Journal of
Information Systems Engineering
and Business Intelligence

Vol.9, No.1, April 2023
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

The Impact of Socioeconomic and Demographic Factors on COVID-19 Forecasting Model

Siti Nur Hasanah [1)*], Yeni Herdiyeni [2)], Medria Kusuma Dewi Hardhienata [3)]

[1)2)3)]*Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Indonesia*
Jalan Agatis Kampus IPB Darmaga, Bogor
[1)]siti.nurhasanah@apps.ipb.ac.id, [2)]yeni.herdiyeni@apps.ipb.ac.id, [3)]medria.hardhienata@apps.ipb.ac.id

Abstract

Background: COVID-19 has become a primary public health issue in various countries across the world. The main difficulty in managing outbreaks of infectious diseases is due to the difference in geographical, demographic, economic inequalities and people's behavior in each region. The spread of disease acts like a series of diverse regional outbreaks; each part has its disease transmission pattern.
Objective: This study aims to assess the association of socioeconomic and demographic factors to COVID-19 cases through cluster analysis and forecast the daily cases of COVID-19 in each cluster using a predictive modeling technique.
Methods: This study applies a hierarchical clustering approach to group regencies and cities based on their socioeconomic and demographic similarities. After that, a time-series forecasting model, Facebook Prophet, is developed in each cluster to assess the transmissibility risk of COVID-19 over a short period of time.
Results: A high incidence of COVID-19 was found in clusters with better socioeconomic conditions and densely populated. The Prophet model forecasted the daily cases of COVID-19 in each cluster, with Mean Absolute Percentage Error (MAPE) of 0.0869; 0.1513; and 0.1040, respectively, for cluster 1, cluster 2, and cluster 3.
Conclusion: Socioeconomic and demographic factors were associated with different COVID-19 waves in a region. From the study, we found that considering socioeconomic and demographic factors to forecast COVID-19 cases played a crucial role in determining the risk in that area.

Keywords: COVID-19, Facebook Prophet , Hierarchical clustering, Socioeconomic and demographic

Article history: Received 29 December 2022, First decision 10 February 2023, Accepted 21 February 2023, Available online 28 April 2023

I. INTRODUCTION

The infection of the SARS-CoV-2 virus causes a highly contagious disease called Coronavirus Disease 2019 or COVID-19 [1]. The first case of COVID-19 was identified in Wuhan City, Hubei Province, China, as pneumonia with an unknown cause in December 2019 [2], [3]. Belonging to the family of respiratory viruses, COVID-19 has a high transmissibility risk [4]. Based on [4], COVID-19 can be transmitted through four major modes of transmission: direct (physical) contact, indirect contact (contact with contaminated surfaces or objects), droplet, and aerosol. On March 11, 2020, the World Health Organization (WHO) announced COVID-19 as a global pandemic [5]. As of February 15, 2023, COVID-19 has infected 677,904,870 people and caused death to 6,784,394 people in the world [6].

Many factors have impacted the transmissibility of infectious disease, not only to be a trait of the biological pathogen but also the result of socioeconomic and demographic factors of an area. [7]. Socioeconomic and demographic factors may affect several aspects of human life, including health-related aspects [8]. According to [9], social structures such as income and poverty play an essential role in determining COVID-19 cases in several European countries. The study from [10] used machine learning techniques such as Random Forest, Quantile Regression, and Hierarchical Regression to understand how socioeconomic and demographic factors affect COVID-19. They found that population density, the proportion of women, and commuting time were the three main factors controlling the spread of COVID-19. Moreover, [11] investigated the statistical association between the prevalence of COVID-19 with socioeconomic, demographic, and health indicators using bivariate and multivariate analysis. The study from [11] found that the prevalence of COVID-19 was negatively associated with the gender ratio and positively associated with life expectancy.

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.9.1.70-83

The differences in socioeconomics and demographics in each region become one of the difficulties in managing infectious diseases. In a forecasting model, the national-level projection is dominated by the number of cases from big and highly populated regions [6]. Many predictions for rural, semi-rural, and small populations produce over or under-forecasting outbreaks due to data skewed toward urban areas [6]. Building a forecasting model that considers the socioeconomic and demographic factors will produce specific information about COVID-19 in the area. Regencies and cities may be seen as "building blocks" for a province based on socioeconomic and demographic characteristics. As a result, clustering techniques are suitable for this task. Regencies or cities in a province can be grouped into clusters based on socioeconomic and demographic similarities. Groups of people with the same demographic, economic, and social status tend to behave similarly because of cultural similarities. Different algorithms, such as hierarchical clustering, can be applied to group the data. Hierarchical clustering has been applied in many fields; authors of [12] applied the hierarchical clustering algorithm in customer segmentation to create a credit card company's marketing strategy. The study from [13] applied hierarchical clustering to analyze the socioeconomic status (SES) of immigrants in Chile. Then [14] applied a hierarchical clustering approach to assessing the relationship between socioeconomic with cardiovascular and metabolic health. Another study by [15] utilized the hierarchical clustering technique to identify the cluster of regions with the same COVID-19 epidemic pattern in Italy. Hierarchical clustering is a technique to group objects based on pairwise distances in an agglomerative (bottom-up) and divisive (top-down) manner [15]. Each observation starts in its cluster with agglomerative clustering, which joins the most similar two clusters into one cluster to create a hierarchy [6]. In this study, we use agglomerative hierarchical clustering because it gives specific details on which observations are most similar. Dendrogram, the algorithm's output, can be utilized to understand the big picture and the groups in the data.

Since COVID-19 began to spread in early 2020, researchers around the world have used various modeling techniques to predict future COVID-19 cases [16]–[21]. Accurate prediction of future cases can be used to anticipate the dispersal of the disease. Machine learning, statistical methods, and deep learning are widely utilized for predicting epidemic evolution [22]. These models use past data to predict future trends and can be adapted to forecast COVID-19 as the cases are reported daily. Based on [23], among time-series models, Facebook Prophet, or called FBProphet, has been utilized in numerous applications across various fields due to its high accuracy in forecasting/prediction. A study from [24] used the Prophet model to predict the positive and fatality of COVID-19 in India during and after the lockdown period. This study shows that the model performs better in predicting COVID-19 during lockdown with 87% accuracy, while after the lockdown relaxation, the model's accuracy dropped to 60%. The study from [25] compared Support Vector Machines, Linear Regression, and FBProphet for predicting COVID-19 in India. Results in [25] have shown that the Prophet method is highly effective in predicting active, death, and cure rates compared with other techniques.

Most studies related to COVID-19 that have been conducted so far mainly focused on finding how socioeconomic and demographic factors affect COVID-19, such as the study from [9]–[11],[26] or only focus on forecasting the daily cases of COVID-19 such as the study in [16]–[21]. Research that combines both analyses has not been widely studied. We combine both analyses for the work discussed in this study by conducting a two-phased modeling approach. First, we collected several regency/city-level socioeconomic and demographic datasets. Using the agglomerative hierarchical clustering approach, we intend to use this data to group regencies/cities into large aggregations based on their socioeconomic and demographic similarities. We will analyze the relationship between socioeconomic and demographic factors to COVID-19 in each group of regencies/cities. Then, we count the daily cases in each group and build a forecasting model using FBProphet to assess how COVID-19 will spread if current restrictions are maintained. From this study, we can identify the most critical regency/city-level socioeconomic variables related to COVID-19 transmission and group individual regencies/cities into clusters based on socioeconomic characteristics. In each cluster, through the forecasting model, we assess the transmissibility of COVID-19 over a short period of time and find the area with high risk where public health measures should be more focused. By conducting these two analysis steps, we hope it can provide better information about COVID-19. Further applications of this study can be used as a principle for the area decision-support systems to assist direct responses to pandemics and to prepare preventative measures against potential future epidemics.

