Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 7, No.1, March 2024, pp. 121 – 133
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

The Use of Satellite Imagery Data for Poverty Clustering at the District Level Administration in Indonesia

Azzahra Dhisa Khamila[1*], Martha Budi Wardani[2], Robert Kurniawan[3]
[1,2] Statistics Study Program, STIS Statistics Polytechnic, Jakarta, Indonesia
[3] Statistical Computing Study Program, STIS Statistics Polytechnic, Jakarta, Indonesia
Email: [1] 212011481@stis.ac.id, [2] 212011824@stis.ac.id, [3] robertk@stis.ac.id

Article Info
Article history:
Received Oct 12th, 2023
Revised Feb 5th, 2024
Accepted Mar 8th, 2024

Keyword:
Clusterization
District Level
K-Means
Poverty
Satellite Imagery

ABSTRACT
Poverty is a problem that will never be separated from every country, including Indonesia. One of the efforts that can be taken to reduce poverty is to carry out comprehensive monitoring of data related to poverty. The use of satellite imagery strongly supports this effort. Data taken to describe poverty in a region are CO, SO2, NO2, Night Time Light (NTL), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), Normalized Difference Built-up Index (NDBI), also per capita expenditure data that can be accessed through the BPS website. Based on the theory, all of these variables negatively affect the poverty of a region except for the NDVI variable. The use of clustering with K-Means method can be implemented in this situation in order to cluster poverty in every district in Indonesia. Then it is supported by a descriptive analysis of each variable in order to describe the distribution of variables in each district in Indonesia. Based on the clustering results, it can be seen that there are 2 clusters, namely cluster 1 which shows a cluster with low poverty and cluster 2 with high poverty. There are a total of 46 districts included in cluster 1, which constitute the majority of economic centers in it's region, and 468 other districts included in cluster 2. The results of this clustering are expected to be used by stakeholders in making decisions according to the characteristics of the district.

Corresponding Author:
Azzahra Dhisa Khamila,
DIV Statistics,
Polytechnic of Statistics STIS,
Jl. Otto Iskandardinata No.64C, City of East Jakarta, Special Capital Region of Jakarta 1330
Email: 212011481@stis.ac.id
DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.25278

1. INTRODUCTION

Poverty is a social problem that is still a challenge for all countries in the world [1]. Poverty is a very complex and widespread problem in developing countries. One of the developing countries that still faces poverty is Indonesia [2]. According to BPS, around 10.14% or 27.54 million people in Indonesia were below the poverty line as of March 2021. People who are below the poverty line are a group that is experiencing a complete lack of basic needs. All efforts are made by the government in alleviating poverty, so as not to create various other problems both in the social, environmental and economic spheres in the long term. Poverty alleviation in Indonesia can be carried out by comprehensive monitoring of various data related to poverty, so that it can be known which areas are included in the poor areas and in terms of which variables must be addressed.

One of the poverty monitoring efforts is the use of satellite imagery data. Satellite imagery is able to provide an accurate scale in observing socio-economic phenomena [3]. An example of using satellite imagery for poverty monitoring is Night Time Light (NTL) [4]. NTL is able to capture the Gross Domestic Product (GDP) based on the intensity of night light in the economic activities carried out [5]. This illustrates that NTL is also able to identify the poverty of a region [6]. The higher the NTL, the higher the economic activity in it.

So that it can identify low poverty in the region. Because poverty is very complex, the NTL variable alone is not enough to be able to identify poverty [7].

During the day, satellite imagery can also identify poverty based on land use or land cover, development growth, and accessibility [8]. One of the image data that is able to capture this information is the Normalized Difference Vegetation Index (NDVI). NDVI has a correlation with poverty [9]. Negative NDVI values describe non-vegetated areas such as waters, while positive NDVI values describe vegetated areas [10]. The lower the NDVI value, the lower the poverty [11]. This is because the vegetation area is very low, which means that there are many standing buildings and dense areas of buildings that describe progress in an area. In addition, the Normalized Difference Built-up Index (NDBI) is also used. Positive NDBI values represent urban areas and negative NDBI values represent non-urban areas [12]. This shows that the higher the NDBI, the lower the poverty in that region.

On the other hand, there is the Land Surface Temperature (LST) which is able to identify poverty. The higher the LST value, the wider the hot area in urban areas [13]. Areas with low LST describe mountainous areas, while areas with high LST describe areas of arid land and limestone mining [11]. Regions with high poverty rates tend to have low LST scores. This shows that areas with high LST have low poverty.

Poverty can also be reflected in the air quality of an area, in which three pollutant elements were used in this study, namely CO, NO_2, and SO_2. Carbon monoxide (CO) has a relationship with economic growth [14]. Nitrogen dioxide (NO_2) has a positive correlation with GDP [15]. Sulfur dioxide (SO_2) has a significant effect on population density and energy consumption [16]. Therefore, CO, NO_2, and SO_2 can be used as approach variables to detect any economic activity taking place. The higher air pollution means that there are many economic activities that cause high economic growth, so that poverty is low.

Apart from using the satellite imagery data described earlier, poverty mapping also requires per capita expenditure data. Expenditure per capita has a negative effect on poverty [17]. The higher the per capita expenditure, the lower the poverty in the region. This is because the region has a high ability to meet their basic needs. On the other hand, high per capita expenditure indicates increased regional welfare [18].

Clustering is grouping observations into one or more groups, so that the observations collected in each group have a high degree of similarity. Clustering is unsupervised , where the application does not need training data, teachers and output targets. The purpose of clustering is to identify data groups so as to produce the properties of the data itself [19] ; [20]. One of the clustering methods with high efficiency and fast use is the k-means method. This method aims to create clusters of objects based on attributes into k partitions.

Detecting poverty level using clustering is an interesting thing to research. Unfortunately, there has been no research that has examined this thoroughly in Indonesia. Research conducted by Putri, Wijayanto, & Sakti [11] only classified poverty in East Java based on the Relative Spatial Poverty Index (RSPI). So how to get a comprehensive poverty cluster in Indonesia and what are the results? To answer this question, this research conducted with the aim of obtaining poverty clusters for districts/cities throughout Indonesia based on a combination of satellite imagery data and aggregate data, which are considered capable of reflecting the level of poverty. This research can contribute to providing insight for the government about the poverty cluster that will enrich or become the basis for them in determining whether areas are poor or not using these variables, which can be used as a reference in forming policies to meet the first SDGs goal, end poverty in all its forms everywhere.

