Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2024, 2024:119

https://doi.org/10.28919/cmbn/8881

ISSN: 2052-2541

# NETWORK PREDICTION BASED ON CLUSTERING: CASE STUDY FOR HUMAN SETTLEMENTS ALONG URBAN ROADS

MOKHAMMAD RIDWAN YUDHANEGARA[1,*], SISWADI[2], SISILIA SYLVIANI[3], KARUNIA EKA LESTARI[1], EDWIN SETIAWAN NUGRAHA[4]

[1]Mathematics Education Department, Universitas Singaperbangsa Karawang, Karawang 41361, Indonesia

[2]Mechanical Engineering Department, Universitas Singaperbangsa Karawang, Karawang 41361, Indonesia

[3]Mathematics Department, Universitas Padjadjaran, Sumedang 45363, Indonesia

[4]Study Program of Actuarial Science, Faculty of Business, President University, Bekasi 17550, Indonesia



**Abstract:** Indonesia is in a category with a very dense population of 279,918,617 people in July 2024. Population density can lead to a decrease in the quality of health services. One of the causes is the high rate of transmission of viral diseases in densely populated areas. For this reason, an innovative strategy is needed to inhibit the spread of the virus. Network clustering and predictive distribution techniques in delivering health logistics in densely populated areas can improve the efficiency and effectiveness of health logistics delivery. The techniques in delivering health logistics in densely populated areas can improve efficiency and effectiveness in dealing with the rate of virus spread. Based on these methods, the problem of delivering health logistics will be easy because zone predictions from network clustering results provide information on the location and density of the area. The method allows medical officers to prioritize the delivery zone for health logistics. This method can also overcome the next wave of viruses, such as

*Corresponding author

E-mail address: mridwan.yudhanegara@staff.unsika.ac.id

Received September 01, 2024

COVID-19 and other infectious diseases.

**Keywords:** health logistics delivery; multinomial distribution; dynamic network; spectral bisection.

**2020 AMS Subject Classification:** 92D30.

## 1. INTRODUCTION

Sustainable Development Goals (SDGs), established by the United Nations (UN), aim to ensure a better and more sustainable life for all people. The SDGs are a set of goals set by the United Nations to achieve a better and more sustainable life for all people. There are 17 interrelated and mutually supportive SDGs to address the various global challenges faced [1] [2].

The SDGs are global and national commitments to improve society's welfare, including 17 global goals and targets for 2030 declared by both developed and developing countries at the UN General Assembly in September 2015. The goals are no poverty; no hunger; healthy and prosperous lives; quality education; gender equality; clean water and proper sanitation; clean and affordable energy; decent work and economic growth; Industry, Innovation and Infrastructure; reduced inequality; sustainable cities and settlements; responsible consumption and production; tackling climate change; marine ecosystems; terrestrial ecosystems; peace, justice and resilient institutions; partnerships to achieve goals [3] [4].

The contribution of this research to the SDGs is specifically on the goal of a healthy and prosperous life. One realization of this goal is through the role of mathematics and statistics in dealing with population density. Population density is a severe problem because Indonesia's population is 279,918,617 people in July 2024. Indonesia's population equals 3.45% of the world's population. Indonesia ranks 4th in the list of countries (and dependencies) by population. The population density in Indonesia is 153 per $Km^2$, and 59.1% of the population lives in urban areas (163,963,233 people in 2023).

High population density can decrease the quality of health services. Transmission of viral diseases in densely populated areas results in a high rate of spread. For this reason, an innovative strategy is needed to inhibit the spread of the virus. One strategy is the delivery of health logistics,

such as medicines and equipment to support health.

The technique to overcome these problems is cluster analysis. Based on previous research, this cluster analysis also makes it easier to solve significant data clustering cases, both in the field of education [5] [6] and the transportation business [7]-[10]. In this paper, the research focuses on the delivery of health equipment. The novelty of this research lies in the prediction network that involves network clustering and predictive distribution. This research is a continuation of previous studies [5]-[10].

Delivering medicines and health equipment in densely populated areas requires an efficient and effective strategy to ensure that the community can access them easily and quickly. Clustering and predictive distribution techniques in delivering such health logistics in densely populated areas are expected to improve efficiency and effectiveness. Thus, the problem of health logistics delivery will be solved quickly because the zone prediction from the clustering network results provides information on congested locations and routes. So that medical officers can prioritize health logistics delivery zones.

## 2. MATERIALS AND METHODS

### *2.1 Data*

Let $\mathbf{X}_t = \left[X_{1,t}X_{2,t}\cdots X_{m,t}\right]^T$ be a random vector representing the number of people living along road segments at time $t$, $X_{i,t}$ is a random variable representing the number of people living along the $i$-th road segment at time $t$. In the network formed from the map in Figure 1, locations or destination areas are represented as nodes and road segments are represented as edges. In this simulation, the network consists of 129 nodes $(\mathrm{V} = \{v_0 = 0, v_1 = 1, v_2 = 2\,, \cdots, v_{128} = 128\})$ and consist of 260 edges $(\mathrm{E} = \{e_1, e_2, \cdots, e_{260}\})$. Each node is labeled from 0 to 128, and edges are marked with the number of people living along each road segment at time $t$. Furthermore, the data was generated 28 times under the assumption of a multinomial distribution with a mass probability function $\mathbf{X}_t \sim \mathrm{Mult}\,(\theta_1, \theta_2, \cdots, \theta_{260}, 1{,}333{,}100)$ [10],

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1{,}333{,}100!}{\prod_{i=1}^{260} x_{i,t}!}\prod_{i=1}^{260}\theta_i^{x_{i,t}}, \tag{1}$$

where $\sum_{i=1}^{260} \theta_i = 1$ and $\sum_{i=1}^{260} x_i = 1{,}333{,}100$.

