CHAPTER I
INTRODUCTION

1.1 Background

Since the 1960s, the world population has more than doubled to more than seven billion. It will probably continue to grow due to an improvement in the quality of life in numerous sectors, such as the availability of food, promotion of good health, and fulfilment of other fundamental human necessities. In the meantime, gross domestic product (GDP) has increased significantly, from a mere $2.3 trillion in the 1960s to over $80 trillion in 2020 (World Bank, 2022).

Scholarly opinions are divided as to whether economic or population growth occurs first. However, it is evident that in the previous three decades, there has been no association between the urbanization rate and the economic growth rate worldwide (Chen et al., 2014). Turok & McGranahan (2019) found no linear relationship between urbanization and economic growth or city size and productivity. Conversely, there is no doubt that urbanization affects land use and land cover (LULC) changes (Nedd et al., 2021; Talukdar, Singha, Mahato, Shahfahad, et al., 2020).

Urbanization typically happens in every nation at different times (Roberts & Kanaley, 2006). According to Hall (1971), urbanization occurs in four stages in developed nations' cities: (1) population moves from rural areas to crowded cities; (2) population growth increases in urban fringe areas; (3) population decreases in the core city to the urban ring; and (4) population increases again in the urban population (Carra & Barthelemy, 2019). In comparison, it could be different when the focus moves to less developed countries. Urban populations in developing nations are projected to expand by at least two billion by 2040 (Overman & Venables, 2010).

Urbanization is a continuing trend that is already beginning to decline in some areas. For the first time in human history, more than half the world's population lives in cities (Brenner & Schmid, 2013) which also has introduced the new term 'an urban age.' As per data released by World Bank, Indonesia has experienced an urban age since 2011. Moreover Dobbs & Remes (2013) found that the urban population globally is expected to grow by another billion people by 2025. Conversely, Kabisch & Haase (2011) found that a shift from growth to decline of urban population is already present in a growing number of cities. These findings suggest that urbanization is a continuing trend already beginning to fall in some areas.

Physically, urban expansion occurs both horizontally and vertically due to the increase in population density and fixed investment (Zhang et al., 2017). The less developed countries are accelerating their pace of urbanization and facing widespread unemployment, poverty, and housing shortages (Shen, 2006). Besides, Urbanization affects urban climate, that is, urban heat islands (Qian et al., 2022) and social vulnerability (Mesta et al., 2022), and also heavily influences the carbon emissions of the city (Zhu et al., 2019). Thus, it is crucial to analyze the growth patterns of cities to predict the future urban climate and carbon emissions. In short, urbanization is a process with several consequences, such as social, economic, or environmental (Dociu & Dunarinţu, 2012).

The studies on how urbanization affects social aspects in urban areas are varied. Many suggest that urbanization does not affect social elements. Buhaug & Urdal (2013) found that urbanization is not associated with a higher risk or frequency of social disorder, and Dempsey, Brown, & Bramley (2012) found that density does not influence residents' propensity to engage in socially sustainable activities. However, Mouratidis (2018) found that compact urban forms enable residents to maintain more extensive networks of close relationships, socialize more frequently with friends and family, receive more robust social support, and enjoy increased opportunities to make new acquaintances. Therefore, urbanization may positively affect social aspects in urban areas.

Regardless of how the social effects are still ambivalent for other scholars. Sato & Zenou (2015), for example, argue that urbanization may have a positive or negative impact depending on the level of urbanization. For a low urbanization level, there is a unique steady-state equilibrium where workers do not interact with weak ties. In contrast, there is a unique steady-state equilibrium with total social interactions for a high level of urbanization.

Effects on economics have been a hot discourse, especially on economic growth locus. Nguyen & Nguyen (2018) found that urbanization positively impacts economic growth, and Pradhan, Arvin, & Nair (2021) found that long-run economic growth depends on the co-development of policies to create a vibrant urban ecosystem. However, Yong Liu, Chen, & Liu (2020) found that urban compactness was negatively correlated with the urban GDP. This suggests that urbanization does not lead to economic growth.