2. RESEARCH METHODS

This research was conducted in all districts or cities in Indonesia, with a total of 514 districts or cities. From this locus, per capita expenditure data were taken, obtained from the BPS website for each province, and several geospatial variables obtained through remote sensing in the form of raster data. The variables referred to include carbon monoxide (CO), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2) obtained from the Sentinel-5P satellite, night-time light intensity (NTL) obtained from the NOAA-VIIRS satellite, Land Surface Temperature (LST) obtained from the MODIS satellite, as well as the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Built-Up Index (NDBI) obtained from the Sentinel-2 satellite. The satellite image data was collected using the Google Earth Engine (GEE), with the point in time the data taken was for a period of one year, namely from 1 January 2021 to 31 December 2021. Likewise, the per capita expenditure data used was at the point in 2021. Overall, the variables used can be seen in table 1.

In data collection, data preprocessing , and processing, the tools used by researchers included the Google Earth Engine (GEE) platform and QGIS and R-Studio software . Google Earth Engine (GEE) is a cloud-based computing platform that can be used for large data collection and processing, with a petabyte capacity, which is usually related to analysis or decision making [21], for example research on land use change detection in Singapore [22] as well as in the Savannah river basin [23]. This platform is capable of providing satellite imagery with very low or cloud-free cloud cover [24].

Table 1. List of variables used.

Variable	Source	Bands	Explanation
Perka	BPS for each province	-	Expenditure per capita for each district/city in 2021
CO	Sentinel-5P	CO column number density	Median value of 2021 data
NO_2		NO_2 column number density	
SO_2		SO_2 column number density	
NTL	NOAA-VIIRS	avg_rad	The median value of the 2021 NTL data
LST	MODIS	LST_day_1km	The median value of ESG data for 2021
NDVI NDBI	Sentinel-2	B8 (NIR) and B3 (Red) B8 (NIR) and B11 (SWIR)	Median value of 2021 data

The Sentinel-5P Tropospheric Monitoring Instrument (TROPOMI) satellite at GEE is divided into 2 instruments, namely the Sentinel-5 Precusor instrument which functions to collect useful data for assessing air quality, and the TROPOMI instrument which is a multispectral sensor that records reflections of wavelengths that are important for measuring the concentrations of carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), as well as several other air contents in the atmosphere [25]. In this study, the CO, NO_2, and SO_2 concentration variables obtained were Near Real Time Imagery (NRTI) with a spatial resolution of 1113.2 meters. CO is a gas that is the main pollutant source in several urban areas, originating from burning fossil fuels, burning biomass, and atmospheric oxidation of methane and other hydrocarbons. NO_2 is a pollutant gas found in the troposphere and stratosphere which is formed due to anthropogenic activities, especially the burning of fossil fuels and biomass, as well as natural processes, such as forest fires, lightning, and microbiological processes in the soil. While SO_2 is a pollutant gas that is almost the same as NO_2, which is produced due to anthropogenic activities and natural processes. SO_2 emissions apart from having a negative impact on health, can also affect the climate by means of Radiative Forcing (RF) through the formation of sulfate aerosols [25].

The Visible Infrared Imaging Radiometer Suite (VIIRS) can capture composite images of the monthly average emission using nighttime data, or in other words the resulting product is Night-Time Light Intensity (NTL), where this data is combined every month with a 463 spatial resolution. .83 meters. NTL is night light that comes from the brightness of cities, farms, industrial areas, fishing boat lights, forest fires, or other areas of human activity that form the image of night light. This data is widely used in research on urban expansion, urban structure, and estimation of socioeconomic status [26].

The MOD11A1 V6 product obtained from the MODIS satellite captures daily Land Surface Temperature (LST) and emissivity values in a 1200 x 1200 kilometer grid. LST is an important variable in the Earth's climate system. This variable provides temperature information in the outer layers of the earth's surface that can be used in various fields, such as climate change, the hydrological cycle, vegetation monitoring, as well as urban climate and environmental studies [27]. LST data in this study were taken using a spatial resolution of 1000 meters.

Sentinel-2 is a wide, high-resolution map with multispectral imaging capabilities with the main objective of supporting research in The Copernicus Land Monitoring program, which includes monitoring of vegetation cover, soil and water, as well as observations of inland waterways and coastal areas. The data taken by this satellite is the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Built-Up Index (NDBI) obtained using a spatial resolution of 5566 meters. NDVI is a spectral reflection that is widely used as a measure that shows the strength of vegetation, a measure of ecological replacement of absorbed photosynthetically active radiation (APAR), as well as Leaf Area Index (LAI) which is the main input of biosphere processes [28]. While NDBI is a linear combination of Near Infrared Band (NIR) (0.76 – 0.90 μm) and Middle Infrared Band (MIR) (1.55 – 1.75 μm), which can be extracted as urban built areas [29].

The QGIS software is used to combine data, visualize, process raster data, and perform zonal statistics to obtain the median value of raster data. While the RStudio software is used for merging, data preprocessing , and data processing. Data preprocessing or data preparation in data mining is a series of activities intended to fulfill two main objectives, namely obtaining data with a good structure or suitable for data mining, and obtaining data capable of providing good performance and quality for modeling or other appropriate matters. with research purposes [30]. The flow of activities from data collection to ready-to-process data carried out in this study can be seen in Figure 1.

1. Data collection, both data sourced from the BPS website and satellite image data obtained through the Google Earth Engine (GEE).
2. The data obtained through GEE is raster data in .tif format, which is then processed using QGIS. Data processing requires shapefiles for district/city boundaries throughout Indonesia as well as zonal

statistics tools to obtain a table of median values for each regional boundary based on the corresponding variable.
3. After the median value of all raster data is obtained, the value table is exported into a Comma Separated Value (CSV) file.
4. Because each raster data will produce one CSV file each, it is necessary to combine the variables with the district/city code primary key which can be done using QGIS software or manually.
5. After all the variables from the raster data have been combined, then it is necessary to combine the satellite image data with data obtained through the BPS website , namely per capita expenditure. This data aggregation is done with software R, using the full_join function which is based on the district/city code.
6. Next is to do data preprocessing, starting from checking variable types, checking missing values , and imputing missing values with the median.
7. After the types of all variables are appropriate and no longer contain missing values , the data is ready to be processed.

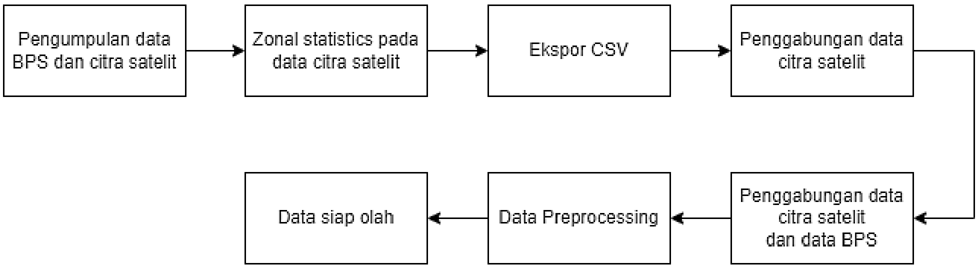


Figure 1. Data Retrieval Flow

2.1. Cluster Analysis

A cluster is a collection of entities that have the same cluster structure. The cluster structure is a representation of an entity so that some of these entities can be grouped into a part of the whole both hierarchically and non-hierarchically. While clustering is a series of activities that aim to find and describe the cluster structure in a dataset [31]. Other sources state that clustering is a grouping method based on a measure of similarity or proximity. Unlike a classification that has a predetermined number of classes, in clustering it is necessary to search for the best number of clusters and there is no exercise or training process to group each data [32].