The number 1,333,100 is the total population living in urban areas in Bandung city along the roads shown in Figure 1. As of June 2024, the total population of Bandung City was 2,569,107 people (source: www.bandungkota.bps.go.id), but only a subset of this data was include in the simulation. Then, the network is taken from the map of Bandung City-West Java Indonesia, see Figure 1. The parameter $\boldsymbol{\theta}$ is randomly generated from the standard uniform distribution $\big(U(0;1)\big)$. The resulting data is used as the weight of the edge in the network, while the weight is the number of people living in the area along the road or edge, with a total of 1,333,100 people.

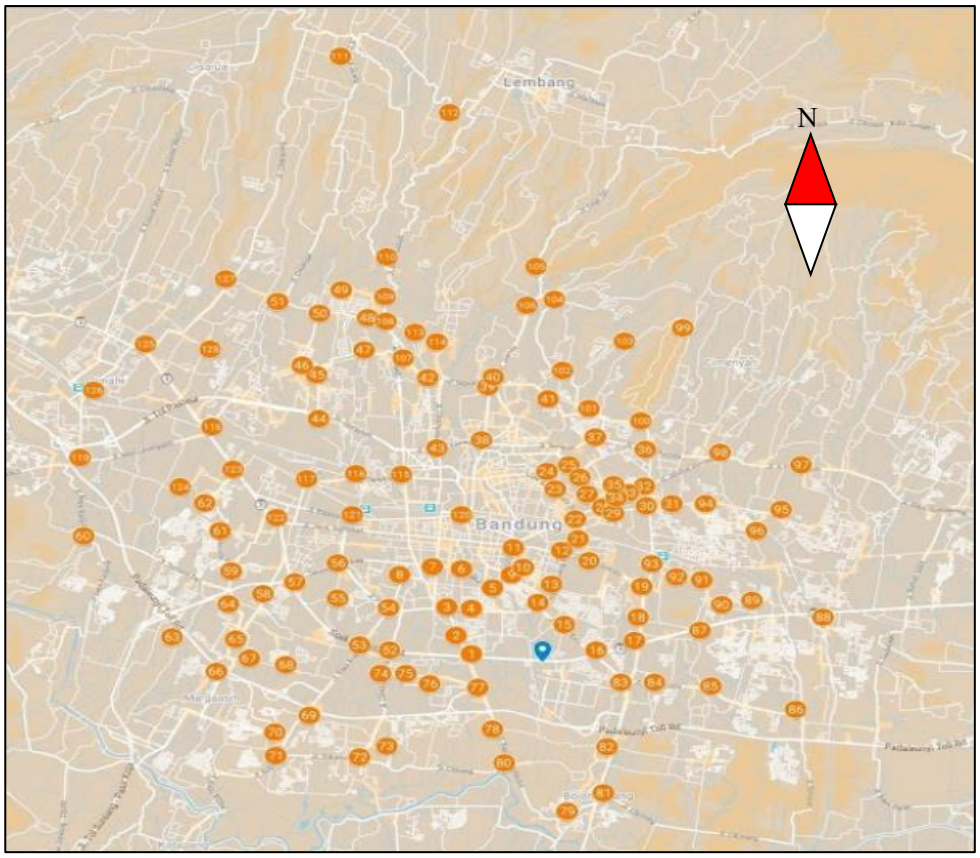


FIGURE 1. Bandung City map from google map application [8] [10]

*2.2 Network Clustering*

The term *network* in mathematics is known as a connected graph. Networks can be classified as either dynamic or static. A dynamic network is a network that continues to grow over time, so its elements can change, such as node or point or vertex, edge, and edge weights [11]. The

condition of a static network is unlike that of a dynamic network. In this research, the dynamic network will be clustered into zones or subnets in the network.

A *spectral bisection* is a clustering method that divides the network into two clusters. The resulting cluster formation is based on the Fiedler vector of the Laplace matrix that corresponds to the nodes in the network. Let $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ where $\mathrm{V} = \{\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_n\}$ be connected graph, and let $\boldsymbol{\varphi}_2$ be a Fiedler vector, If given $\mathcal{r} \geq 0, \mathcal{r} \in \mathbb{R}$, defined $\mathrm{V}_1 = \{\mathrm{v}_i \in \mathrm{V}: \varphi_{2_i} \geq -\mathcal{r}\}$ then the induced subgraph of $\mathrm{V}_1$ is connected. For $\leq 0, \mathcal{r} \in \mathbb{R},$, the induced subgraph $\mathrm{V}_2 = \{\mathrm{v}_i \in \mathrm{V}: \varphi_{2_i} \leq -\mathcal{r}\}$ is also connected [12]. The spectral bisection algorithm is described in Algorithm 1 [6].