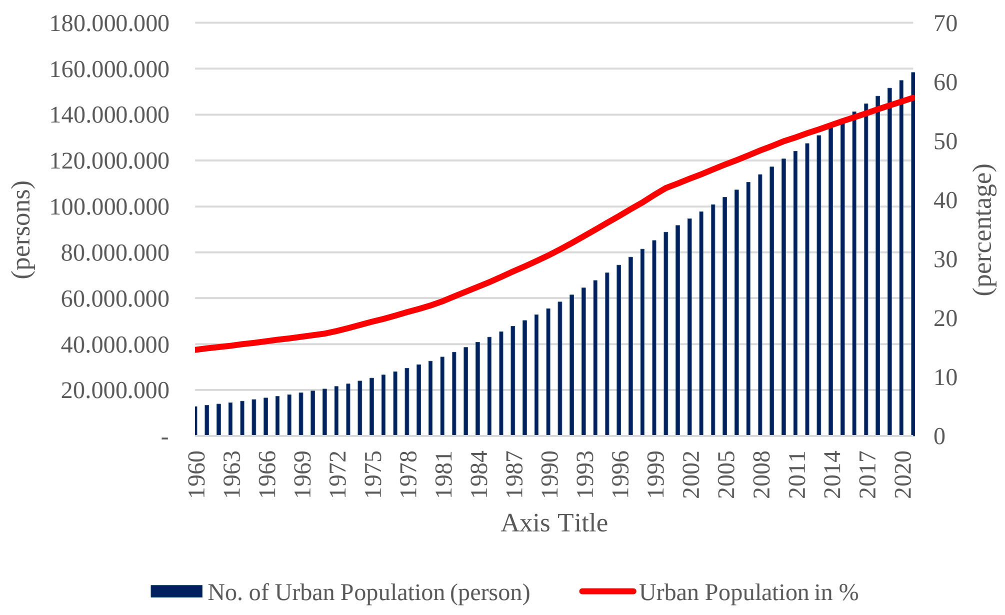
On the other hand, Gross & Ouyang (2020) propose a simple explanation for the relationship between urbanization and economic growth as a push-and-pull Harris-Todaro framework. People will be more inclined to relocate to cities due to the possibility of working and boosting their income due to urbanization. As a consequence, residual urbanization has an impact on economic growth.

All the scholars agree that urbanization harms the environment. Many studies focus on different topics. For example, Grimm et al. (2008) found that urban areas are hot spots that drive environmental change at multiple scales. McKinney (2002) found that urbanization harms native ecosystems. Then, Bai et al. (2017) found that the impacts of urbanization on the environment are amplified or accelerated. Besides, Tan Hoi (2020) reveal that in the case of Hanoi, the consequences of urbanization are poor air quality, more power consumption, and waste problem. To conclude, urbanization hurts the environment.

There is a complex relationship between urbanization and development. Zhao & Zhang (2009) argue that urbanization progress contributes to income and the patterns of development. Furthermore, mixed findings can be found when we discuss the relationship between urbanization and urban development. First, Schneider & Woodcock (2008) propose that the most common type of city is the high-growth city with rapid, fragmented development. Another study conducted by Wu (2006) found that environmental amenities' spatial distribution strongly influences development patterns and community characteristics. Therefore, we cannot definitively say the most common types of urban development.

Indonesia, the world's fourth most populous country, has witnessed significant urbanization over the past several decades (Figure 1.1). According to World Bank data, the number of urban residents in Indonesia has increased dramatically from fewer than 20 million (or less than 10 percent) in 1960 to approximately 160 million in 2021. Since 2011, more than 50 percent of the overall population of Indonesia has settled in urban areas. There is a chance that this phenomenon will continue to grow in the coming years.

Understanding changes in LULC monitoring requires precise and current data and reduces most of the difficulties associated with understanding continuous differences. It assists researchers and decision-makers in monitoring change, observing patterns, and managing land resources efficiently (Nedd et al., 2021). This study investigates the pattern and changes of land use and land cover of Surabaya metropolitan since 2000 until 2020 and predict the possible future changes.



(Source: data.worldbank.org)

FIGURE 1.1
Urban Population in Indonesia (1960-2021)

1.2 Problem Statement

Owing to the continuously increasing urbanization, stakeholders acting as policymakers must manage it well. Knowing the state of the land use patterns is essential for developing policies to improve urban management. Additionally, LULC changes undoubtedly follow certain patterns. Hence, studying LULC change, and its future prediction is necessary.

1.3 Objective

The ultimate objective of this study is to predict LULC changes in the metropolitan of Surabaya and the surroundings. To achieve this objective, several sub-objectives were carried out in stages in the previous study, which included clustering cities and regencies in Indonesia based on statistical data by utilizing the statistical analysis of principal component analysis (PCA) and cluster analysis. In addition, another objective that needs to be achieved is evaluating LULC changes from 2000 to 2021 using remote sensing data.

1.4 Scope

This study focuses on the metropolitan region of Surabaya and the surrounding districts of Sidoarjo, Mojokerto, and Gresik. In addition, the coastline and sea to the north and east of Surabaya are included in the scope of this study. The study region spans 7.1828-7.4556 south latitude and 112.4721-112.8648 east longitude from an astronomical perspective.