Cluster analysis can be carried out using two methods, namely hierarchical and non-hierarchical methods. In the hierarchical method, the formation of groups or clusters is carried out in a predetermined order, which in its application there are two types of grouping sequences, namely agglomerative and divisive. Whereas in the non-hierarchical method, the formation of groups or clusters is done by combining or separating clusters without following a certain sequence structure, by trying to maximize or minimize several evaluation criteria [33]. One example of non-hierarchical cluster analysis is the K-means method , which is used in this study.

Entity grouping is based on the proximity or similarity of certain characteristics used. So an important step in cluster analysis is the measurement of proximity distances. There are two types of distance measurements, namely the Manhattan distance and the Euclidean distance , which in this study used the Euclidean distance measurement . The formula for calculating the Euclidean distance from the point **x** $(x_1, x_2, \text{etc})$and **y** $(y_1, y_2, \text{etc})$is as follows.

$$d = \sqrt{\sum_{j=1}^{n} \left(x_j - y_j\right)^2} \qquad (1)$$

Where:

d = distance of point x and y
x_j = x value in the jth variable
y_j = y value in the jth variable
n = number of observation variables

2.2. K-Means method

K-means method is a method in non-hierarchical cluster analysis that is often used. In this method, each point will be grouped into a cluster with the nearest centroid from the corresponding point [33]. The clustering process in the K-means method follows the flow: choosing the number of k clusters, setting the center point (centroid) as the cluster center in various ways, either randomly or not, calculating the distance of each point to the cluster center point to determine which cluster the point is included in, standardizing data, recalculating the center point of each cluster with the new standardized value, and recalculating the distance of each point to the new center point [34]. In this study the K-Means method was used for reasons of low complexity in terms of time to iterate and space to store data, and did not need to depend on a particular order. The clustering flow that is carried out refers to Figure 2.

Figure 2. Clusterization Flow

The first step taken is the stage of descriptive analysis or data exploration of all the variables used. Descriptive analysis is an analytical method used to describe the characteristics of some or all of the data objectively [35]. After conducting descriptive analysis, it is necessary to standardize the data because there are different units for each variable used, as well as to standardize the variable values so that they are not too spread out.

After the data is standardized, the next step is to determine the possible k values, then cluster with each k that is proposed. Selection of the number of k clusters can be done using the Elbow and Silhouette methods. After the clusters are formed, the selection of optimal k can be reviewed based on several validation measures, which in this study used the dunn index . If the cluster with the optimum k value has been obtained, the activity can be continued by analyzing the results of the clustering. Then it can be forwarded to interpretation and drawing conclusions.

2.3. Optimal K Value Validation

One validation measure for the optimal k value that can be used is the Dunn Index . Dunn Index (DI) serves to show how well the clustering algorithm can ensure that the variance between classes can be as small as possible, and the variance within each class can be as large as possible [36]. The dunn index calculation formula is as follows.

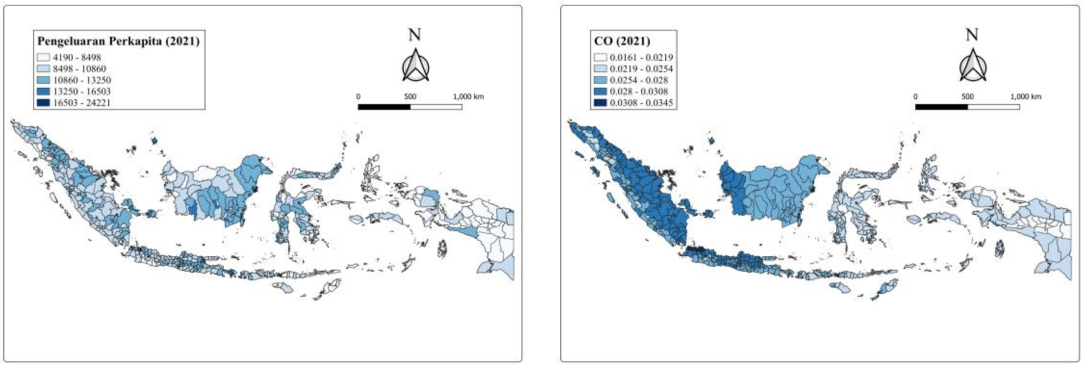
$$\mathrm{DI} = \frac{\min_{1 \le i \le j \le K} d(c_i, c_j)}{\max_{1 \le s \le K} \Delta_s} \tag{2}$$

Where :

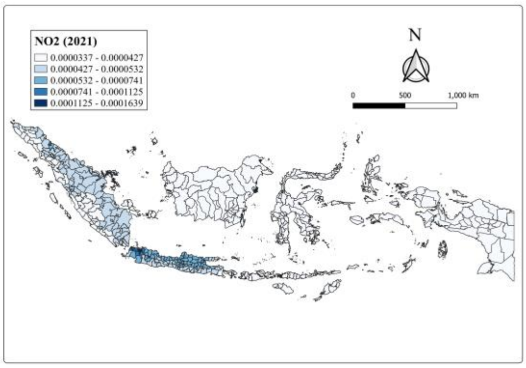
c_i =centroid point of the i-th cluster
c_j = centroid point of the jth cluster
Δ_s = maximum distance from two points in the clusterc_s

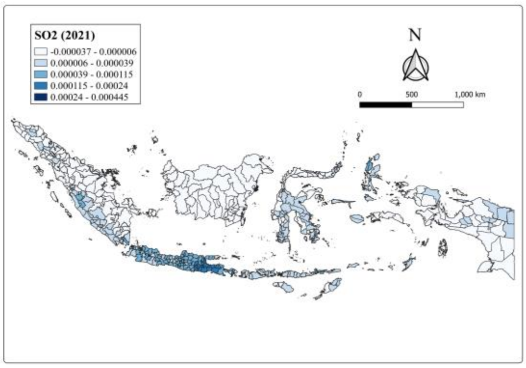
3. RESULTS AND ANALYSIS

After all the data has been obtained and merged, then the data preprocessing process is carried out . If the data structure is in accordance with the needs of analysis, then proceed with descriptive analysis to see the distribution of each variable in the data. Based on the results of the descriptive analysis carried out on the per capita expenditure variables, CO, NO_2, SO_2, NTL, LST, NDVI and NDBI, the following information is obtained.