ALGORITHM 1. Spectral bisection recursively:

1) Get the matrix Laplace $\mathbf{L}_\mathrm{G}$ of $\mathrm{G}(g_e)$.
2) Get the Fiedler vector $\boldsymbol{\varphi}_2$ from the second smallest eigenvalue $\lambda_2$.
3) Calculate the median $\mathrm{me}_{\boldsymbol{\varphi}_2}$.
4) Find the community members by selecting $\mathrm{V}_1 = \{\mathrm{v}_\mathrm{i} \in \mathrm{V}: \boldsymbol{\varphi}_{2_i} < \mathrm{me}_{\boldsymbol{\varphi}_2}\}$ and $\mathrm{V}_2 = \{\mathrm{v}_\mathrm{i} \in \mathrm{V}: \boldsymbol{\varphi}_{2_i} > \mathrm{me}_{\boldsymbol{\varphi}_2}\}$.
5) Get the matrix Laplace of subnet $\mathrm{G}'(g_e)$.
6) Determine the community members according stage 2, stage 3, and stage 4.

A numerical approach is used to determine the second smallest eigenvalue $(\lambda_2)$ of $\mathbf{L}_\mathrm{G}$ [10]. This algorithm performs clustering in stages; the first step is to divide a network into two clusters, and the next step, with an iterative process, is to cluster each cluster so that $2^{\mathcal{k}}, \mathcal{k} \in \mathbb{N}$ clusters of a network are obtained.

*2.3 Predictive Distribution*

The conditional probability distribution for observation $\mathbf{x}_{t+k}$ given observation $\mathbf{D}_{t+(k-1)} = \left[\mathbf{x}_1\ \mathbf{x}_2 \cdots \mathbf{x}_{t+(k-1)}\right]^T$

(2) $$p\left(\mathbf{x}_{t+k}\middle|\mathbf{D}_{t+(k-1)}\right)=\int p(\mathbf{x}_{t+k}|\boldsymbol{\theta})f\left(\boldsymbol{\theta}\middle|\mathbf{D}_{t+(k-1)}\right)d\theta.$$

The derivation of the predictive distribution function for multinomials has been discussed in Indratno et al. [8]. Based on Equation 2, it is obtained

(3) $$p\left(\mathbf{x}_{t+k}\middle|\mathbf{D}_{t+(k-1)}\right)=\frac{n!\Gamma\left(\sum_{i=1}^{m}\alpha_i'\right)}{\Gamma\left(n+\sum_{i=1}^{m}\alpha_i'\right)}\prod_{i=1}^{m}\frac{\Gamma(x_i+\alpha_i')}{x_i!\Gamma(\alpha_i')},$$

where $\alpha_i'=\alpha_i+\sum_{t=1}^{l}x_{i,t}$, and $\mathbf{D}_{t+(k-1)}=\left[\mathbf{x}_1\,\mathbf{x}_2\cdots\mathbf{x}_{t+(k-1)}\right]^T$. Equation 3 is the probability function of the predictive distribution which is the probability mass function of the Dirichlet-multinomial distribution with an expected value

(4) $$E\left[X_{i,t+k}\right]=n\,\theta_i=\;n\,\left[\frac{\alpha_i}{\alpha_0}\right],i=1,2,\cdots,m.$$

*2.4 Statstical Test*

The statistical tests used for data analysis follow Indratno et al. [8]. Let $L_l(\boldsymbol{\theta})=\prod_{t=1}^{l}\left(\frac{n!}{\prod_{i=1}^{m}x_{i,t}!}\prod_{i=1}^{m}\theta_i^{x_{i,t}}\right)$ be a multinomial likelihood function, for $n,m,l\gg 1$, then random variable $\left(\log L_l(\boldsymbol{\theta})-\sum_{t=1}^{l}\log n!\right)$ has a normal distribution with mean $\ddot{\mu}=\sum_{t=1}^{l}\sum_{i=1}^{m}x_{i,t}\log\theta_i-\sum_{t=1}^{l}\sum_{i=1}^{m}\log x_{i,t}!$, and variance $\ddot{\sigma}^2=\frac{m}{(m-1)}\sum_{t=1}^{l}\sum_{i=1}^{m}\left(x_{i,t}\log\theta_i-\frac{\sum_{i=1}^{m}x_{i,t}\log\theta_i}{m}\right)^2-\sum_{t=1}^{l}\sum_{i=1}^{m}\left(\log x_{i,t}!-\frac{\sum_{i=1}^{m}\log x_{i,t}!}{m}\right)^2$. The formula states that the random variable $\left(\log L(\boldsymbol{\theta})-\sum_{t=1}^{l}\log n!\right)\sim N(\ddot{\mu},\ddot{\sigma}^2)$ for testing the fit of the model in the multinomial case with assumptions $\mathrm{H}_0$: $\left(\log L_{l+1}(\boldsymbol{\theta})-\sum_{t=1}^{l+1}\log n!\right)$ follows a normal distribution with mean $\ddot{\mu}_l$ and variance $\ddot{\sigma}_l^2$. The statistical test is