II. METHODS

This study is divided into three main stages: data collection, analysis of the relationship between socioeconomic-demographic factors and COVID-19 using hierarchical clustering, and forecasting COVID-19 in each cluster using FBProphet, as shown in Fig. 1. The main stages of this study are described in detail in the subsections.

A. Stage 1: Data Collection

In this study, we used data from East Java, one of the provinces in Indonesia. The COVID-19 data were retrieved from April 21, 2020, to November 28, 2022, from the daily reports of the East Java Government [27]. Each regency's socioeconomic and demographic data were collected from the annual report of the Central Bureau of Statistics of East Java Province [28]. The socioeconomic and demographic variables used in this study are based on the variables used in several previous studies [11], [26], [29], [30]. But not all of them are used; the final variables are adjusted to the available data in East Java Province. Finally, 17 socioeconomic and demographic variables were identified to group regencies/cities and analyze the relationship between socioeconomic-demographic factors with COVID-19. Further information about the variables is given in Table 1.

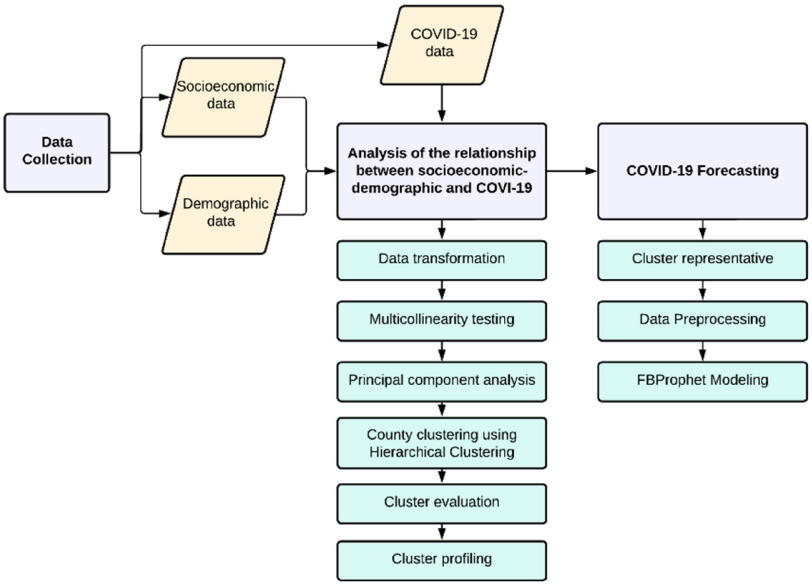

Fig. 1 The research flowchart

TABLE 1
LIST OF SOCIOECONOMIC AND DEMOGRAPHIC VARIABLES

No.	Variable	Description
1	Tot_pop	The number of populations by county
2	Pop_density	Population density is the number of individuals per unit geographic region
3	%_elderly_pop	Percentage of the population aged 65 years or over per total population
4	Gender_ratio	The ratio between the number of males and females
5	%_higher_ed	Percentage of the population with a college degree
6	Reg_min_wage	The minimum wage by county
7	Num_hospital	The number of hospitals
8	Unemployment_rate	The ratio between job seekers and the workforce
9	HDI	Human development index
10	GDI	Gender development index
11	%_poor_pop	Percentage of the poor population
12	Expenditure_per_cap	The household's final consumption of goods and services for one month per total number of the household members
13	%_beneficiary_pop	The percentage of the family who receives government assistance budget
14	GDP_reg	Gross domestic product regional
15	%_urban_pop	The number of urban populations per total population
16	Num_of_comp_LM	Number of companies in large and medium industries
17	Num_of_comp_MS	Number of companies in micro and small industries

B. Stage 2: Analysis of the Relationship of Socioeconomic-Demographic Factors with COVID-19

East Java is the largest province on Java Island; it is very diverse demographically and geographically. Hence, it is worthwhile to analyze the disease's spread based on regions in this province with similar socioeconomic and demographic characteristics. Several steps in examining the relationship between socioeconomic and demographic with COVID-19 are as follows [6], [15], [12]:

1) Data transformation

The socioeconomic and demographic data have different measurement units, so data transformation is needed to ensure that no variable dominates another. In this study, we apply standardization to ensure the variables are rescaled and follow the normal distribution Z~N (0.1). We use the formula in (1) [12] to standardize the data.

$$Z_{ij} = \frac{x_{ij} - \bar{x}_i}{S_{xi}} \quad (1)$$

Where Z_{ij} is the data in standardization format, x_{ij} is the j[th] data on the i[th] variable, $\bar{x}_i$ is the mean of the i[th] variable, and S_{xi} is the standard deviation of the i[th] variable.

2) Multicollinearity testing

Most socioeconomic and demographic data contain correlated information. Therefore, we filter the socioeconomic and demographic data by applying multicollinearity testing. Multicollinearity is a higher linear relationship between variables in data [31]. In multivariate analysis, multicollinearity will cause a destructive impact. Based on [32], when multicollinearity increases, it will be hard to interpret the effect of each variable because of the relationship between them. In this study, we identify multicollinearity by examining the correlation matrix of the variables and calculating the Variance Inflation Factor (VIF) of each variable. The formula to calculate the correlation could be represented as (2) [33].

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}} \quad (2)$$

where r is the correlation coefficient, n is the number of observations, X is the first variable, and Y is the second variable. According to [34], if the value of $|r| \geq 0{,}7$, indicate a strong linear relationship in the data. We also examine the VIF value of each variable that could be calculated as (3) [35].

$$VIF = \frac{1}{1 - R^2} \quad (3)$$

where R^2 is the r-squared value from a regression between variable X and other variables. Based on [35], the value of $VIF \geq 10$, indicates multicollinearity in the data. There are several ways to solve the multicollinearity problem such as excluding the variable that indicates collinearity or performing principal component analysis [31]. This study uses a principal component analysis to deal with the multicollinearity problem.

3) Principal Component Analysis

Principal component analysis, or PCA, is a method to transform a set of correlated observations into uncorrelated variables [36]. In PCA, we try to create several new variables, which are the combinations of the initial variables. The new variables are named principal components (PCs). We calculate the eigenvector and eigenvalue from the covariance matrix to obtain PC. The eigenvalue is computed using (4) [37].

$$|C - \lambda I| = 0 \quad (4)$$

where C is a $M \times M$ covariance matrix, M is the total of variables, λ is a $M \times 1$ eigenvalue matrix, and I is an $M \times M$ identity matrix. Then, the eigenvector can be calculated by substituting the eigenvalue obtained into (5) [37].

$$Ca - \lambda a = 0 \quad (5)$$

where a is a $M \times 1$ eigenvector matrix.