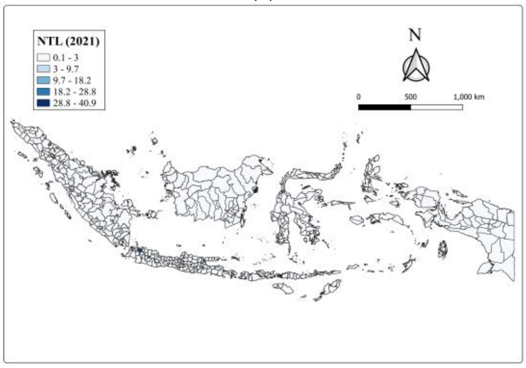


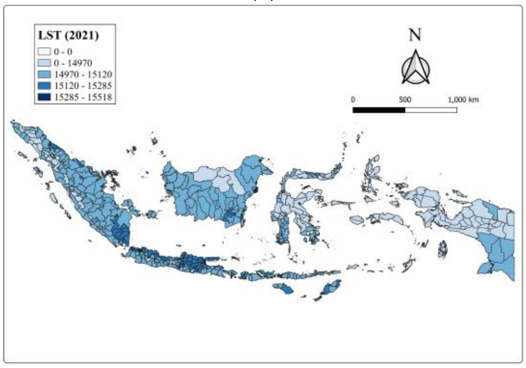
(a) (b)



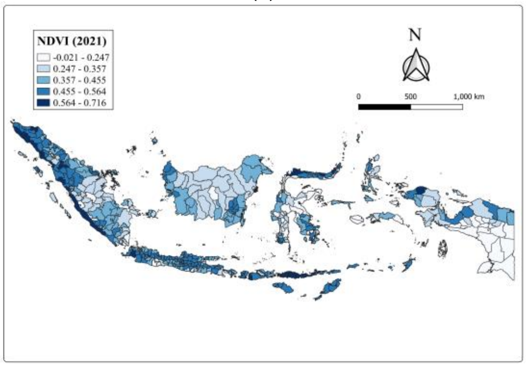


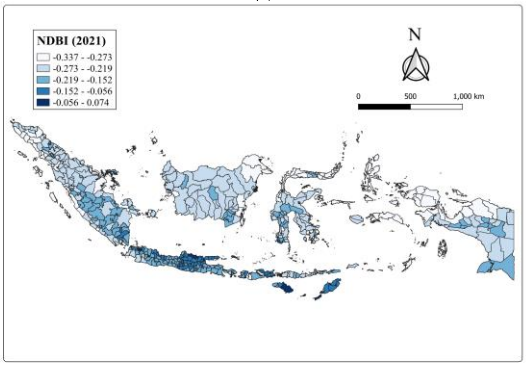
(c) (d)





(e) (f)





(g) (h)

Figure 3. Thematic map of the distribution of variables at the district or city level in Indonesia in 2021 ((a) Per Capita Expenditure Variables; (b) CO Variables; (c) NO 2 Variables; (d) SO 2 Variables; (e) NTL Variables ; (f) ESG Variables (g) NDVI Variables (h) NDBI Variables).

Based on Figure 3, it can be seen that the distribution of expenditure per capita in each district/city in Indonesia varies. Regions with high per capita expenditure are also areas with high welfare. Java Island is dominated by quite dark colors, considering that Java Island is the most populous island with many urban areas. This is in line with the results of a study conducted by Çağlayan, E. and Astar, M. [37] that urban spending will be about two times higher than rural spending.

If we review the elements of pollutants, it can be seen that Java Island is the area with the highest NO_2 and SO_2 levels compared to other regions. Meanwhile, CO levels are quite high on the islands of Sumatra and Java, especially DKI Jakarta, West Java and its surroundings. The high levels of CO and NO_2 in these areas are due to the large use of motorized vehicles [38]. Meanwhile, the high levels of SO_2 are caused not only by the large number of vehicles used, but also by volcanic activity.

Through the thematic maps of NTL distribution, it can be seen that NTL in Indonesia is low in almost all districts or cities. However, DKI Jakarta is the area with the highest NTL when compared to other provinces. This shows that economic activity at night that occurs in Jakarta has a higher intensity compared to other regions, because NTL can be used as an approach variable to measure the extent to which the transportation network explains the dynamics of economic activity [39].

Through the thematic maps of LST distribution, it is clear that LST in Indonesia is diverse, and tends to be high in urban areas, the majority of which are on the island of Java. This result is in line with the theory that was first explained in 1818 regarding the Urban Heat Island (UHI), namely a phenomenon when the atmospheric and surface temperatures will be hotter in urban areas than rural areas [40]. It was explained that there are two types of UHI, namely UHI in the atmosphere and UHI on the surface. UHI in the atmosphere is measured by air temperature, while UHI on the surface is measured by Land Surface Tenperature (LST) [41]. According to research conducted by Estoque RC, et al [42], it was found that there is a close relationship between impermeable surfaces or densely built areas such as cities and green spaces to LST. In urban areas, the average LST will be 3°C higher than in green spaces.

On the thematic map of NDVI distribution, it can be seen that there are still many areas in Indonesia that have high NDVI values, the majority of which are located on the outskirts of Sumatra Island. The high NDVI in the region is due to the fact that there are still many forest or vegetation areas. On the other hand, even though it is the most densely populated area, NDVI in Java is still relatively high. However, based on research conducted by Benedict and Jaelani, Lalu M. [43], the NDVI value in Java continues to decline from year to year as forest areas and other vegetation areas decrease due to population growth or natural disasters.

On the thematic map of NDBI distribution, it can be seen that Java Island is the region with the most dark areas, which means that NDBI in Java Island is higher compared to other areas. Normalized Difference Built-Up Index (NDBI) is a variable that is known to be effective in detecting changes in the built-up area through remote sensing techniques [44]. With a high NDBI value on the island of Java, it means that there have been many area changes to become built-up areas. In addition to reviewing each variable used with thematic maps, a histogram of each variable is also formed to determine its distribution.