(5) $$z_{stat}=\frac{\left(\log L_{l+1}(\boldsymbol{\theta})-\sum_{t=1}^{l+1}\log n!\right)-\ddot{\mu}}{\ddot{\sigma}},$$

where $\log L_{l+1}(\boldsymbol{\theta})-\sum_{t=1}^{l+1}\log n!$ is random variable under the assumption $\mathrm{H}_0$ [8]. The decision-making criteria $\mathrm{H}_0$ is not rejected when $-z_{\frac{\alpha}{2}}\le z_{stat}\le z_{\frac{\alpha}{2}}$. The margin of error for the confidence interval is $(1-\alpha)$. Let $\left(\log L_{l+1}(\boldsymbol{\theta})-\sum_{t=1}^{l+1}\log n!\right)\sim N(\ddot{\mu}_l,\ddot{\sigma}_l^2)$ be a random variable, where $\ddot{\mu}_l=\sum_{t=1}^{l}\sum_{i=1}^{m}x_{i,t}\log\theta_i-\sum_{t=1}^{l}\sum_{i=1}^{m}\log x_{i,t}!$ and $\ddot{\sigma}_l^2=\frac{m}{(m-1)}\sum_{t=1}^{l}\sum_{i=1}^{m}\left(x_{i,t}\log\theta_i-\frac{\sum_{i=1}^{m}x_{i,t}\log\theta_i}{m}\right)^2-\sum_{t=1}^{l}\sum_{i=1}^{m}\left(\log x_{i,t}!-\frac{\sum_{i=1}^{m}\log x_{i,t}!}{m}\right)^2$, then margin

of error $(\varepsilon)$ for the confidence interval $(1-\alpha)$ *is* $\varepsilon = z_{\left(1-\frac{\alpha}{2}\right)}\ddot{\sigma}_l$.

## 3. RESULTS AND DISCUSSION

The results of data analysis, ranging from network clustering to predictive distribution and statistical tests, in this study using Python. The results of the simulation of weighted network formation and its clustering from the Bandung city map for the flow of health logistics distribution can be seen in Figures 2 to 7. The statistical test results can be seen in Table 1.

The first simulation is network clustering into two subnets or zones; it is divided into two parts. The first is done by generating as much as 14 history data, namely data from Day 1 to Day 14. The history data is used to predict data on Day 15. The second is done by generating as much as 13 historical data, namely data from Day 16 to Day 28. Data from Day 15 to Day 28 is used to predict data on Day 29.

The results of the data analysis based on network clustering and predictive distribution are presented in Figure 2 and Figure 3. The network shows two health logistics distribution zones, the blue zone and the orange zone. These two zones are areas that medical officers will consider in distributing health logistics such as medicines and medical equipment.