4) Regency/City Clustering Using Hierarchical Clustering

The obtained principal components from the previous step will be used as input for clustering regencies/cities. In this study, we choose Agglomerative Hierarchical Clustering to group regencies/cities in East Java. We select the Ward method to determine the object that will be combined to form a cluster. We need to calculate SSE using (6) [38] in the Ward method.

$$SSE = \left(x_j - \bar{x}\right)'\left(x_j - \bar{x}\right) \quad (6)$$

where x_j is the data of the j[th] multivariate object, and $\bar{x}$ is the mean of all objects. When merging two clusters, for example, clusters P and Q, in Ward's Linkage method, we try to minimize the increase in SSE, defined as the distance between clusters P and Q. This distance is mathematically expressed as (7) [39].

$$I_{(PQ)} = SSE_{PQ} - \left(SSE_P + SSE_Q\right) \quad (7)$$

5) Cluster Evaluation

To evaluate the clustering result, we calculate the cophenetic coefficient. According to [40], the cophenetic coefficient could be represented as (8).

$$c = \frac{\sum_{i=1}^{j}(x_{(i,j)}-\bar{x})(t_{(i,j)}-\bar{t})}{\sqrt{\left[\sum_{i=1}^{j}(x_{(i,j)}-\bar{x})^2\right]\left[\sum_{i=1}^{j}(t_{(i,j)}-\bar{t})^2\right]}} \quad (8)$$

where $x_{(i,j)}$ is the distance of the i[th] object to the j[th] object, $t_{(i,j)}$ is the cophenetic distance of the i[th] object and the j[th] object, $\bar{x}$ is the mean of $x_{(i,j)}$, and $\bar{t}$ is the mean of $t_{(i,j)}$. To determine the number of clusters, we utilize silhouette analysis. In silhouette analysis, we compare an object's similarity in its cluster to other clusters. The silhouette coefficient is defined as (9) [41].

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (9)$$

where $s(i)$ is the ith silhouette coefficient, $a(i)$ is the average distance between the ith object and other objects in cluster a, $b(i)$ is the average distance nearest other clusters.

6) Cluster Profiling

As the final step in analyzing the relationship between socioeconomic-demographic and COVID-19, we will summarize the profiles of each cluster. We categorize the socioeconomic-demographic and COVID-19 data into five categories: Very High (VH), High (H), Medium (M), Low (L), and Very Low (VL). The categorization is based on the mean and standard deviation of the data. In this step, we will analyze the spread of COVID-19 in each cluster and find the specific socioeconomic-demographic character of each cluster.

C. Stage 3: COVID-19 Forecasting

The processes for forecasting daily cases of COVID-19 in each cluster are based on the procedures in [24], which are as follows:

1) Cluster Representative

Each cluster consists of more than one time-series data, while only one data set can be modeled in forecasting. For this reason, it is necessary to form data representing the cluster. In this study, we calculate the daily average as the representative of data in a cluster. The calculation process follows (10).

$$Rep_i = average\{x_1, x_2, \dots, x_k\} \quad (10)$$

where i is the length of the data and k is the number of cluster members. This representation data will be used for forecasting COVID-19 at the cluster level.

2) Data Pre-processing

In data pre-processing, we perform two techniques: data smoothing and transformation. The daily cases of COVID-19 in East Java often drop to a relatively low level or even zero on specific days during the standard period. We utilize a rolling average approach to minimize the noise in time-series data. Rolling average implies replacing the value of time-series data on time t with the average value of its n neighbors. We will use n to be 7 in this study. As the final technique in the pre-processing step, we need to transform the data to minimize the skewness of the data and make the distribution close to a normal distribution. We apply the Box-Cox transformation to the daily COVID-19 data. Equation (11) calculates the transformation as follows [42]:

$${y_t}^{(\lambda)} = \begin{cases} \frac{{y_t}^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log y_t & (\lambda = 0) \end{cases} \quad (11)$$

where ${y_t}^{(\lambda)}$ is data after transformation, and y_t is data at time t. The transformed data will be utilized as the input for forecasting COVID-19 using the FBProphet model.

3) FBProphet Modelling

FBProphet is a forecasting technique developed by the Data Scientist team of Facebook. This technique comprises three main components: trend, seasonality, and holiday. According to [43], FBProphet is formulated as (12).

$$y(t) = g(t) + s(t) + h(t) + \in_t \quad (12)$$

where $g(t)$ is a trend function that models non-periodic changes in time-series data, $s(t)$ is a seasonality function that represents the periodic changes in data (e.g., daily, weekly, or yearly), $h(t)$ is the holiday's effect on the data, and $\in_t$ is the other changes that are not accommodated in the model. The transformed data in each cluster will be utilized as the input for modeling. We apply hyperparameter tuning with the grid search method to find the best value of each

parameter in the model. The model's performance is measured by calculating the Mean Absolute Percentage Error (MAPE). According to [44], the mathematical expression of MAPE is shown in (13).

$$MAPE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t / y_t|}{n} \times 100 \quad (13)$$

where y_t is the actual value, and $\hat{y}_t$ is the predicted value. The model with the best parameter value will forecast daily COVID-19 cases in each cluster for the next 30 days.

III. RESULTS

A. Data Exploration

Fig. 2 shows the daily cases of COVID-19 in East Java recorded from April 21, 2020, to November 28, 2022. The blue color presents the confirmed cases, and the orange represents confirmed cases on a 7-day rolling average. The daily cases increased dramatically after June 2021 and continued throughout September. Then we started to see a downward trend toward January 2022. A couple of weeks after that, the cases begin to rise again. The daily cases of COVID-19 in East Java peaked on February 17, 2022, with 8977 cases. The 7-day rolling average did not change the pattern of the disease, but it smooths the data and distinguishes the sequence's noise.

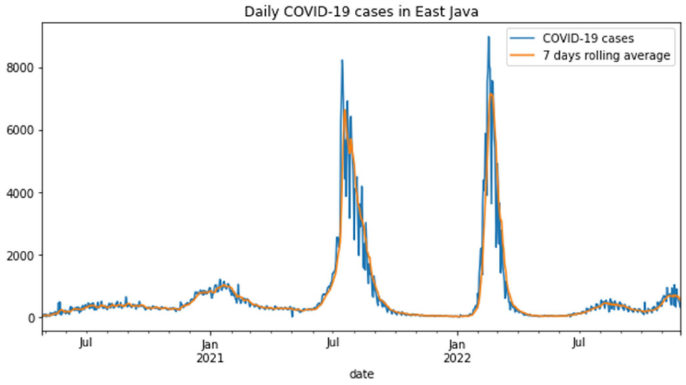

Fig. 2 The daily cases of COVID-19 in East Java

Based on the total cases of COVID-19 on November 28, 2022, the regencies/cities in East Java categorize into five categories. The category of each regency/city is shown in Fig. 3. Surabaya City has the highest total amount of COVID-19 cases, with a total of 142.487 cases. Meanwhile, Pamekasan Regency has the lowest cases, totaling to 3306.