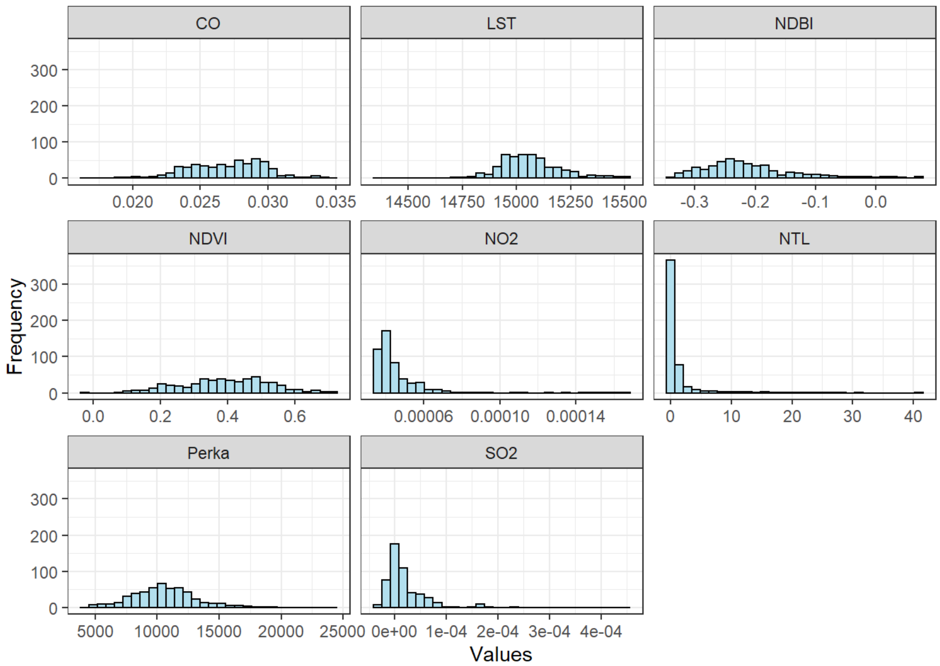


Figure 4. Histogram of all variables used.

Based on Figure 4, it can be seen that the distribution of CO, LST, and NDVI data at the district or city level in Indonesia tends to be normally distributed. Meanwhile, the variables NDBI, NO_2, SO_2, NTL, and per capita expenditure tend to be right skewed or have a positive slope, which means that the majority of these variables in regencies or cities in Indonesia have low values. Furthermore, it is also examined whether there is a relationship between the variables used. This examination is carried out by making a correlation chart which results are as follows.

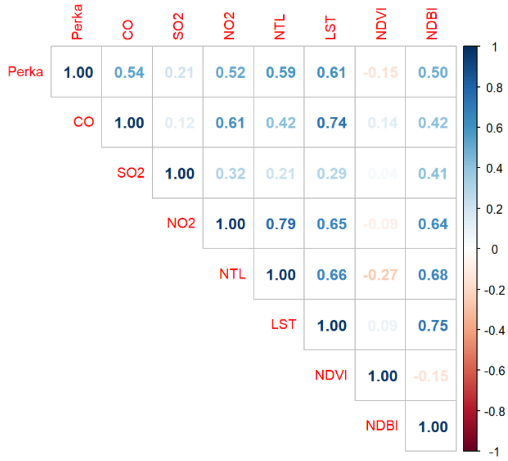


Figure 5. Graph of the correlation of each variable used.

Through the correlation graph in Figure 5 it is known that there is no strong correlation between the variables used. While the biggest correlation is in NO_2 and NTL variables. After carrying out a series of descriptive analyzes above, the stages of data standardization were carried out, which were then selected for the number of clusters using the Elbow and Silhouette methods. The results of cluster selection are presented in table 2.

Table 2. Results of the number of cluster selection based on the Elbow and Silhouette methods.

	Cluster Selection Method	
	Elbows	Silhouettes
Number of Clusters	3	2

Based on the Elbow method, the number of good clusters is 3 clusters. Meanwhile, based on the Silhouette method, a good number of clusters is 2 clusters. So that the clustering will be carried out twice with the K-means method using the sum of k = 2 and k = 3. The results of the clustering can be seen in Figure 6.

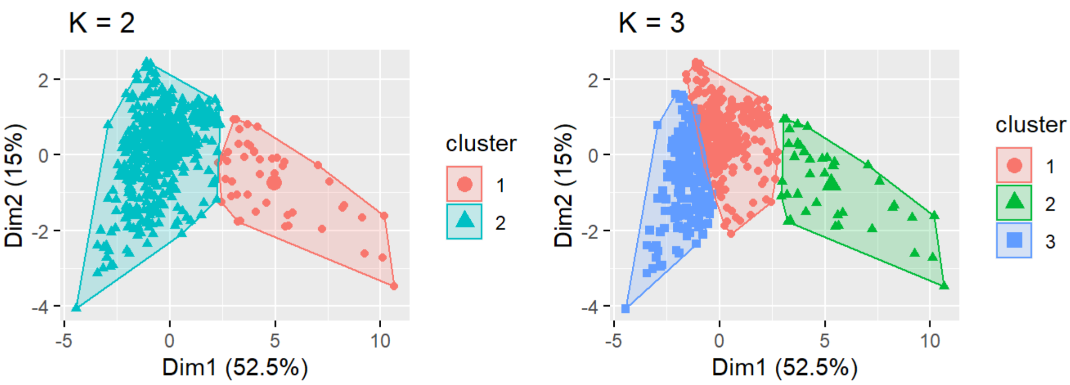


Figure 6. Cluster graph formed with k = 2 and k = 3.

When viewed from the graph of the clustering results, it appears that by using a number of clusters of 2 there is a large imbalance in the number of observations included in clusters 1 and 2, of which the majority are included in cluster 2. Meanwhile, if using a number of clusters of 3, the observations previously included in cluster 2 divided into 2 clusters. To determine the use of the optimal number of k clusters, it is necessary to calculate the Dunn Index . The calculation results are presented in table 3.

Table 3. Dunn Index values in clusters with k = 2 and k = 3.

	Number of Clusters	
	2	3
Dunn Index	0.0867	0.0263

Dunn Index value indicates a better clustering result. It can be seen that using 2 clusters will result in a larger DI value. So that of the 514 regencies or cities that were used as observations, they were divided into 2 clusters. The first cluster contains 46 districts/cities and the second cluster contains 468 districts/cities. The following is the average value of the variables from each cluster.

Table 4. The average value of each variable based on the cluster.

Cluster	Perka	CO	SO2	NO2	NTL	LST	NDVI	NDBI
1	170.271	119.426	105.776	20.546	260.251	200.251	-0.6873	210.945
2	-0.1674	-0.1174	-0.104	-0.2019	-0.2558	-0.1968	0.06755	-0.2073

From table 4 it can be concluded as follows:

1. Cluster 1: Regencies/cities that have high per capita expenditure, high CO, high SO_2, high NO_2 , high NTL, high LST, low NDVI and high NDBI. Based on the theory, this criterion can describe that cluster 1 is a group of districts/cities with low poverty. Consists of 46 regencies/cities.
2. Cluster 2: Regencies/cities that have low per capita expenditure, low CO, low SO_2, low NO_2 , low NTL, low LST, high NDVI, and low NDBI . Based on the theory , this criterion can describe that cluster 2 is a group of districts/cities with high poverty . Consists of 468 regencies/cities.