This study focuses on the prediction of LULC in Surabaya metropolitan area using ANN-CA after an initial study consisting of two stages. The first stage was the classification process of cities and regencies into groups, and the second stage was the prediction of LULC using the MLPNN method. This study concluded by comparing the ANN-CA and MLPNN methods. Finally, policy recommendations on the results of this study were presented at the end.

This study concentrated on the classification of land use and land cover, as well as changes from the past to the present. Then, at the same place, projections of prospective changes in the future were produced. Besides, it will only consider digital elevation models (DEM) and derivative maps in the form of aspects and slopes as change drivers. In addition, the road network linking various areas within the research region was considered.

1.5 Benefits

Changes in land use patterns that are the subject of research facilitate the monitoring of alterations in human activities in specific regions. In addition, by doing this monitoring, it is possible to determine if there is an increase or decrease in human activities that may be anticipated and managed, allowing for the formulation of policies that promote a more environmentally friendly lifestyle. For instance, a location's decision-making process will be simplified while determining whether to develop an area or preserve it as a native vegetation. In addition, locations with excessive rapid changing environment would be more easily identified. Monitoring these alterations aid planners in spatial planning.

Moreover, observing the pattern of changes in land use from the past to the present would disclose unique patterns, allowing for the projection of future possibilities at the same region. Utilizing the advancement of remote sensing technologies make long-term monitoring of changes more cost- and time-efficient.

1.6 Research Framework

The research framework of this study can be found in FIGURE 1.2.

1.7 Research Methodology

1.7.1 Data Collection

The process of data collection will be split into two sections, neither of which can be executed concurrently. Initially, statistical data collecting, and calculations are performed. Then, spatial data will be gathered depending on the recommendations of the statistical stage's outcomes. Following the completion of data collecting and statistical data analysis, the second phase will be executed.