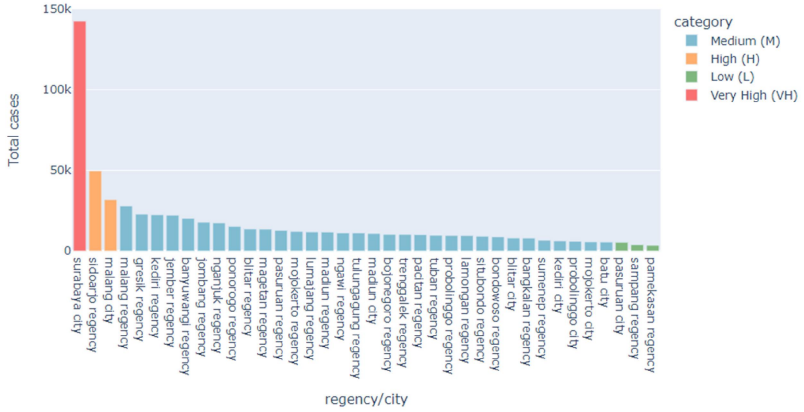

Fig. 3 The category of COVID-19 in Each Regency/City

B. Analysis of Socioeconomic-Demographic Factors with COVID-19

Most of the socioeconomic and demographic data in this study were correlated with each other. Some variables have correlation coefficients greater than 0.7 and VIF values greater than 10. For example, population density is highly correlated with some variables, such as the percentage of people with college degrees with a correlation coefficient of

0.91. Population density also has a VIF value of 25,895. These conditions indicate multicollinearity in the data. The correlation matrix and the VIF value of each variable are shown in Fig. 4 and Table 2.

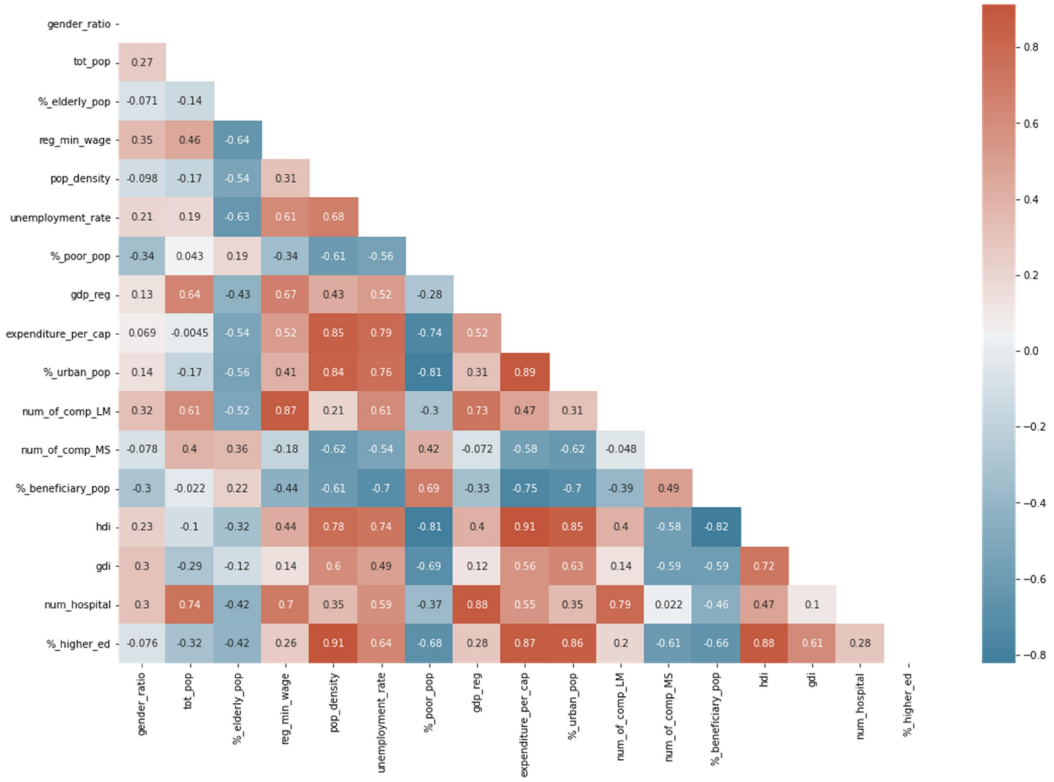

Fig. 4 The correlation matrix of socioeconomic and demographic data

TABLE 2
THE VARIANCE INFLATION FACTOR (VIF) OF SOCIOECONOMIC AND DEMOGRAPHIC DATA

Features	VIF
%_higher_ed	38,599
HDI	38,449
Pop_density	25,895
Expenditure_per_cap	20,833
Num_hospital	18,062
%_urban_pop	15,648
Num_comp_LM	11,525
Tot_pop	10,438
Gdp_reg	9,441
Reg_min_wage	9,056
%_poor_pop	7,593
%_elderly_pop	7,187
GDI	6,743
Unemployment_rate	6,087
%_beneficiary_pop	4,426
Num_of_comp_MS	3,102
Gender_ratio	2,911

To address the multicollinearity problem, we applied the principal component analysis. Principal component analysis tries to reduce the strongly correlated variables into new uncorrelated ones (principal components). We inspect the Scree plot in Fig. 5 to choose the number of principal components (PCs) to retain. PC1 described the most considerable variance among all PCs (51.05%); it highly correlated with most variables such as expenditure_per_cap, %_urban_pop, HDI, %_higher_ed, pop_density, unemployment_rate, %_poor_pop and %_beneficiary_pop. The second PC could explain about 20.9 % of the variance and highly correlated with tot_pop and num_of_comp_LM. We retained only two first PC that cumulatively explained 71.96% of all variance.

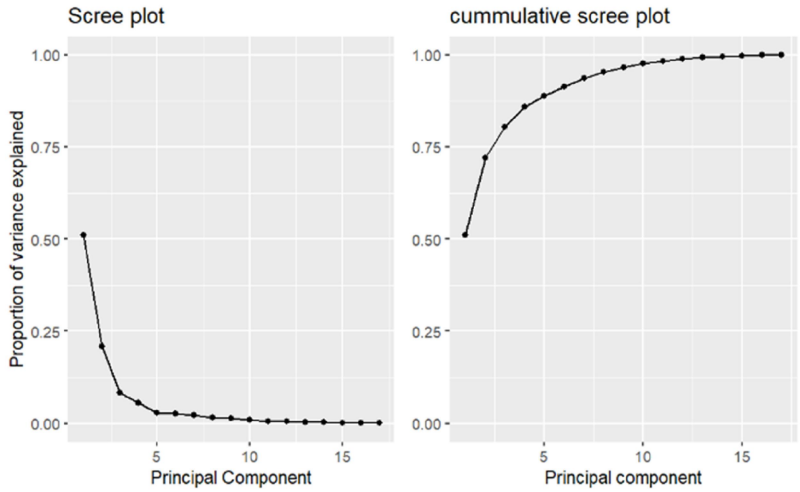

Fig. 5 The scree plot of the principal component

The two PCs were utilized as the input for the clustering process. The whole clustering process is depicted through the dendrogram in Fig. 6. Using the cophenetic correlation coefficient, we evaluated the goodness of Euclidean distance in measuring the dissimilarity between regencies/cities. The cophenetic coefficient in this study was 0.78, implying a strong correlation between the Euclidean distance and the resulting cophenetic distance. Based on the dendrogram in Fig. 6, the number of clusters formed may be from 2 to 38. The optimal number of clusters can be determined through the silhouette coefficient in Fig. 7.