Regencies or cities in Indonesia belonging to cluster 1, or districts/cities with low poverty include Banda Aceh, Medan City, Jambi, Palembang, Bandar Lampung, South Jakarta, East Jakarta, Central Jakarta, West Jakarta, North Jakarta, Bekasi, Bogor City, Bandung City, Cirebon City, Bekasi City, Depok, Cimahi, Magelang City, Surakarta, Salatiga, Semarang City, Pekalongan City, Tegal City, Yogyakarta City, Sidoarjo, Mojokerto, Gresik, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City, Tangerang, Tangerang City, Cilegon, Serang City, South Tangerang, Denpasar City, Mataram, Kupang City, Pontianak City, Banjarmasin, and Makassar. While the remaining districts/cities are included in cluster 2 or districts/cities with high poverty.

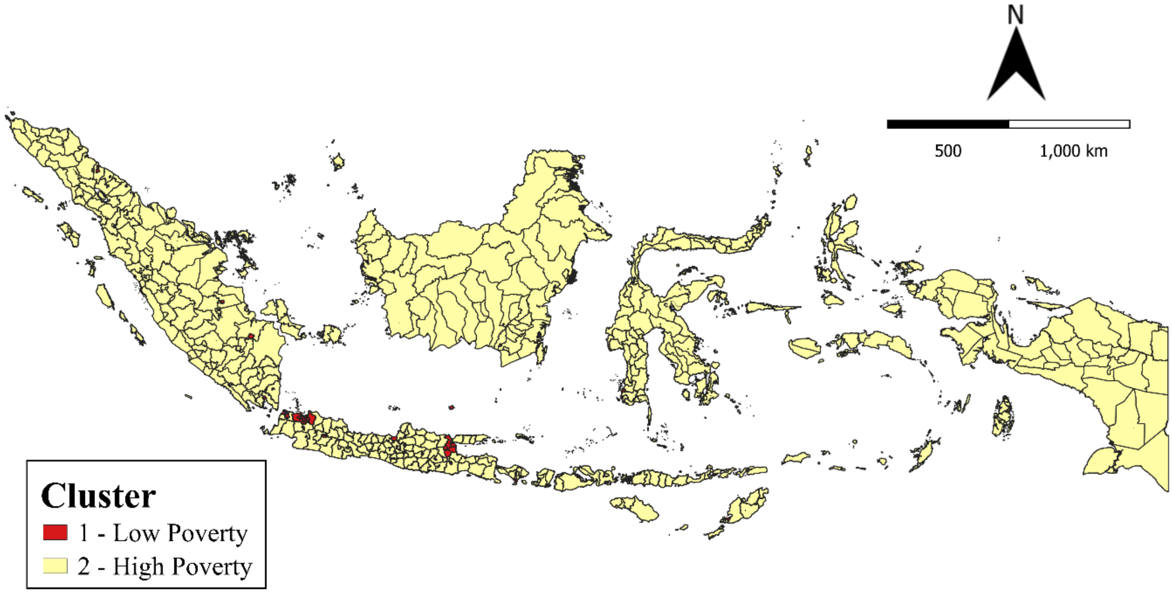


Figure 7. Map of Poverty Clustering Result in District/Cities in Indonesia.

Based on the map in figure 7, the majority of areas belonging to the low poverty cluster are districts/cities as economic centers for the surrounding districts/cities. The large number of economic activities occurring in these regencies/cities has resulted in increased levels of air pollution. These results are in line with research conducted by Wang, et al. [45] who stated that an increase in economic activity as part of urbanization has a significant positive effect on atmospheric pollution. In other words, cities which are economic centers tend to have a negative effect on air pollution. This statement is also supported by the results of research conducted by Adebayo, et al. [46] in Indonesia, which states that there is a significant positive interaction between economic growth and pollutant gas emissions. It shows that economic development in Indonesia is in the scale-effect phase, which is characterized by economic growth and environmental degradation growing simultaneously. This statement is proven by the fact that one of the cities included in the low poverty cluster which is Jakarta, the economic center of Indonesia, has high air pollution and was even the city with the worst air quality in the world.

In addition, areas that are included in the low poverty cluster tend to have high night light intensity. This statement is in line with the results of research conducted by Pan, Fu, and Zheng [47] in China, that there is a significant positive relationship between NTL and GDP at the prefectural and district levels. This means that the higher the GDP, the higher the light output at night, as measured by NTL. In another study, conducted by Afrianto [48] located in East Java, it showed the same results that NTL had a significant positive effect on GRDP as indicated by an increase of 1 watt/cm2.str in NTL intensity would increase GRDP by 1.400943 billion rupiahs. This research states that the city of Surabaya is the pole of economic growth in East Java, which then spreads to the surrounding areas. This is in line with the cluster result which shows that Surabaya City and its surroundings (Sidoarjo, Mojokerto, Gresik, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, and Madiun City) are included in first cluster, low poverty. However, Xu, et al. [49] stated that the use of NTL to predict poverty should be more careful, especially in groups of areas that are included in low poverty. This is because the presence of black dots in the area can indicate hidden poverty.

Areas with low poverty also tend to have high land surface temperatures. This statement is in line with the results of research conducted by Kanga, et al. [50] in Bangalore, India, that after comparing temperature changes from 10 urban locations and calculating the average value, it is known that an increase in urbanization will be followed by an increase in LST. This means that cities with a higher level of economic activity compared to rural areas will have higher surface temperatures. In Indonesia, Jakarta and West Java are the provinces with the most urbanization destinations until now. Based on research conducted by Nurwanda and Honjo [51] in Bogor City, it is predicted that in 2027 there will be an increase in built-up land due to urbanization from the center of Bogor City to suburban areas, followed by the expansion of warmer areas. The LST of Bogor City in 2027 is predicted to increase continuously due to the increase in urban areas, where it is obvious that more economic activity will occur in urban areas. These findings are support the cluster result which shows that Bogor City is included in the low poverty cluster.

In research by Kanga, et al. [50] also stated that an increase in surface temperature in an area would be followed by reduced vegetation and increased buildings as a result of urbanization. This is consistent with the characteristics of the low poverty cluster which has a low land cover index and a high built-up area change index. Those results are in line with research conducted by Wiratama et al. [52] in Kalimantan, which is stated that there is a significant influence of infrastructure development on poverty alleviation, which is indicated by an increase of 1 point in the Physical Infrastructure Index (PII) can reduce Kalimantan's poverty rate by 0.07 percent. In line with this research, in the poverty cluster that was formed, it was proven that Pontianak City and Banjarmasin City were included in the low poverty cluster. These two cities are the economic centers of Kalimantan so they have more complete infrastructure compared to other districts/cities in Kalimantan, considering that Kalimantan is a large island that is still dominated by villages that have not developed much.