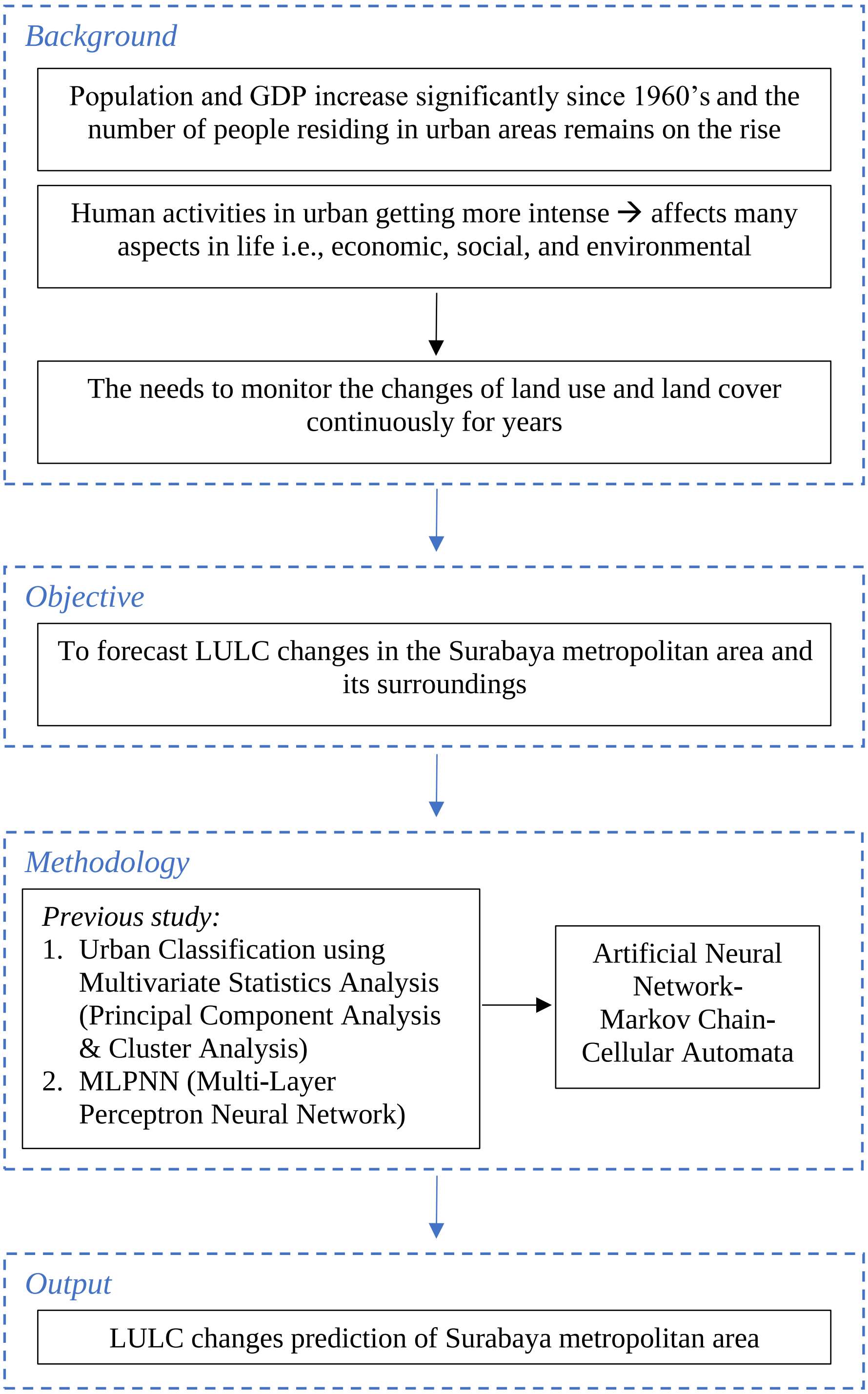


FIGURE 1.2
Research Framework

1.7.1.1 Statistical Data

The necessary statistical information was retrieved from the website of the Central Bureau of Statistic of Indonesia (CBS/*BPS Indonesia*). Source of data is the Regency/City document in Figures, which is published annually. From 2000 to 2020, data were collected for all relevant variables in all areas of research. The details listed in the TABLE I.1.

TABLE I.1
List of variables used in multivariate analysis

No	Variable	Data Duration (years)	Year (start–end)
1	Total area (km^2)	18	2003-2020
2	Number of districts	19	2002-2020
3	Number of villages	19	2002-2020
4	Total population	20	2001-2020
5	Population density (per km^2)	18	2003-2020
6	Number of households	16	2003-2019
7	Poverty line (Rp)	11	2010-2020
8	Number of poor people	12	2009-2020
9	Percentage of poor people (%)	12	2009-2020
10	Human development index	11	2010-2020
11	Harvested paddy area (Ha)	19	2002-2020
12	Paddy production (ton)	19	2002-2020
13	Number of large and medium industries	11	2007-2018
14	Number of employees in large and medium industries	4	2008-2011
15	Number of micro and small industries	1	2019-2019
16	Number of employees in micro and small industries	1	2019-2019
17	Number of industries	16	2005-2019
18	Number of employees in industries	14	2006-2019
19	Number of electric customers	19	2002-2020
20	Number of electric production (KWh)	6	2013-2018
21	Total consumption of electricity (KWh)	16	2004-2020
22	Total length of roads (Km)	18	2003-2020
23	Total number of registered motor vehicle	15	2003-2019
24	Monthly average expenditure per capita (Rp)	6	2015-2019
25	Gross regional domestic product by industrial origin at current market price (Million Rp)	20	2001-2020

1.7.1.2 Spatial Data

The spatial data used in this study were derived from secondary data available on multiple open-source platforms and were freely utilized. Data comprised satellite imagery with a medium resolution, road network maps, and a digital elevation model (DEM). The spatial data used in this study are described in TABLE I.2.

TABLE I.2
Spatial Data

Data	Spatial Resolution	Purpose	Year	Source
SRTM DEM	30m	spatial driver	2014	USGS https://earthexplorer.usgs.gov/
Landsat 7 Satellite Images	30m	to retrieve LULC maps	2000 & 2003	USGS https://earthexplorer.usgs.gov/
Landsat 8 Satellite Images	30m	to retrieve LULC map	2013 & 2021	USGS https://earthexplorer.usgs.gov/
Transportation Network	-	spatial driver	Existing	Open Street Map https://www.openstreetmap.org/

(Source: Analysis)

1.7.2 Analysis

This study's analysis was separated into two primary sections: statistical analysis and spatial analysis. Initially, statistical analysis was comprised of two interdependent components. PCA analysis was used to identify the primary variables (called factors) while preserving the primary properties of the data as a whole. Because the overall data type was numerical, PCA was performed. The results of the PCA would be utilized in the second statistical analysis procedure, hierarchical cluster analysis. This second step grouped cities and regencies based on data characteristics, meaning that characteristics within a cluster tend to be similar.

Secondly, the spatial analysis was conducted upon completion of the statistical analysis and interpretation of the data. As depicted in table IV.2, the spatial analysis leveraged data derived from open sources accessible on multiple website platforms.

1.7.2.1 Principal Component Analysis

The primary objective of the PCA is to reduce the number of variables while retaining the variables that influence the data, as reflected by the data variance value. Different data units influence the weighting of data variants during the initial stage of the PCA, which is data normalization. For instance, data on the number of villages (without units) with values ranging from 5 to 25 have a far narrower data range than the data on monthly per capita income with a range in the hundreds of thousands of rupiah. If statistics are not normalized, it will be evident that the monthly per capita income dominates the final result of calculation. To prevent this bias, the data were normalized.