Based on Fig. 7, forming three clusters will produce the most robust clusters marked by the highest silhouette coefficient. So, the regencies/cities in East Java will be grouped into three clusters based on the similarity in the socioeconomic and demographic characteristics. We visualize the three clusters through the biplot in Fig. 8. Each cluster is associated with specific socioeconomic and demographic variables. Most regencies in cluster 1 are positively correlated with variables such as %_elderly_population, %_beneficiary_pop, %_poor_pop, etc. The city and regency in cluster 2 highly correlate with total_pop, gdp_reg, num_hospital, num_of_comp_LM, reg_min_wage, etc. Meanwhile, the cities in cluster 3 are positively associated with variables such as GDI, %_higher_ed, etc.

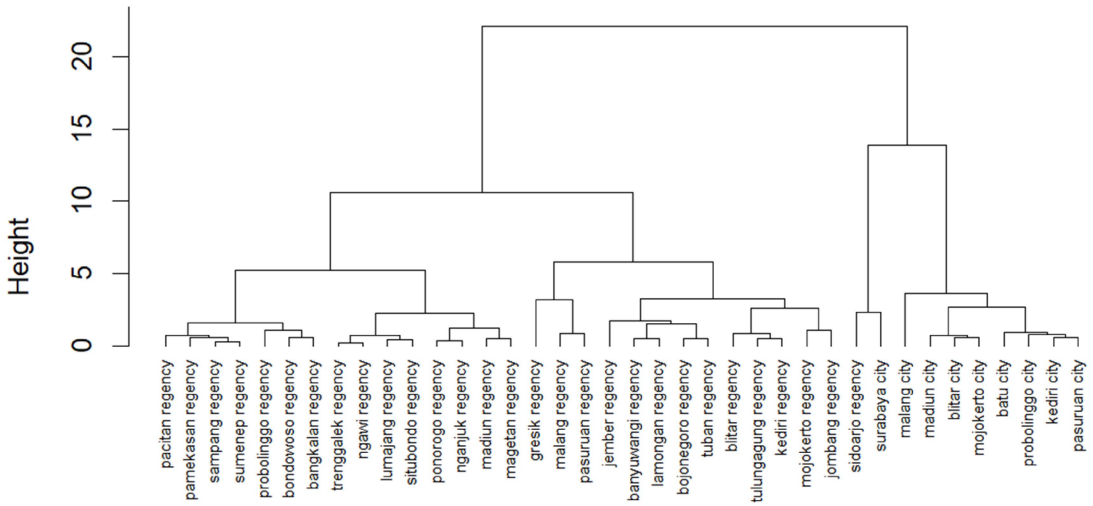

Fig. 6 The dendrogram of agglomerative hierarchical clustering

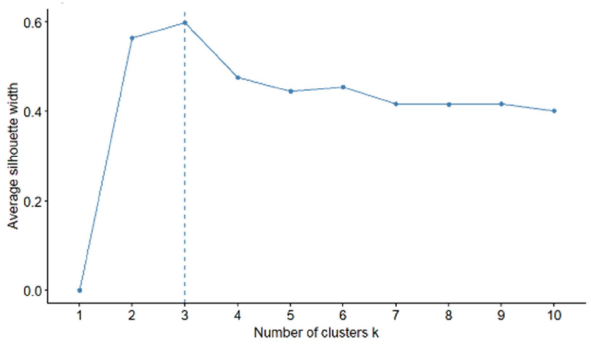

Fig. 7 Silhouette coefficient

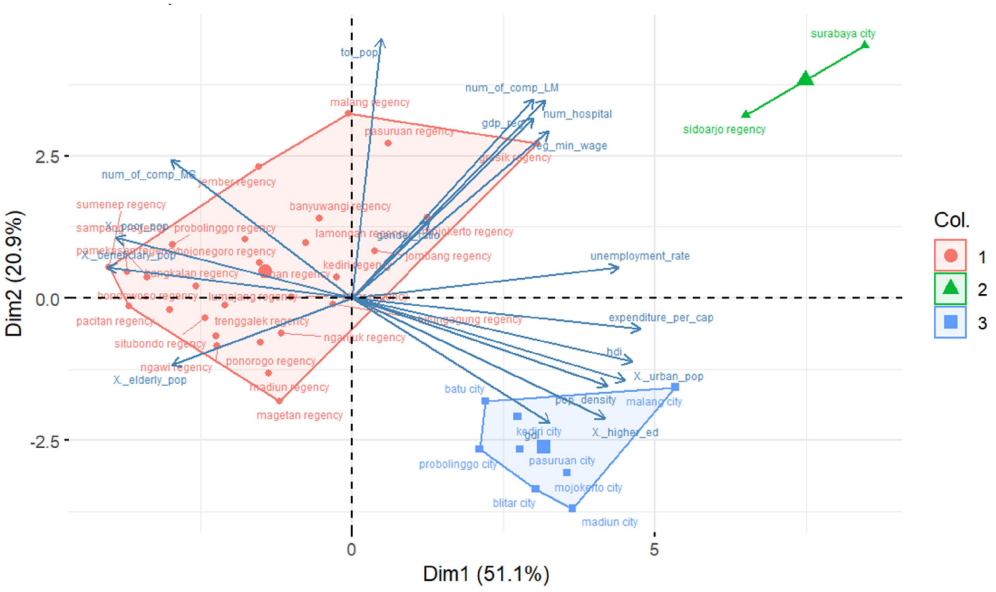

Fig. 8 The biplot of the cluster

Table 3 shows the complete profile of each cluster. The table shows the average cluster value for all socioeconomic, demographic, and COVID-19 data. We also presented the category of each data.

TABLE 3
CLUSTER PROFILE

Features	Cluster 1	Cluster 2	Cluster 3
Number of regencies/cities	28	2	8
Gender_ratio	99.28 (M)	99.83 (H)	99.01 (L)
Tot_pop	1,200,990.71 (M)	2,486.10 (H)	284,853.,75 (L)
%_elderly_pop	14.81 (H)	8.67 (L)	11.77 (M)
Reg_min_wage	2,381,035.08 (L)	4,372,030.52 (H)	2,462,286.07 (L)
Pop_density	814.64 (L)	5,757.50 (H)	4,892.12 (M)
Unemployment_rate	4.72 (L)	10.27 (H)	7.12 (M)
%_poor_pop	13.17 (H)	5.58 (L)	6.26 (M)
Gdp_reg	33,404.03 (L)	274,363.58 (H)	23,273.95 (L)
Expenditure_per_cap	941,606.50 (L)	1,827,782.5 (H)	1.503,522 (M)
%_urban_pop	41.37 (L)	99.21 (H)	99.36 (H)
Num_of_comp_LM	129,78 (M)	914,50 (H)	51,87 (L)
Num_of_comp_MS	27,193.35 (H)	14.730.50 (M)	4,819.87 (L)
%_beneficiary_pop	28.08 (H)	10,,41 (L)	17,15 (M)
HDI	69,88 (L)	81.48 (H)	78.10 (M)
GDI	89.52 (L)	94.08 (M)	94.60 (H)
Num_hospital	7.43 (L)	31 (H)	6 (L)
%_higher_ed	4.59 (L)	12.54 (H)	12.31 (H)
Covid-19 per total population	11,1497 (L)	36,7012 (H)	35,2997 (H)

A brief description of each cluster profile is as follows:

1) Cluster 1

Cluster 1 has the highest percentage of the elderly population, poor population, the number of companies in micro and small industries, and the percentage of the beneficiary population. With this characteristic of socioeconomic and demographic, the average COVID-19 per capita in this cluster is categorized as low. Based on Fig.3, the total cases of COVID-19 in most of the members of this cluster are classified as low and medium.