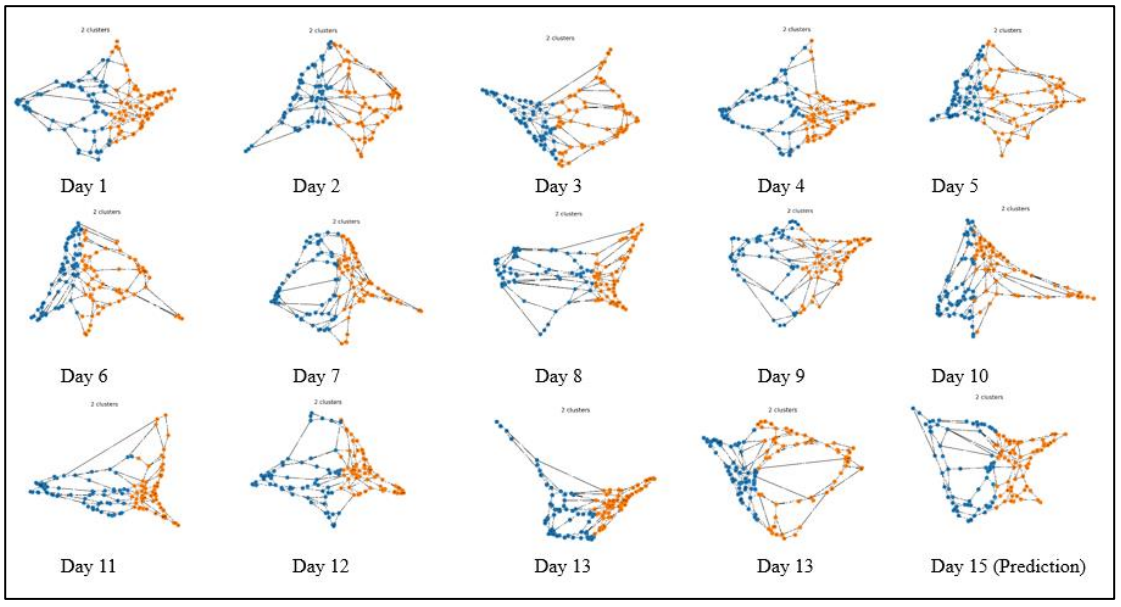


FIGURE 2. The network prediction in the first simulation for two zone

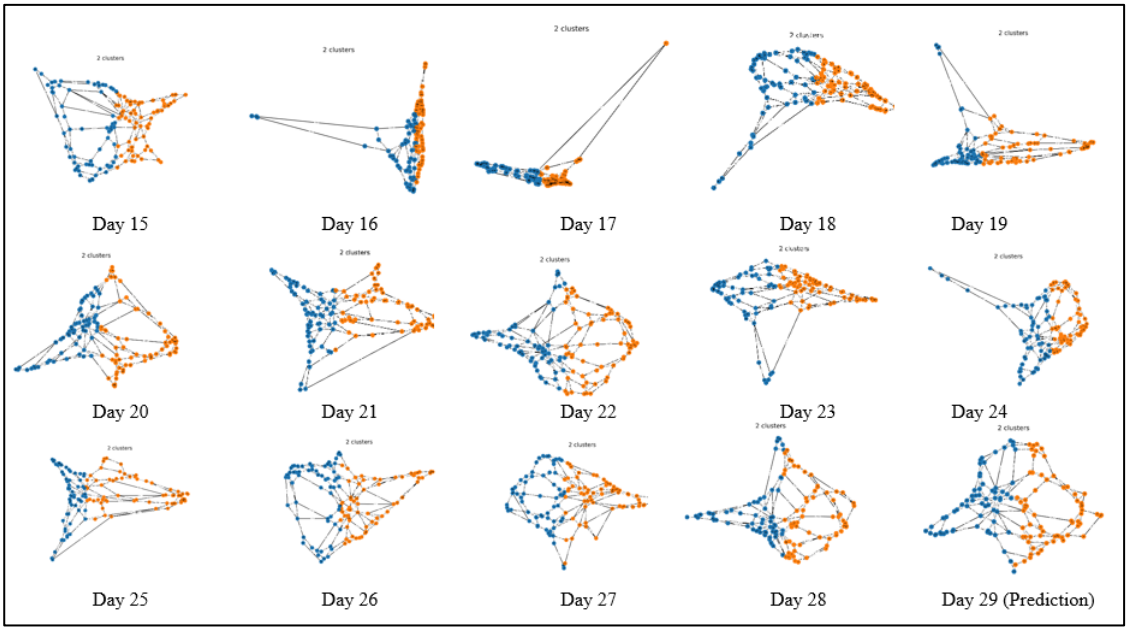


FIGURE 3. The network prediction in the first simulation for two zone after prediction

The second simulation is network clustering into four zones. Like the first simulation, it is divided into two parts. The results of the data analysis based on network clustering and predictive distribution are presented in Figures 4 and 5. In this simulation, the network consists of four health logistics distribution zones: blue zone, orange zone, red zone, and green zone.