In general, the PCA stage consists of four steps in addition to the data normalization procedure performed during data cleaning. The PCA steps can be observed in Figure 4.1. The calculations in this study employed R program with Rstudio support (IDE). This program is freely accessible and open source. In particular, the *factoextra* package was the focus of this study's analysis procedure.

FIGURE 1.3 demonstrates that the PCA begins with calculating the covariance matrix to establish the relationship between variables. If the link between two variables is too close (the value is 1), the variables may be redundant. Positive or negative covariance values could exist. A positive value indicates that two variables concurrently rise or decrease in value or that they are connected. Conversely, if the value is negative, the relationship will be inverse.

The covariance matrix is an n × n symmetric matrix, where n is the number of variables. The diagonal from upper left to lower right represents the variance of each data set. The remainder is the covariance value of each, which is symmetrical because the nature of covariance (a,b) is identical to that of covariance (a,b) (b,a). Therefore, this research yielded a 40-by-40 covariance matrix.

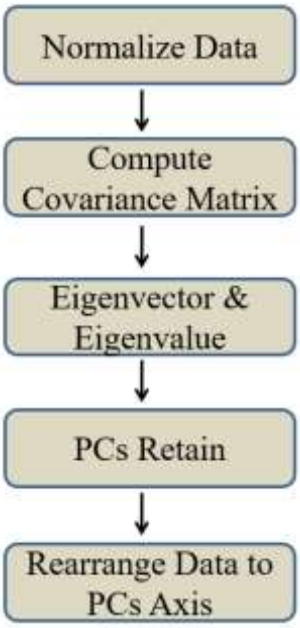


FIGURE 1.3
Steps in PCA calculation

The PCA's second step was to identify the primary components (PC). This stage consisted primarily of performing calculations on the covariance matrix to obtain eigenvectors and eigenvalues. PC was an extraction or summary of the starting variable. PC1 represented the most information that could be recovered from the calculation, followed by PC2, and so on, until no more information could be extracted. Typically, the results of this procedure are presented in the form of a scree plot, which also indicates the amount of variation covered by each PC. This procedure produces a PC with decreased dimensions without losing a substantial amount of vital data.

In the form of a two-dimensional Cartesian graphic, PC1 is derived from a line that optimizes variance (mean of the sum of the squares of the distance between the projection point and the original data). PC2 is the outcome of the same computation as PC1 but with a perpendicular orientation to PC1. PC3 and other PCs perform the same operation until the value equals the number of variables in the initial data set.

Next, eigenvalues and eigenvectors always coexist, and their sum remains constant. Therefore, there will be 10 eigenvectors if there are 10 eigenvalues. This study produced forty eigenvalues and forty eigenvectors. Eigenvectors are the direction of the axis along which the greatest amount of information (variance) exists, also known as the principal component. Moreover, eigenvalues are coefficients associated with eigenvectors that represent the variance of each principal component. The variance value can be determined by dividing the percentage of eigenvalues in each PC by the sum of eigenvalues in all PCs. The variation is expressed in percentage units.

The third stage of the PCA involved determining how many PCs were preserved for the subsequent round of analysis. At this point, the process consists of reducing the dimensionality of the existing data. The initial n-dimensional data was reduced to p dimensions. In this study, dimension reduction was determined by the minimum coverage value of at least 50 percent of the data variance. Therefore, the cumulative number of PCs to be maintained must be at least 50 percent.

The final step involved reorienting the data from the original axis to the axis that reflects the major component. This can be quantitatively determined by multiplying the transpose of the initial data set by the transpose of the feature vector.

PCA delivered the number of the most influential elements and the amount of data variance represented by the numerical solution. The PCA findings were also shown in a scree plot for simplified understanding and reading. The value of the contribution of each variable to each principal component was an additional outcome that was crucial for interpreting data. This information can be tabulated by displaying the value of each variable in each PC.

1.7.2.2 Cluster Analysis

The results of the PCA were used into the second statistical analysis, cluster analysis. Clustering used the same tools as principal component analysis (PCA), namely Rstudio with a modified script. As a result, a dendrogram was created to illustrate the degree of proximity between each group with the same or closest characteristics.

Clustering is a method used to divide observations into distinct groups, where "within a group" refers to the observation data that is most identical, and "different groups" demonstrate something that is most likely divergent from the other groups.

Clustering can be separated into k-means clustering and hierarchical clustering. The distinction between these two strategies is based on whether or not the number of groups or classes was decided beforehand. When the number of groups or clusters is known beforehand, k-means clustering is the most appropriate technique. If, however, the number of classes or clusters has not been defined at the outset or is unknown, hierarchical clustering is preferable. In this work, R will be used for the final clustering analysis and calculation.