Besides that, the characteristics of regions that are included in the low poverty cluster will have high per capita spending, as the results of research by Cammeraat [17] which state that public spending for social purposes will have a significant negative effect on poverty and inequality. According to research conducted by Ningtias and Anwar [53] located in Makassar City, it is stated that the higher percapita expenditure, the further one is from the cycle of poverty. If we look at the real condition, Makassar City owns the most expensive cost of living outside Java. It because Makassar City is the center of economic growth in the Eastern Indonesia Region. This information supports the research findings that Makassar City is included in the low poverty cluster.

The characteristics mentioned above are inversely proportional to the situation in the second cluster, where the economic activity that occurs in the area is not as massive as what happened in the districts/cities in the first cluster. So that the characteristics of the regencies/cities in the second cluster include air pollution levels that are not too high, low night light intensity, low land surface temperature, land cover index tends to be high, index of change in built-up areas is low, and low per capita expenditure. It is the regions that are included in the second cluster that need more attention from the government.

4. CONCLUSION

Cluster analysis using the K-Means method can produce comprehensive mapping information for poor districts/cities in Indonesia which is useful for evaluating and improving policies or decision-making processes for stakeholders, such as the government and the private sector, based on the variable characteristics of the two clusters. Even though information about mapping with this method is obtained quickly and easily, improvements are needed so that better mapping results are obtained with the hope that later decisions taken by stakeholders will be more on target. Therefore, suggestions for further research are to add or change variables that can clarify the different conditions of each cluster and use other methods that might produce better clustering.

REFERENCE

[1] Sano AVD, Nindito H. Application Of K-Means Algorithm For Cluster Analysis On Poverty Of Provinces In Indonesia. ComTech. 2016; 7(2): 141-150.

[2] Idrus S, Rosida L. Poverty In Indonesia : Critical Review. *Archives of Business Research.* 2020; 8(6): 127-132.

[3] Shi K, Chang Z, Chen Z, Wu J, Yu B. Identifying and Evaluating Poverty Using Multisource Romote Sensing and Point of Interest (POI) Data : A Case Study of Chongqing, China. *J. Clean. Prod*. 2020; 255, 120245.

[4] Ghose T, Anderson SJ, Elvidge CD, Sutton PC. Using Nighttime Satellite Imagery as a Proxy Measure of Human Well-Being. Sustainability. 2013; 5: 4988-5019.

[5] Zhao N, Liu Y, Cao G, Samson EL, Zhang J. Forecasting China's GDP at the Pixel Level Using Nighttime Light Time Series and Population Images. GISci. Remote Sens. 2017; 54: 407-425.

[6] Yu B, et al. Urban Built-up Area Extraction from Log-Transformed NPP-VIIRS Nighttime Light Composite Data. *IEEE Geosci*. 2018; 15: 1279-1283.

[7] Yin J, Qiu Y, Zhang B. Identification of Poverty Areas by Remote Sensing and Machine Learning : A Case Study in Guizhou, Southwest China. *ISPRS Int. J. Geo-Inf.* 2020; 10: 11.

[8] Sapena M, Ruiz LA, Taubenbock H. Analyzing Links between Sapio-Temporal Metrics of Built-up Areas and Socio-Economic Indicators on a Semi-Global Scale. *ISPRS Int. J. Geo-Inf*. 2020; 9: 436.

[9] Dawson T, Sandoval JS, Sagan V, Crawford T. A Spatial Analysis of Relationship between Vegetation and Poverty. *ISPRS Int. J. geo-Inf*. 2018; 7: 83.

[10] Maurya R, Gupta PR, Sharma MK. Builtding Extraction from Very High Resolution Multispectral Images Using NDVI Based Segmentation and Morphological Operators. *In Proceedings of the IEEE-International Conference on Advances In Engineering, Science and Management (ICAESM-2012), Nagapattinam, India*. 2012; 577-581.

[11] Putri SR, Wijayanto AW, Sakti DA. Developing Relative Spatial Poverty Index Using Integrated Remote Sensing and Geospatial Big Data Approach: A Case Study of East Java, Indonesia. *ISPRS Int. J. Geo-Inf*. 2022; 11: 275.

[12] Zha Y, Ghao J, Ni S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Rem. Sens*. 2003; 24 (3): 583–594.

[13] Mia B, Bhattacharya R, Woobaidullah ASM. Correlation and Monitoring of Land Surface temperature, Urban Heat Island with land Use-land Cover of Dhaka City using Satellite Imageries. *Int. J. Res. Geogr*. 2017; 3: 10-20.

[14] Wang B, Tian J, Yang P, He B. Multi-Scale Features of Regional Poverty and The Impact of Geographic Capital : A case Study of Yanbian Korean Autonomous Prefecture in Jilin Province, China. *Land*. 2021; 10: 1406.

[15] Han C, Gu Z, Yang H. EKC Test of the Relationship between Nitrogen Dioxide Pollution and Ecconomic Growth-A Spatial Economic Analysis Based on Chinese City Data. *Int. J. Environ. Res. Public Health*. 2021; 18: 9697.

[16] Bakash K, Akmal T, Ahmad T, Abbas Q. Investigating the Nexus among Sulfur Dioxade Emmision, Energy Consumption, and Economic Growth: Empirical Evidence from Pakistan. *Environ. Sci. Pollut. Res.* 2022; 29: 7214-7224.

[17] Cammeraat E. The relationship between different social expenditure schemes and poverty, inequality and economic growth. *International Social Security Review.* 2020; 73(2): 101-123.

[18] Wulandari I, Pratama AAN. Analisis Pengaruh Dana ZIS (Zakat, Infak, Sedekah), Pertumbuhan Ekonomi, Angka Harapan Hidup, Rata-Rata Lama Sekolah dan Pengeluaran Perkapita Terhadap Tingkat Kemiskinan di Indonesia Periode 2010-2021. *Jurnal Ilmiah Ekonomi Islam.* 2022; 8(3): 3301-3309.

[19] Han J, Kamber M, Pei J. Data Mining Concept and Techniques. *Elsevier*. 2012.

[20] Zaki MJ, Meira W. Data Mining and Analysis. Fundamental Conceps and Algorithms. *Cambridge University Press*. 2014.

[21] Kumar L, Mutanga O. Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. *Remote Sens*. 2018.

[22] Sidhu N, Pebesma E, Camara G. Using Google Earth Engine to Detect Land Cover Change: Singapore as a use case. *European Journal of Remote Sensing*. 2018; 486-500.

[23] Zurqani HA, Post CJ, Mikhailova, EA, Schlautman MA, Sharp JL. Geospatial analysis of land use change in the Savannah River Basin using Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*. 2018; 175-185.

[24] Papilaya PPE. Aplikasi Google Earth Engine dalam Menyediakan Citra Satelit Sumber Daya Alam Bebas Awan. *Jurnal Penelitian Kehutanan.* 2022; 16(2): 96-103.

[25] GEE. *Sentinel-5P.* June 4, 2023. https://developers.google.com/earth-engine/datasets/catalog/sentinel-5p.