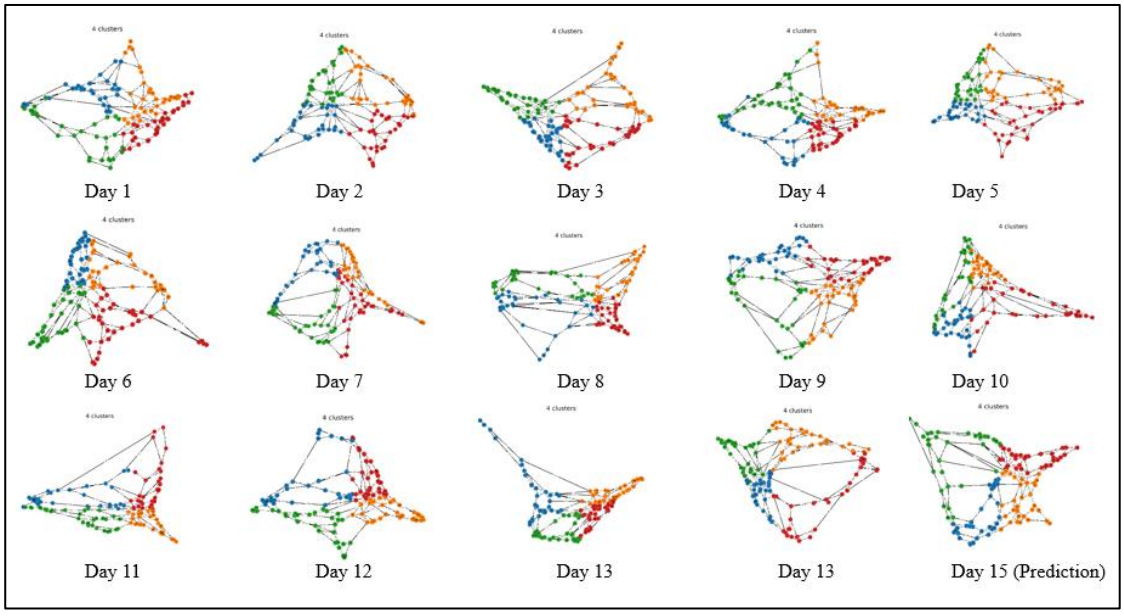


FIGURE 4. The network prediction in the second simulation for four zone

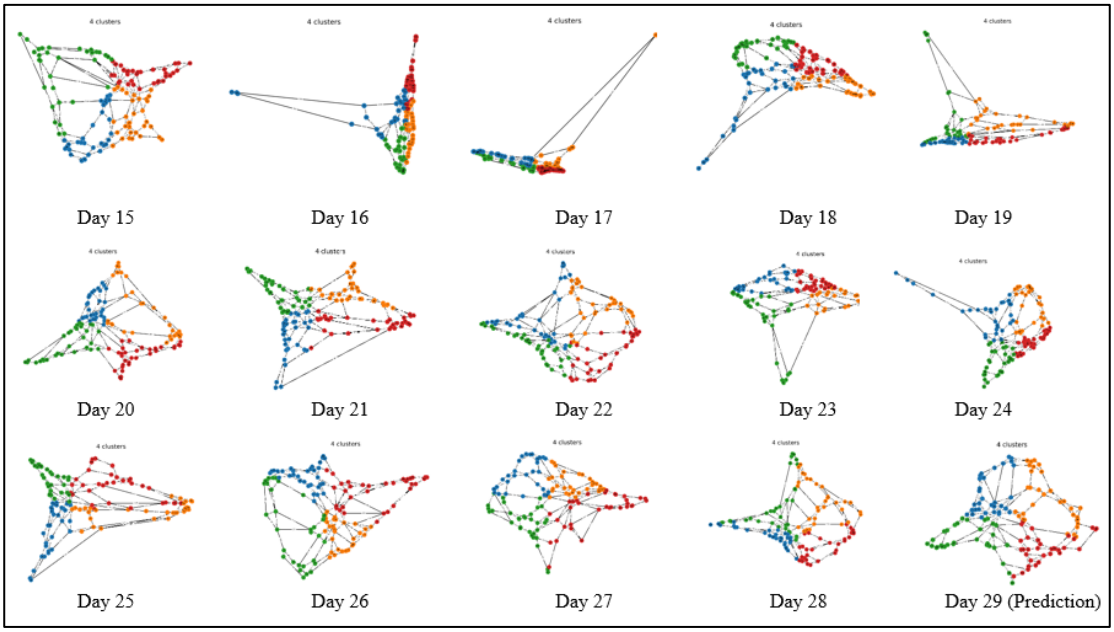


FIGURE 5. The network prediction in the second simulation for four zone after prediction

The third simulation involves network clustering into eight zones. As before, the first and second simulations are divided into two parts. The results of the data analysis based on network clustering and predictive distribution are presented in Figures 6 and 7. In this simulation, the network consists of eight health logistics distribution zones: blue zone, orange zone, red zone, green zone, pink zone, grey zone, brown zone, and purple zone.