Essentially, clustering utilized distance to divide the set data into groups. There were five major clustering stages. First, each point was placed within its respective cluster so that each group comprises itself. Based on the distance matrix computation, the second step was to assign the two nearest locations into one group. This second stage has the effect of reducing the number of clusters by one. Next, the distance between the new and old clusters was recalculated and stored as a new distance matrix for use in the last step. Finally, the first and second procedures were repeated until a big cluster, including all the data sets, was formed.

How distance is calculated differs between methods of grouping. This research was conducted using Ward's criterion. This technique is often referred to as minimum variance. Ward's criterion minimized the total variance within the cluster and identifies cluster pairs that result in a small increase in the total variance of the merged cluster.

This study concentrated on cluster analysis utilizing Ward's methods. The clustering procedure was based on the PCA results from the preceding phase of study. The objective of hierarchical clustering with wards was that to minimize the number of squared differences between each cluster. The results were represented by a *dendogram*, and the researcher might decide, based on the number of *dendogram* branches, at what level of clusters to be retained.

1.7.2.3 LULC Classification

The primary objective of a LULC analysis is to capture land use patterns in an area through times. Then, a comparison was conducted to determine the pattern of changes over time at the same location. Finally, the pattern of change was obtained as a model and utilized to create future predictions so that several alternatives could be suggested to anticipate the phenomena of change that might occur.

This study's LULC maps were obtained directly from Google Earth Engine (GEE), a free platform accessible to anyone who registers with the platform. A further advantage of using this platform is that users only require a browser and a computer connected to the internet network. Users' scripts are easily executable without the requirement of a powerful computer because the computing is performed in the cloud by the GEE server. Users must write scripts based on their demands, and the results can be provided in a variety of methods; for instance, the results of this study were directly saved in *.tiff format and stored on a cloud drive server.

At the beginning of the procedure, a masking layer is called for and created for use in the image's initial cloud cover reduction. This method employed the BQA band offered by each satellite. The next step was to request the necessary satellite imagery, in this case, Landsat 7 and 8 (according to the year required by this study). GEE retrieved the data directly from the United States Geological Survey (USGS) website.

The next step was browser-based picture visualization. At this step, the shape of the image display could be modified by altering the band combination of each image. For instance, Landsat 8 uses a combination of 5-4-3 bands (for the combination of Red-Green-Blue bands) to display true colour image, whereas other combinations display false colours, which can be useful for the sampling process. This study split the LULC class into four categories, as shown in the TABLE I.3, along with a description of each category.

TABLE I.3
LULC Classes

No	LULC Class	Description
1	Vegetation	All types of vegetation and green areas, trees with a canopy
2	Water	Lake, sea, pond, river, etc.
3	Urban	All built-up areas, roads, asphalt, etc.
4	Agriculture/Barren	Paddy field, barren land, shrubs, etc.

Next, moving to the sampling for each land cover class. This sampling was conducted manually by digitization and was then categorized by land cover type. Sampling was accomplished by selecting one or a combination of the four possibilities available; points, lines, rectangles, and polygons. This sampling was restricted to a maximum sample size of 5000 pixels. The notion might be applied to the study place until it has evenly permeated the area. After it has been completed for the full land cover class, it was turned into a single variable, the land cover class.

The next step was to evaluate the precision using the confusion matrix. The accuracy evaluation was based on the usage of Google-owned, server-accessible high-resolution satellite imagery. Users could digitize the image to compare the findings of the previous classification procedure with this image because the land cover was plainly visible.

The confusion matrix is a comparison of the values of the estimates in the LULC classification based on data that are believed to accurately represent actual field representations (in this study, high-resolution satellite imagery at the study site). As there are four types of land cover classes, the confusion matrix is a 4 x 4 matrix. Due to the estimated and real values, the diagonal matrix value is frequently near 1. The accuracy is determined by dividing the number of valid pixels (total diagonal sides) by the total number of samples used in the calculation process. This quantity is expressed as a percentage.

The final phase of LULC classification was exported to the specified file format (GeoTIFF). This could be accomplished concurrently by issuing commands to accessible scripts. The results then were saved to the cloud drive based on the specified location. The GeoTIFF format was chosen because this file type is easily readable by many spatially aware data processing applications, such as QGIS.

The LULC map was the outcome of Landsat 7 and Landsat 8 analysis as the primary data input. The Shuttle Radar Topography Mission (SRTM), Digital Elevation Model (DEM) was then utilized to create elevation maps, slope maps, and aspect maps, which were employed as spatial drivers in the formulation of the change matrix. Finally, road data was extracted from OSM (openstreetmap.org).