2) Cluster 2

Cluster 2 consists of only two members, Surabaya City, and Sidoarjo Regency. This cluster has a significantly different cluster than the other clusters. Cluster 2 has the highest average of most socioeconomic data, such as regional minimum wage, GDP regional, expenditure per capita, the number of companies in large and medium industries, HDI, and percentage of the population with higher education. This cluster also has the highest total population, population density, and unemployment rate. Cluster 2 is categorized as high based on the COVID-19 cases per capita. Based on Fig.3, the total cases of COVID-19 in cluster 2 are ranked as high and very high.

3) Cluster 3

The members of cluster 3 are all cities in East Java, except for Surabaya city. Cluster 3 has the lowest total population but the second highest population density. Most of this cluster's socioeconomic and demographic data are categorized in the medium. Based on Fig. 3, the total cases of COVID-19 in this cluster are classified into very diverse categories, starting from low to high. The total COVID-19 cases per capita in this cluster are in the high category.

C. COVID-19 Forecasting

After the clustering process, the next step is calculating the representatives of COVID-19 cases in each datum for further modeling. We calculated the average of the member of each cluster. For example, cluster 2 consists of two members. The daily information for cluster 2 on a specific date is calculated from the average of the two members on that day. The representative of each cluster is shown in Fig. 9.

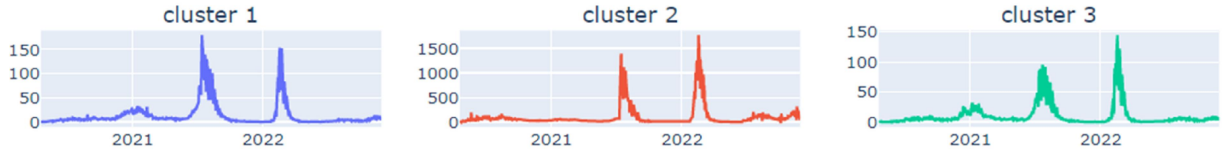

Fig. 9 COVID-19 representative in each cluster

Before further modeling, we conducted the smoothing and box-cox transformation on representative data in each cluster. Fig. 10 shows the data after smoothing and transformation.

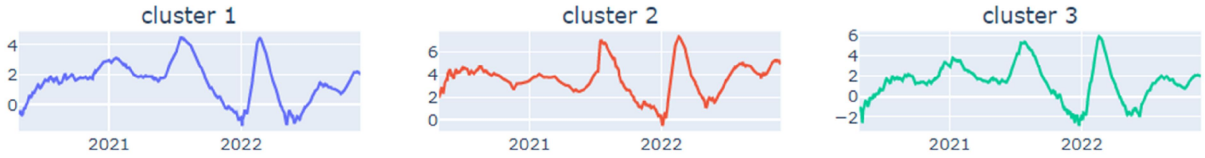

Fig. 10 COVID-19 data after smoothing and transformation

After the smoothing and transformation process, we applied the Prophet models to each cluster. The 706 days of the data are used for the training model (initial data), 30 days as testing data, and 30 days as forecast horizon. In addition, we conducted hyperparameter tuning in every cluster to obtain the best parameter. The best parameter for each model in a cluster and the error of the model are shown in Table 4.

TABLE 4
THE BEST HYPERPARAMETER FOR FORECASTING THE MODEL IN EACH CLUSTER

Cluster	Changepoint prior scale	Seasonality prior scale	Changepoint range	MAPE
1	0.5	10	0.8	0.0869
2	0.5	5	0.8	0.1513
3	0.5	10	.8	0.1040

To forecast the daily cases of COVID-19 in all clusters, we set the parameter value using the best value obtained. The output of the forecasting model in each cluster is summarized in Table 5.

TABLE 5
THE OUTPUT OF THE PROPHET MODEL IN EACH CLUSTER

ds (date)	Cluster 1		Cluster 2		Cluster 3	
	yhat (predicted)	y (actual)	yhat (predicted)	y (actual)	yhat (predicted)	y (actual)
2020-04-27	0.4681	0.5153	9.6611	7.0714	0.1924	0.2142
2020-04-28	0.4793	0.6173	9.9846	8.2857	0.1961	0.3392
2020-04-29	0.4863	0.5663	10.3049	7.7857	0.2000	0.3214
2020-04-30	0.4990	0.6683	10.7198	10.3571	0.2068	0.3035
2020-05-01	0.5895	0.6173	13.1679	14.8571	0.2940	0.2857
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2022-12-24	33.9520	-	298.5409	-	22.5210	-
2022-12-25	34.8471	-	254.6042	-	16.8561	-
2022-12-26	34.9837	-	296.8099	-	23.4338	-
2022-12-27	35.8536	-	298.6668	-	23.9868	-
2022-12-28	36.2661	-	299.4400	-	24.5559	-

As the final step, the data and the forecasted result from the Prophet model are visualized in Fig. 11. Note that the figure's y-axis represents a 7-day rolling average of COVID-19 cases within the cluster.

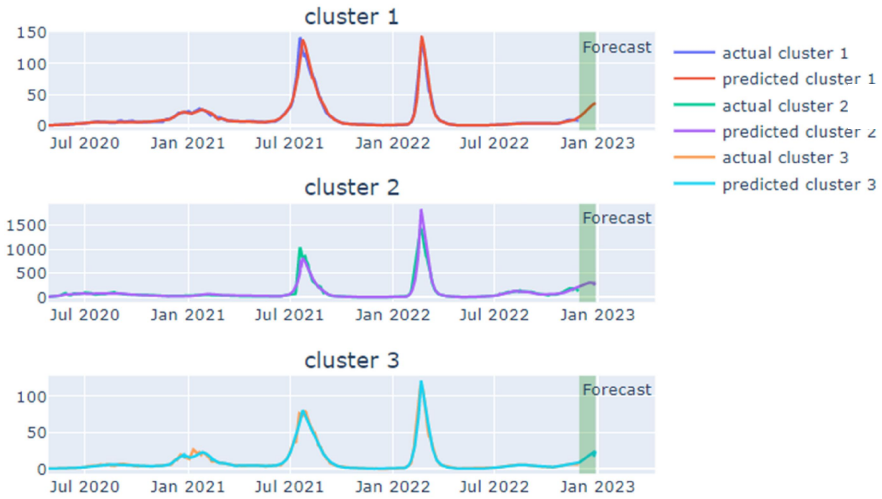

Fig. 11 The Prophet forecasts for each cluster

IV. DISCUSSION

With the analysis of the relationship between socioeconomic and demographic factors to COVID-19, we found differences in COVID-19 cases in regions with different socioeconomic and demographic characteristics. Higher COVID-19 cases were found in areas with better socioeconomic status and dense populations. This finding is associated with a study by [29]; a densely populated area is positively correlated with the spread of infection, which is related to population mobility and exposure to close social contact, allowing for a higher spread of COVID-19. In this research, cluster 2 has better socioeconomic conditions than other clusters, marked by the high value of GDP per capita, expenditure per capita, human development index, and regional minimum wage. Cluster 2 also has the highest average COVID-19 cases per capita, which means this cluster has a higher transmissibility risk of COVID-19 cases than the other clusters. A high GDP indicates high economic activity in that area, where people are more physically active, and population mobility increases. So, the possibility of virus infection is more significant if no restrictions apply. Furthermore, the area with a better economy can conduct more tests and reach a broader community to disclose more cases.