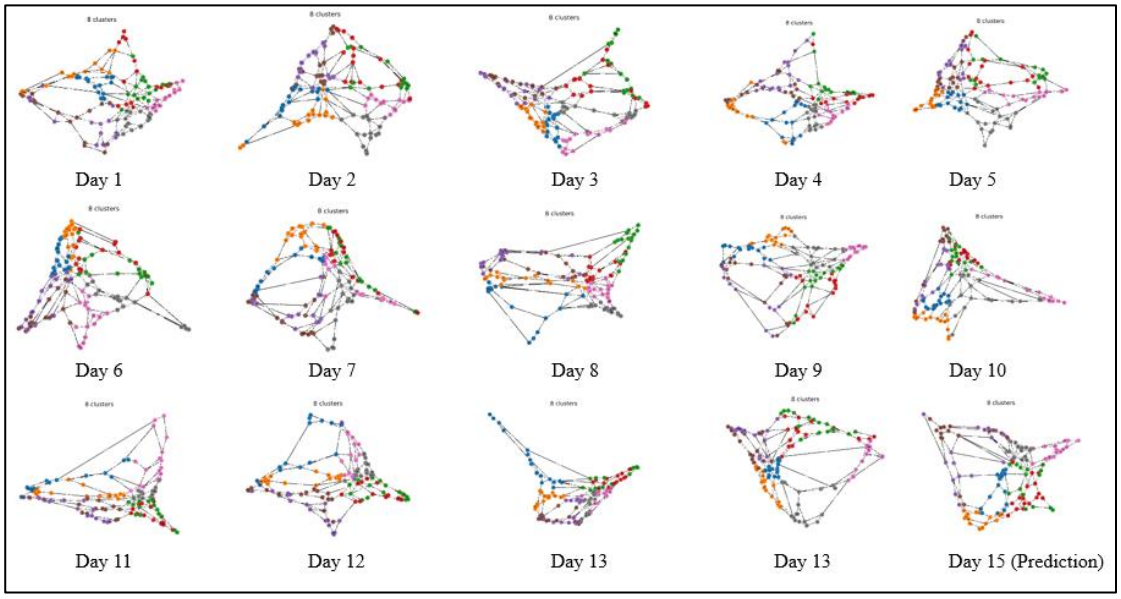


FIGURE 6. The network prediction in the third simulation for eight zone

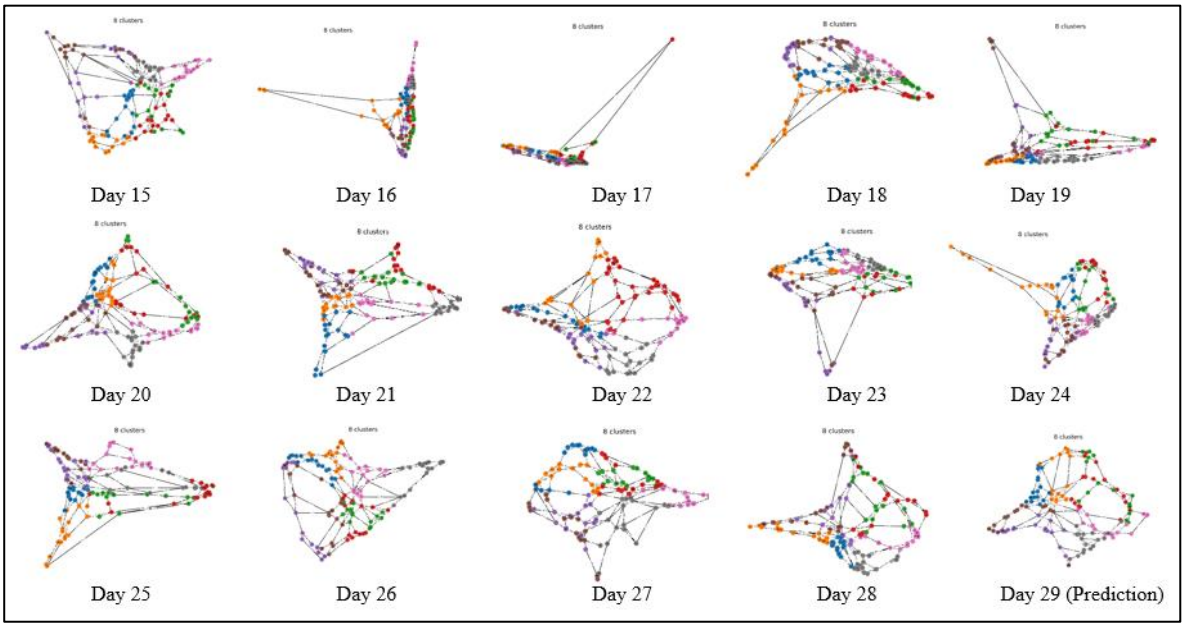


FIGURE 7. The network prediction in the third simulation for eight zone after prediction

Zone prediction benefits medical officers who do not know how to distribute logistics in an area tomorrow, as the data is only limited to today. Figure 8 presents one of the predicted network results.

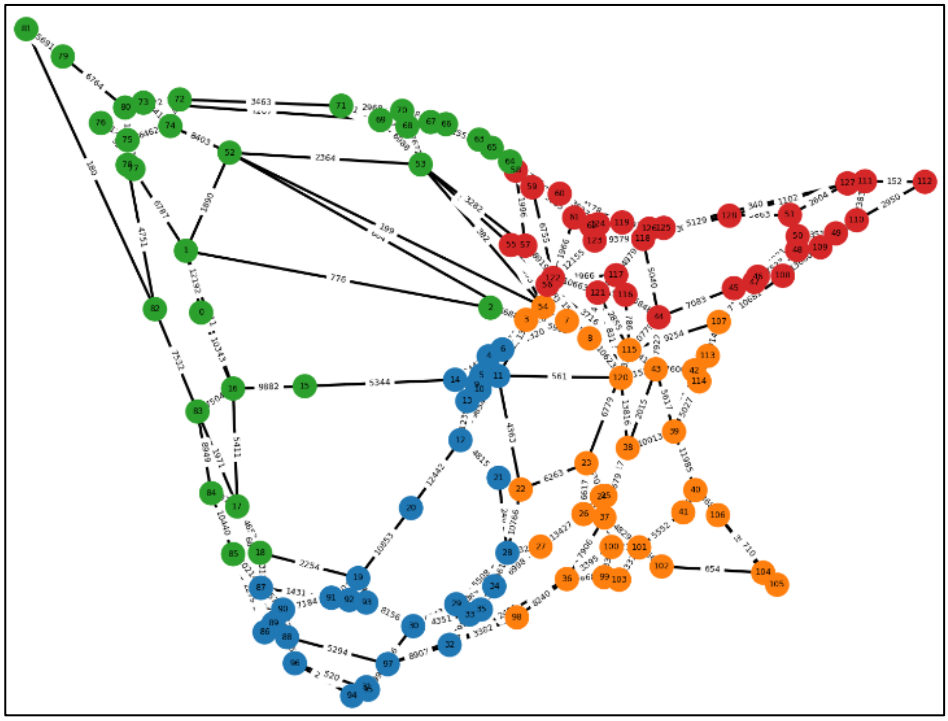

FIGURE 8. Network prediction which is divided into four zones

The statistical test results of the simulation are presented in Table 1.

TABLE 1. Hypothesis testing with $\alpha = 0.05$ and $z_{critical} = 1.96$ or $-1.96$

| Parameter $(\ddot{\mu}_l, \ddot{\sigma}_l^2)$ | Data Analysis | $z_{statitics}$ | $H_0$ | Error $(\varepsilon)$ |
|---|---|---|---|---|
| $D_{1\text{-}14}$ | $D_{1\text{-}14}$ versus $D_{15}$ | $-0.11$ | not rejected | 0.14 |
| $D_{15\text{-}28}$ | $D_{15\text{-}28}$ versus $D_{29}$ | $-0.23$ | not rejected | 0.18 |

$D_{14}$ = Day 14 data, $D_{1\text{-}14}$ = Day 1 to Day 14 data

The unique thing about the results of data analysis is that when more historical data is used to predict, the error is small. The results of this test statistic can be seen in Table 2.

TABLE 2. Hypothesis testing with $\alpha = 0.05$ and $z_{critical} = 1.96$ or $-1.96$

| Parameter $(\ddot{\mu}_l, \ddot{\sigma}_l^2)$ | Data Analysis | $z_{statitics}$ | $H_0$ | Error $(\varepsilon)$ |
|---|---|---|---|---|
| $D_{1\text{-}28}$ | $D_{1\text{-}28}$ versus $D_{29}$ | $-0.02$ | not rejected | 0.09 |

Consider Table 1 and Table 2. In each simulation, the history data used to predict is 14 in Table 1, and the history data used in Table 2 is 28. We can see that the resulting error value in Table 2 is smaller. This finding is consistent with previous studies on the same model [6] [8] [10], so the results of this study can be considered reasonable. Thus, the new method found in this research is expected to support the realization of SDGs in Indonesia.