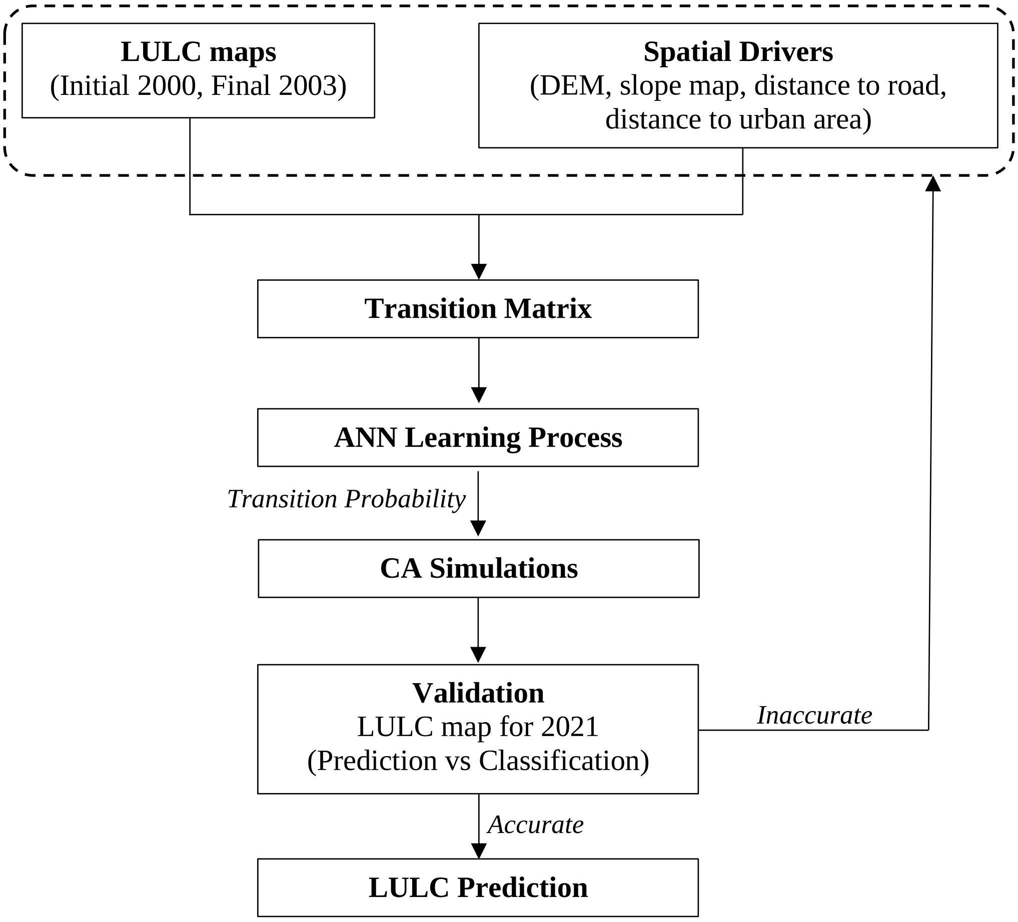
1.7.2.4 LULC Prediction

Using the opensource program QGIS, the LULC projection used a plugin, notably MOLUCSE (https://plugins.qgis.org/plugins/molucse/). This study specifically employed the artificial neural network (ANN) supplied by the plugin. The plugin executed the following procedure, in brief:

1. takes a raster of the land use categories for the time periods A (past), B (current), and raster of the explanatory variables or factors.
2. builds a model that forecasts changes in land usage from the past to the present.
3. based on the existing level of land use, current circumstances, and a derived model, forecasts future changes in land use.

The prediction of LULC process covers mainly six steps as illustrated in FIGURE 1.4. It can be seen that the input of all steps consists of two data types, LULC maps and spatial drivers. The LULC maps used here were LULC map from the year 2000 as an initial input while the one from 2003 as a final input. Besides, spatial drivers were added to the input which derived from the DEM and road network. All the input data might have the same size in terms of pixel number. Otherwise, the process could not be proceeded to the next step.

The transition probability for LULC was computed using ANN. This was accomplished by mimicking several classes within LULC. Meanwhile, CA was used to simulate changes in land use based on the probabilities of transition given by ANN learning processes.



(Source: Saputra & Lee, 2019, modified)

FIGURE 1.4
Procedure for LULC Analysis

The initial stage was to determine the simulated inputs for an ANN. This study utilized four land cover classifications and five distinct spatial factors. The ANN was fed nine spatial variables in total. Each pixel would have nine spatial variables due to the usage of cell-based or pixel-based simulation in this analysis. Note that the spatial resolution of all input data is 30 arc seconds.

