South-Central Timor) that show low spatial density, but high proportional and total cases are included in Cluster C. Due to their size, these areas need more extensive intervention.

Clustering performance was evaluated using the Silhouette Score, which resulted in a value of 0.480. It suggests a reasonably good separation between clusters. The correspondence analysis shown in the following section depends on these clustering results.

3.3. *Correspondence Analysis*

A contingency table was constructed by adding the number of diseases over two years in each of the clusters previously obtained, resulting in the following contingency table:

Table 1. Contingency table

	TB	Dengue	AIDS
Cluster A	10086	4338	1308
Cluster B	1695	1441	385
Cluster C	4069	736	319

From the contingency table, a correspondence will then be checked by calculating the correspondence matrix by dividing each entry in the contingency table by the total entries in the contingency table. After determining the dependency, a chi-square test will be conducted to determine whether there is a dependence between the disease variable and the cluster variable. The hypotheses used are as follows:

H_1: Disease variable and cluster variable are independent of each other
H_0: Disease variable and cluster variable are not independent

Using $\alpha = 0{,}05$, a p-value of 7.06×10^{-202} was obtained. Since the p-value < 0.05, we reject H_0, indicating that the disease variable and the cluster variable are dependent.

3.4. *Inertia*

In correspondence analysis, the inertia matrix represents the total variance of the data, which is distributed across the principal dimensions. The inertia matrix helps to assess the significance of each dimension, with higher values indicating dimensions that capture more meaningful variation in the data. The percentage of inertia obtained is as follows:

Table 2. Percentage of inertia (%)

Dimension	1	2	3
1	100	0	0
2	0	0	0
3	0	0	0

A 100% inertia indicates that all the variance in the data is accounted for by the dimensions derived from the analysis. This means that the components identified through techniques such as Singular Value Decomposition (SVD) fully explain the relationships and patterns within the data. Therefore, the analysis has captured all the relevant information, and no additional dimensions are required to represent the data comprehensively.

3.5. *Map Result*

Since the inertia for one dimension alone is already 100, a one dimension plot is used. The map result of column asymmetric in one dimension obtained is as follows:

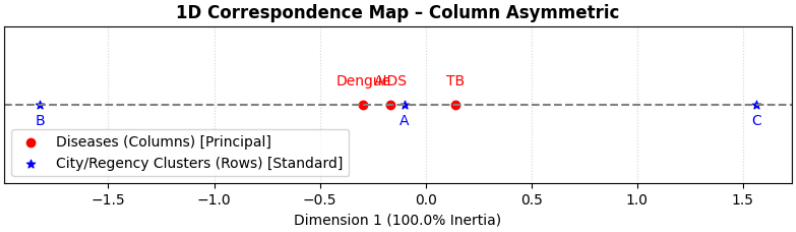

Figure 4. Column Asymmetric Map

Diseases TB, Dengue (DBD), and AIDS are strongly associated with Cluster A. Cluster A exhibits a relatively high population density compared to other districts in ENT, as shown in **Appendix B**. As presented in Appendix A, High population density creates crowded living conditions that accelerate the spread of infectious diseases such as TB, Dengue, and AIDS. This explains the strong association between Cluster A and these diseases.

4. Conclusion

We clustered 21 regencies and 1 city in East Nusa Tenggara to 3 clusters (A, B, and C) with K-Means algorithm based on area and population in each region. This clustering has quite good results because the silhouette coefficient is 0.48. All the diseases (TBC, AIDS, and DHF) are strongly related to cluster A because it has a high density of population. It means that the area and population of the region could affect the spread of various infectious diseases.

References

[1] T. B. Nutman, "Tropical infectious diseases: diagnostics, therapeutics, and vaccines, in *Hunter's Tropical Medicine and Emerging Infectious Diseases*, 10th ed., Philadelphia: Elsevier, **2020**, pp. 54–69.

[2] BPS–Statistics Indonesia, *East Nusa Tenggara Province in Figures 2024*. Kupang: Badan Pusat Statistik, 2024.

[3] World Bank, *Information and Communication Technologies: A World Bank Group Strategy*. Washington, DC: World Bank, 2002.