[26] Lou G, Chen Q, He K, Zhou Y, Shi Z. Using Nighttime Light Data and POI Big Data to Detect the Urban Centers of Hangzhou. *Remote Sensing*. 2019; 1-20.

[27] Li, Zhao-Liang, et al. Satellite-derived land surface temperature: Current status and perspective. *Remote Sensing of Environment*. 2013; 14-37.

[28] Zhou X, Guan H, Xie H, Wilson JL. Analysis and optimization of NDVI definitions and areal fraction models in remote sensing of vegetation. *International Journal of Remote Sensing*. 2009; 721-751.

[29] Malik MS, Shukla JP, Mishra S. Relationship of LST, NDBI, and NDVI using Landsat-8 data in Kandaihimmat Watershed, Hoshangabad, India. *Indian Journal of Geo Marine Sciences*. 2019; 25-31.

[30] Danubianu M. *Step by Step Data Preprocessing for Data Mining A Case Study*. Proceedings of the International Conference on Information Technologies (InfoTech-2015). Bulgaria "Stefan cel Mare" University if Suceava. 2015; 1-8.

[31] Mirkin, B. Clustering: A Data Recovery Approach. CRC Press. 2013.

[32] Prianto C, Bunyamin S. Pembuatan Aplikasi Clustering Gangguan Jaringan Menggunakan Metode K-Means Clustering. Bandung: Kreatif Industri Nusantara. 2020.

[33] Madhulatha TS. An Overview on Clustering Methods. *IOSR Journal of Engineering*. 2012; 719-725.
[34] Rokhimah W, Kismiantini. Analisis Clustering Tingkat Kerawanan Wilayah Terhadap Kasus Penyakit di Kabupaten Sleman dengan Metode K-Means. *Jurnal Kajian dan Terapan Matematika.* 2022; 25-37.
[35] Kemp SE, Ng M, Hollowood T, Hort J. Descriptive Analysis in Sensory Evaluation, Chapter 1: Introduction to Descriptive Analysis. https://doi.org/10.1002/9781118991657.ch1. 2018.
[36] Rezaee MJ, Eshkevari M, Saberi M, Hussain O. GBK-means clustering algorithm: An improvement to the K-means algorithm based on the bargaining game. *Knowledge-Based System*. 2021; 1-12.
[37] Ebru Ç, Astar M. A Microeconometric Analysis of Household Consumption Expenditure Determinants for Both Rural and Urban Areas in Turkey. *American International Journal of Contemporary Research.* 2012; 27-34.
[38] Indriningtyas S, Hasandy LR, Dewantoro BEB. Dinamika Konsentrasi Emisi Gas Karbon Monoksida (Co) Selama Periode Psbb Menggunakan Komputasi Berbasis Cloud Pada Google Earth Engine. *Majalah Ilmiah Globe*. 2021; 23: 35-42.
[39] Yudhistira MH, Indriyani W, Pratama AP, Sofiyandi Y. Transportation Network and Changes in Urban Structure: Evidence from the Jakarta Metropolitan Area. *Research in Transportation Economics*. 2018; 52-63.
[40] Howard L. The Climate of London: Deduced from Meteorological Observations Made in the Metropolis and at Various Places Around it. London: Harvey and Darton, J. and A. Arch, Longman, Hatchard, S. Highley [and] R. Hunter. 1833.
[41] EPA. Reducing Urban Heat Island: Compendium of Strategies - Cool Pavements. Washington DC: Environmental Protection Agency. 2008.
[42] Estoque RC, Murayama Y, Myint SW. Effects of Landscape Composition and Pattern on Land Surface Temperature: An Urban Heat Island Study in the Megacities of Southeast Asia. *Science of the Total Environment*. 2017; 1-10.
[43] Benedict, Jaelani LM. A Long-term Spatial and Temporal Analysis of NDVI Changes in Java Island Using Google Earth Engine. *IOP Conference Series: Earth and Environmental Science*. 2021; 1755-1315.
[44] Shofy YF, Adi W. The Impact of the Trans-Java Toll Road Development on Spatial Planning in the Northern Region of Java Island: A Study Utilizing NDBI and Google Earth Images. *Indonesian Journal of Earth Sciences*. 2023; A611-A611.
[45] Wang S, Gao S, Li S, Feng K. Strategizing the relation between urbanization and air pollution: Empirical evidence from global countries. *Journal of Cleaner Production*. 2020; 1-9.
[46] Adebayo, T.S., Akinsola, G.D., Kirikkaleli, D. et al. Economic performance of Indonesia amidst CO2 emissions and agriculture: a time series analysis. *Environmental Science Pollution Research 28*. 2021; 47942–47956.
[47] Pan W, Fu H, Zheng P. Regional Poverty and Inequality in the Xiamen-Zhangzhou-Quanzhou City Cluster in China Based on NPP/VIIRS Night-Time Light Imagery. *Sustainability.* 2020; 1-20.
[48] Afrianto, F. East Java Province GRDP Projection Model Using Night-Time Light Imagery. *East Java Economic Journal, 6(2)*. 2022; 208-223.
[49] Xu J, Song J, Li B, Liu D, Cao X. Combining night time lights in prediction of poverty incidence at the county level. *Applied Geography*. 2021; 1-7.
[50] Kanga S, et al. Understanding the Linkage between Urban Growth and Land Surface Temperature—A Case Study of Bangalore City, India. *Remote Sensing*. 2022; 1-17.
[51] Nurwanda, A., & Honjo, T. The prediction of city expansion and land surface temperature in Bogor City, Indonesia. *Sustainable Cities and Society, 52*. 2020; 101772.
[52] Wiratama, Budhi Fatanza, Robert Kurniawan, Mulyanto, Mohamad Andrian Isnaeni, Bagus Sumargo, and Prana Ugiana Gio . Measuring the physical infrastructure development as poverty reduction program in Kalimantan, Indonesia. *Cities*. 2023; 1-13.
[53] Ningtias, E. N. A., & Anwar, A. F. Mengukur Dampak Pengangguran, Tingkat Pendidikan, Upah Minimum, dan Pengeluaran Per Kapita terhadap Kemiskinan di Kota Makassar. *Bulletin of Economic Studies (BEST)*. 2021; 1(1).

AUTHOR'S BIBLIOGRAPHY

Azzahra Dhisa Khamila is a student of the DIV Statistics study program at the Polytechnic of Statistics STIS. Currently, she is in her final year of college with a specialization in social and population statistics.

Martha Budi Wardani is a student of the DIV Statistics study program at the Polytechnic of Statistics STIS. Currently, she is in her final year of college with a specialization in social and population statistics.

Robert Kurniawan is a lecturer and researcher at the Polytechnic of Statistics STIS, Jakarta. The results of his research are published in reputable national and international journals. In addition, he is active in writing statistics and data science books that are published nationally.