The second stage was to assess the differences between the initial and final LULC maps. Every pixel would undergo a two-way raster comparison. This procedure concludes with a transition matrix depicting the shift in pixel scale from one variable class to another for the study area.

The subsequent phase involved modelling with ANN. At this point, the topology of the neural network consists of three major components: input, hidden, and output layers. The input layer was the previously described spatial variable that was represented by a neuron. Each neuron was transformed to a value that falls between 0 and 1. Meanwhile, the hidden layer was the signal received by the *j*-th neuron from the input layer at the *k*-th cell at time *t*, as denoted by $net_j(k,t)$. The hidden layer was expressed by equation (1). The hidden layer could be expressed as follows:

$$net_j(k,t) = \sum_i w_{i,k} x'_i(k,t) \qquad (1)$$

where;

$w_{i,k}$ = the weight between the input and the hidden layers, and

$x'_i(k,t)$ = the *i*-th scaled characteristic associated with the *i*-th neuron in the input layer relative to the *k*-th cell at time *t*.

Furthermore, the output layer could be written according to equation (2). The result of the equation was a value which means that the higher the value, the higher of transition probability from the initial type of LULC to the *l*-th type. LULC change was calculated by comparing the values of transition probability such that LULC converted from the current type to the form with the highest value of transition probability. If the transition probability for the same type of LULC was the highest, the condition of the corresponding point remained constant.

$$P(k,t,l) = \sum_j w_{j,l} \frac{1}{1+e^{-net_j(k,t)'}} \qquad (2)$$

where;

$P(k,t,l)$ = at time *t*, the chance of conversion from the current class to the *l*-th type of LULC for the *k*-th cell, and

$w_{j,l}$ = the difference in density between the hidden and output layers.

The fourth step of this procedure was a CA simulation utilizing the transition probabilities acquired in the previous stage. This CA simulation took the value of a cell and the values of its neighbouring cells into consideration. CA also repeatedly run iterations or calculations that will decide whether or not there was a change in a single cell at the final stage time *t*. The set threshold determined whether a change was possible or not. If the result exceeds the threshold, then an alteration would occur, and vice versa.

The validation computation constituted the fifth step of the prediction process. In this study, the validation procedure compared the simulated value to the actual value, namely the LULC coming from the classification procedure, using the kappa coefficient (LULC 2021). The kappa coefficient could be determined using equation (3), and could be calculated in considerable detail using equations (4) and (5).

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \quad (3)$$

$$p_0 = \sum_{i=1}^{c} p_{ij} \quad (4)$$

$$p_e = \sum_{i=1}^{c} p_i T_p T_j \quad (5)$$

where;

p_0 = the percentage of matches observed,

p_e = the fraction of occurrences predicted to occur,

p_{ij} = the *i*-th and *j*-th cell of the contingency table,

p_iT = the sum of all cells in *i*-th row, and

$_pT_j$ = the sum of all cells in *j*-th column.

Note:

A contingency table is a matrix that illustrates the relationship between the *i*-th and *j*-th cells by describing the distribution of variables utilized in the study. This matrix's representation provides an explanation of the degree of agreement on the criteria in each cell. The Kappa coefficient is represented as a percentage in this matrix.

The final step was the LULC prediction for the predefined years if all preceding stages match the minimum conditions. Using iteration calculations, this forecast translated to multiples of years based on the initial input. In this study, one cycle corresponds to three years (initial year 2000 and final year 2003).

1.8 Research Systematic

This research is divided into five interdependent chapters, as described below.

CHAPTER I INTRODUCTION

The first chapter provides an introduction where the general scope of the study and related issues given. Then it narrows down to the problem statement and research questions. Besides, this chapter also give the scope and research framework.

CHAPTER II LITERATURE REVIEW

Literature review provides the detail theories and studies conducted by scholars. It synthesizes and presents from the general ideas to the details which are used in this study. It is limited to the research on land use and land cover classification and prediction.

CHAPTER III STUDY AREA OVERVIEW

This chapter mainly focus on reasoning why the study location is chosen and the urgency why it should be conducted in this specific area. This chapter delivers the statistical data compilation taking from the official bureau in Indonesia.

CHAPTER IV ANALYSIS

This chapter begins with the summary of previous study conducted by author which related to this study. The study includes multivariate analysis using PCA and cluster analysis and the prediction of LULC in Surabaya using MLPNN. Then, the next sub-chapter presents the analysis of LULC using ANN-CA in the same location. Finally, the comparison between the methods will conclude in the end of this chapter.

CHAPTER V CONCLUSION

In the conclusion, the study's summary and limitations are presented. Additionally, the recommendations and future works conclude the chapter.