## 4. CONCLUSION

This new method for predicting networks and the resulting zones will facilitate the process of health logistics delivery in densely populated areas, especially in urban areas. The approach used in this method is straightforward, namely, using network clustering and predictive distribution so that everyone can easily apply it. One of the benefits of the results of this research is that it can be used in tackling the next COVID-19 pandemic and other disease pandemics.

The limitation of this research is that data analysis still uses Python, and special software still needs to be made. So, it requires a high level of accuracy and a lot of time. For future research, the method will be combined by adding a correspondence analysis [13]-[17]. After that, special software will facilitate data analysis with the same cases.

## ACKNOWLEDGMENT

The authors thank LPPM Universitas Singaperbangsa Karawang for the research funding through the “Hibah Bersaing-Penelitian Dasar 2024” Scheme and assistance, i.e., Miss. Julia Permata, Mr. Doni Ramdan, and Mr. Dian Pebriana.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

## REFERENCES

[1] United Nations Economic Commission for Europe, Road map on statistics for sustainable development goals, United Nations, 2017.

[2] P. Jones, M. Wynn, D. Hillier, et al. The sustainable development goals and information and communication technologies, Indones. J. Sustain. Account. Manage. 1 (2017), 1. https://doi.org/10.28992/ijsam.v1i1.22.

[3] United Nations Department of Economic and Social Affairs, Sustainable development goals (SDGs) and the promise of a transformative agenda, United Nations, 2022.

[4] C. Kroll, A. Warchold, P. Pradhan, Sustainable development goals (SDGs): Are we successful in turning trade-offs into synergies?, Palgrave Commun. 5 (2019), 140. https://doi.org/10.1057/s41599-019-0335-5.

[5] M.R. Yudhanegara, K.E. Lestari, Clustering for multi-dimensional data set: a case study on educational data, J. Phys.: Conf. Ser. 1280 (2019), 042025. https://doi.org/10.1088/1742-6596/1280/4/042025.

[6] M.R. Yudhanegara, K.E. Lestari, Network clustering method for preventing the spread of COVID-19 in Indonesian schools, Commun. Math. Biol. Neurosci. 2023 (2023), 34. https://doi.org/10.28919/cmbn/7922.

[7] M.R. Yudhanegara, S.W. Indratno, R.K.N. Sari, Clustering for Item Delivery Using Rule-K-Means, J. Indones. Math. Soc. 26 (2020), 185–191. https://doi.org/10.22342/jims.26.2.871.185-191.

[8] S.W. Indratno, K.N. Sari, M.R. Yudhanegara, Optimization in item delivery as risk management: Multinomial case using the new method of statistical inference for online decision, Risks 10 (2022), 122. https://doi.org/10.3390/risks10060122.

[9] M.R. Yudhanegara, S.W. Indratno, R.R.K.N. Sari, Clustering for items distribution network, J. Phys.: Conf. Ser. 1496 (2020), 012019. https://doi.org/10.1088/1742-6596/1496/1/012019.

[10] M.R. Yudhanegara, S.W. Indratno, RR.K.N. Sari, Dynamic items delivery network: prediction and clustering, Heliyon 7 (2021), e06934. https://doi.org/10.1016/j.heliyon.2021.e06934.

[11] H.D. Bedru, S. Yu, X. Xiao, et al. Big networks: A survey, Computer Sci. Rev. 37 (2020), 100247. https://doi.org/10.1016/j.cosrev.2020.100247.

[12] U. Elsner, Graph partitioning-a survey, Technical Report, Technische Universität Chemnitz, 1997.

[13] K.E. Lestari, M.R. Utami, M.R. Yudhanegara, Sequential exploratory design by performing correspondence analysis to investigate procedural fluency of undergraduate student, AIP Conf. Proc. 2588 (2023) 050004. https://doi.org/10.1063/5.0111974.

[14] K.E. Lestari, U.S. Pasaribu, S.W. Indratno, et al.The comparative analysis of dependence for three-way contingency table using Burt matrix and Tucker3 in correspondence analysis, J. Phys.: Conf. Ser. 1245 (2019), 012056. https://doi.org/10.1088/1742-6596/1245/1/012056.

[15] K.E. Lestari, U.S. Pasaribu, S.W. Indratno, Graphical depiction of three-way association in contingency table using higher-order singular value decomposition Tucker3, J. Phys.: Conf. Ser. 1280 (2019), 022035. https://doi.org/10.1088/1742-6596/1280/2/022035.

[16] K.E. Lestari, M.R. Utami, M.R. Yudhanegara, Exploratory analysis on adaptive reasoning of undergraduate student in statistical inference, Int. J. Instruction 15 (2022), 535–554. https://doi.org/10.29333/iji.2022.15429a.

[17] K.E. Lestari, M.R. Utami, M.R. Yudhanegara, Empirical study of mathematical investigation skill on graph theory, Mat. Teach. Res. J. 16 (2024), 4-27.