An area with a higher number of companies in large and medium industries also tends to have higher COVID-19 cases. Large and medium companies have a large number of workers, thus enabling more intense social contact. Some tasks are also impossible to do without direct interaction between workers. The condition is related to higher COVID-19 cases found in the industrial area. The research from [45] stated that the workplace is one factor that significantly

impacts the spread of the disease. Manufacturing, construction, and wholesale trade are the riskiest sector in COVID-19 transmission.

This research also found that COVID-19 cases are higher in cities than in rural areas. [46] gave four main reasons why COVID-19 is affecting cities more than rural areas: accentuating density, mass living conditions, connectivity, and exposed activity. According to [46], densely populated metropolitan areas tend to have more extensive public transportation, more personal interaction occupations, and higher housing congestion and income inequality than rural areas. [47] stated that urban congestion and interaction are beneficial to social and economic life, but they are also vectors for disease propagation.

The forecasting model we built for each cluster using FBProphet models the daily cases of COVID-19. The predictive model in a cluster with lower COVID-19 cases (cluster 1 and cluster 3) has a lower error (MAPE) than the one in a cluster with higher COVID-19 cases (cluster 2). In the model's characteristic, cluster 2, which COVID-19 hardest hit, needs a lower parameter value of seasonality prior scale. According to [48], the seasonality prior scale is a parameter to control the flexibleness of the seasonality component. A higher value allows the seasonality to adjust for large fluctuations, and smaller values reduce the size of the seasonal part. The rest of the parameters tend to be the same in all clusters.

We note some trends based on the visualization represented by Fig.11. The cluster with better socioeconomic conditions (cluster 2) experienced the highest daily new cases of COVID-19 over the period of time. This cluster only has two infection waves: during June - August 2021 and January – March 2022; meanwhile, the other two clusters have three infection waves. Across all clusters, the daily number of COVID-19 cases is projected to accelerate towards December 2022 holiday season. Also, the number of new cases in the low-income cluster (cluster 1) is expected to increase significantly in December. A cluster with the lowest total population but the second densely cluster was observed to have smaller increases than other clusters.

The result provided in this research is a great starting point and could step in as a basis for evaluating other approaches to managing infectious disease, COVID-19. But this study still has some limitations; we only used the socioeconomic and demographic indicators of the data available to the public, which sometimes do not quantify all related phenomena to grasp and clarify the results. Incorporating additional features into the clustering model may help to provide a better cluster. Then, in cluster analysis, we did not consider the spatial effect in determining the cluster of regencies/cities. For the time-series forecasting model, we did not account the effect of the COVID-19 variant and vaccination program. And the COVID-19 data used are not the actual data, and the available data are estimated to be lower than the actual data. This is because the screening process is unavailable for all individuals in a population, and the tested patients are treated differently by different local health systems. And in forecasting COVID-19, this study does not consider the importance of other controlling aspects, such as climatic change, vaccination rate, environmental conditions, the variant of the virus, lockdown intervention, and other policies. Future research could further examine the influence of these factors.

V. CONCLUSIONS

Hierarchical clustering has succeeded in dividing 38 regencies/cities in East Java into three clusters. We found that the infection waves have induced a spike in cases more concentrated in areas with specific socioeconomic and demographic characteristics. Socioeconomic and demographic factors, such as population density, urban population, regional GDP, the number of companies in large and medium industries, and the human development index, are positively associated with COVID-19 cases. In this study, we found that the highest cases of COVID-19 were found in cluster 3, the metropolis area with high population density and GDP. The forecasting model using Prophet can predict the daily cases of COVID-19 well. And the model with the best parameter from the hyperparameter tuning process is superior in forecasting COVID-19 cases in all clusters. Based on the Prophet model, it is estimated that the number of daily cases of COVID-19 will increase in all clusters. The MAPEs of the final models are 0.0869; 0.1513; and 0.1040, respectively, for cluster 1, cluster 2, and cluster 3.

Author Contributions: *Hasanah*: Conceptualization, Methodology, Writing - Original Draft, Editing, Data Collection. *Herdiyeni*: Conceptualization, Methodology, Supervision, Review. *Hardhienata*: Conceptualization, Methodology, Supervision, Review.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: origin, transmission, and characteristics of human coronaviruses," *J. Adv. Res.*, vol. 24, pp. 91–98, 2020, https://doi.org/10.1016/j.jare.2020.03.005.

[2] H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang, and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): current status and future perspectives," *Int. J. Antimicrob. Agents*, vol. 55, no. 5, p. 105951, 2020, https://doi.org/10.1016/j.ijantimicag.2020.105951.

[3] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *Lancet*, vol. 395, no. 10223, pp. 470–473, 2020, https://doi.org/10.1016/S0140-6736(20)30185-9.

[4] N. H. L. Leung, "Transmissibility and transmission of respiratory viruses," *Nat. Rev. Microbiol.*, vol. 19, no. 8, pp. 528–545, 2021, https://doi.org/10.1038/s41579-021-00535-6.

[5] WHO, "WHO Director-General's opening remarks at the media briefing on COVID-19 - March 11 2020," 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed August 11, 2022).

[6] C. Nicholson, L. Beattie, M. Beattie, T. Razzaghi, and S. Chen, "A machine learning and clustering-based approach for county-level COVID-19 analysis," *PLoS One*, vol. 17, no. 4 April, pp. 1–24, 2022, https://doi.org/10.1371/journal.pone.0267558.

[7] F. R. Lashley, "Factors Contributing to the Occurrence of Emerging Infectious Diseases," *Biol. Res. Nurs.*, vol. 4, no. 4, pp. 258–267, 2003, https://doi.org/10.1177/1099800403251238.

[8] R. B. Hawkins, E. J. Charles, and J. H. Mehaffey, "Socioeconomic status and COVID-19–related cases and fatalities," *Public Health*, vol. 189, pp. 129–134, 2020, https://doi.org/10.1016/j.puhe.2020.09.016.

[9] S. Sannigrahi, F. Pilla, B. Basu, A. S. Basu, and A. Molter, "Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach," *Sustain. Cities Soc.*, vol. 62, no. July, p. 102418, 2020, https://doi.org/10.1016/j.scs.2020.102418.

[10] Y. H. Ying, W. L. Lee, Y. C. Chi, M. J. Chen, and K. Chang, "Demographics, socioeconomic context, and the spread of infectious disease: the case of COVID-19," *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, 2022, https://doi.org/10.3390/ijerph19042206.

[11] R. P. Rajkumar, "The relationship between demographic, socioeconomic, and health-related parameters and the impact of COVID-19 on 24 regions in India: Exploratory cross-sectional study," *JMIR Public Heal. Surveill.*, vol. 6, no. 4, 2020, https://doi.org/10.2196/23083.

[12] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *J. City Dev.*, vol. 3, no. 1, pp. 12–30, 2021.