[4] A. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, **vol. 31**, no. 8, pp. 651–666, July 2010.

[5] A. Wanto et al., Data Mining : Algoritma dan Implementasi. Yayasan Kita Menulis, 2020.

[6] F. Sudweeks, *Development and Leadership in Computer-Mediated Collaborative Groups*. Ph.D. dissertation. Murdoch, WA: Murdoch Univ., 2007. [Online]. Available: Australasian Digital Theses Program.

[7] Amalina, T., Bima, D., Pramana, A., & Sari, B. N, "Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food", *Jurnal Ilmiah Wahana Pendidikan*, **vol. 8**, no. 15, p. 574–583, Sept. 2022.

[8] Purnima Bholowalia and Arvind Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN", *International Journal of Computer Applications*, **vol. 105**, no. 9, p. 17-24, Nov. 2014.

[9] Kroonenberg, P. M., & Greenacre, M. J, "Correspondence Analysis", *Encyclopedia of Statistical Sciences*, 2nd ed. New Jersey: Wiley, 2004.

[10] Greenacre. M, Correspondence Analysis in Practice, 3rd ed, New York: Chapman and Hall/CRC, 2017

Appendix A
Cluster Province Members

City/ Regency	Population	Area (km²)	Spatial_Cluster	TB	Dengue	AIDS	Total Cases	Cases per 100k	Cases per km²
Ende	278581	2085.24	A	14	62	52	128	45.947.139	0.061384
East Manggarai	290790	2389.53	A	7	168	19	194	66.714.811	0.081188
Southwest Sumba	322073	1383.31	A	7885	445	81	8411	2.611.519.749	6.080.344
West Manggarai	270917	3129.00	A	34	835	73	942	347.707.970	0.301055
Manggarai	328758	1343.83	A	41	322	101	464	141.137.250	0.345282
Sikka	335360	1671.65	A	673	1633	156	2462	734.136.450	1.472.796
East Flores	288310	1748.52	A	231	31	52	314	108.910.548	0.179580
Kupang City	466632	159.33	A	57	473	529	1059	226.945.430	6.646.583
Alor	221536	2928.56	A	1118	332	157	1607	725.390.004	0.548734
North Central Timor	271277	2623.20	A	26	37	88	151	55.662.662	0.057563
Belu	231008	1127.25	B	49	231	123	403	174.452.833	0.357507
Malaka	190994	1109.16	B	139	43	36	218	114.139.711	0.196545
Ngada	171736	1735.64	B	7	325	18	350	203.801.183	0.201655
Rote Ndao	150521	1286.45	B	103	5	24	132	87.695.405	0.102608
Central Sumba	90521	1789.66	B	68	51	19	138	152.450.813	0.077110
Nagekeo	166063	1396.16	B	8	91	20	119	71.659.551	0.085234
Sabu Raijua	93330	460.96	B	5	181	19	205	219.650.702	0.444724
Lembata	141391	1268.11	B	36	46	79	161	113.868.634	0.126961
West Sumba	152414	757.41	B	1280	468	47	1795	1.177.713.333	2.369.919
South Central Timor	474521	3933.15	C	658	72	132	862	181.656.871	0.219163
Kupang	376837	5136.51	C	76	205	92	373	98.981.788	0.072617
East Sumba	255498	6984.01	C	3335	459	95	3889	1.522.125.418	0.556843

Appendix B
Population density in ENT (people/km²)

City/Regency	Spatial_Cluster	2023	2024
East Flores	A	165	167
Ende	A	134	135
Sikka	A	201	204
East Manggarai	A	122	124
Southwest Sumba	A	233	238
Manggarai	A	245	249
North Central Timor	A	103	105
West Manggarai	A	87	88
Alor	A	76	77
Kupang City	A	2929	2980
Rote Ndao	B	117	119
Lembata	B	111	113
Nagekeo	B	119	121
West Sumba	B	201	205
Malaka	B	172	174
Ngada	B	99	100
Central Sumba	B	51	52
Belu	B	205	209
Sabu Raijua	B	202	206
Kupang	C	73	74
South Central Timor	C	121	122
East Sumba	C	37	37