[13] B. Cabieses, H. Tunstall, and K. Pickett, "Understanding the socioeconomic status of international immigrants in Chile through hierarchical cluster analysis: A population-based study," *Int. Migr.*, vol. 53, no. 2, pp. 303–320, 2015, https://doi.org/10.1111/imig.12077.

[14] J. E. Mirowsky *et al.*, "A novel approach for measuring residential socioeconomic factors associated with cardiovascular and metabolic health," *J. Expo. Sci. Environ. Epidemiol.*, vol. 27, no. 3, pp. 281–289, 2017, https://doi.org/10.1038/jes.2016.53.

[15] A. Maugeri, M. Barchitta, G. Basile, and A. Agodi, "Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, 2021, https://doi.org/10.1038/s41598-021-86703-3.

[16] H. T. Rauf *et al.*, "Time series forecasting of COVID-19 transmission in Asia Pacific countries using deep neural networks," *Pers. Ubiquitous Comput.*, 2021, https://doi.org/10.1007/s00779-020-01494-0.

[17] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos, Solitons and Fractals*, vol. 140, p. 110212, 2020, https://doi.org/10.1016/j.chaos.2020.110212.

[18] S. Sah, B. Surendiran, R. Dhanalakshmi, S. N. Mohanty, F. Alenezi, and K. Polat, "Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India," *Comput. Math. Methods Med.*, vol. 2022, 2022, https://doi.org/10.1155/2022/1556025.

[19] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," *Chaos, Solitons and Fractals*, vol. 139, p. 110058, 2020, https://doi.org/10.1016/j.chaos.2020.110058.

[20] M. Lounis, "Predicting active , death and recovery rates of COVID-19 in Al- geria using Facebook' Prophet model," no. March, 2021, https://doi.org/10.20944/preprints202103.0019.v1.

[21] S. Belkacem, "COVID-19 data analysis and forecasting: Algeria and the world," pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2007.09755.

[22] S. F. Ardabili *et al.*, "COVID-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, no. 10, p. 249, 2020, https://doi.org/10.3390/a13100249.

[23] C. Xu, "A comparative study: time-series analysis methods for predicting COVID-19 case trend," *Degree Proj. Comput. Sci. Eng.*, 2021.

[24] V. Tulshyan, D. Sharma, and M. Mittal, "An eye on the future of COVID'19: prediction of likely positive cases and fatality in India over a 30 days horizon using Prophet Model," *Disaster Med. Public Health Prep.*, no. May, 2020, https://doi.org/10.1017/dmp.2020.444.

[25] A. K. Gupta, V. Singh, P. Mathur, and C. M. Travieso-Gonzalez, "Prediction of COVID-19 pandemic measuring criteria using support vector machine, prophet and linear regression models in Indian scenario," *J. Interdiscip. Math.*, vol. 24, no. 1, pp. 89–108, 2021, https://doi.org/10.1080/09720502.2020.1833458.

[26] Y. Yoshikawa and I. Kawachi, "Association of socioeconomic characteristics with disparities in COVID-19 outcomes in Japan," *JAMA Netw. Open*, vol. 4, no. 7, pp. 1–13, 2021, https://doi.org/10.1001/jamanetworkopen.2021.17060.

[27] Satgas COVID-19 Jatim, "Peta Sebaran COVID-19 Jatim," *Jatim Tanggap COVID-19*, 2022. https://infocovid19.jatimprov.go.id/ (accessed Aug. 10, 2022).

[28] "BPS Provinsi Jawa Timur." https://jatim.bps.go.id/.

[29] A. Buja, M. Paganini, S. Cocchio, M. Scioni, V. Rebba, and V. Baldo, "Demographic and socioeconomic factors, and healthcare resource indicators associated with the rapid spread of COVID-19 in Northern Italy: An ecological study," *PLoS One*, vol. 15, no. 12 December, pp. 1–13, 2020, https://doi.org/10.1371/journal.pone.0244535.

[30] S. Sannigrahi, F. Pilla, B. Basu, A. S. Basu, and A. Molter, "Examining the association between socio-demographic composition and COVID-19 fatalities in the European region using spatial regression approach," *Sustain. Cities Soc.*, vol. 62, no. January, 2020, https://doi.org/10.1016/j.scs.2020.102418.

[31] N. Ulinnuh and R. Veriani, "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular Menggunakan Metode Complete Linkage, Average Linkage dan Ward," *J. Nas. Inform. dan Teknol. Jar.*, vol. 5, 2020.

[32] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed., vol. 87, no. 4. Annabel Ainscow, 2019.

[33] K. Pearson, "notes on the history of correlation," *Biometrika*, vol. 13, no. 1, p. 25, 1920, https://doi.org/10.2307/2331722.

[34] B. Ratner, "The correlation coefficient: Its values range between 1/1, or do they," *J. Targeting, Meas. Anal. Mark.*, vol. 17, no. 2, pp. 139–142, 2009, https://doi.org/10.1057/jt.2009.5.

[35] N. Shrestha, "Detecting multicollinearity in regression analysis," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, 2020, https://doi.org/10.12691/ajams-8-2-1.

[36] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," *J. Signal Inf. Process.*, vol. 04, no. 03, pp. 173–175, 2013, https://doi.org/10.4236/jsip.2013.43b031.

[37] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space ," *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901, https://doi.org/10.1080/14786440109462720.

[38] R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Pearson Education, 2014.

[39] T. Strauss and M. J. Von Maltitz, "Generalising ward's method for use with manhattan distances," *PLoS One*, vol. 12, no. 1, pp. 1–21, 2017, https://doi.org/10.1371/journal.pone.0168288.

[40] S. Saraçli, N. Doğan, and I. Doğan, "Comparison of hierarchical cluster analysis methods by cophenetic correlation," *J. Inequalities Appl.*, vol. 2013, pp. 1–8, 2013, https://doi.org/10.1186/1029-242X-2013-203.

[41] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, https://doi.org/10.1016/0377-0427(87)90125-7.

[42] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *J. R. Stat. Soc. Ser. B*, vol. 26, no. 2, pp. 211–243, 1964, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

[43] S. J. Taylor and B. Letham, "Forecasting at Scale," *Am. Stat.*, vol. 72, no. 1, pp. 37–45, 2018, https://doi.org/10.1080/00031305.2017.1380080.

[44] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction Time Series Analysis and Forecasting*, 2nd ed. New Jersey (US): John Wiley & Sons, 2015.

[45] M. Murti *et al.*, "COVID-19 workplace outbreaks by industry sector and their associated household transmission, Ontario, Canada, January to June, 2020," *J. Occup. Environ. Med.*, vol. 63, no. 7, pp. 574–580, 2021, https://doi.org/10.1097/JOM.0000000000002201.

[46] J. Matheson, M. Nathan, H. Pickard, and E. Vanino, "Why has coronavirus affected cities more than rural areas?," *Economic Observatory*, 2020. https://www.economicsobservatory.com/why-has-coronavirus-affected-cities-more-rural-areas (accessed Dec. 07, 2022).

[47] M. Nathan, "The city and the virus," *Medium*, 2020. https://maxnathan.medium.com/the-city-and-the-virus-db8f4a68e404 (accessed February 18, 2023).

[48] "Prophet Diagnostics." https://facebook.github.io/prophet/docs/diagnostics.html (accessed February 18, 2023).