

Categorizing Toll Gates using Data of Sentinel-1, Sentinel-2, Sentinel-5P, NOAA-20 Satellite, and Objects on Google Map: A Case Study in Indonesia

Moertini, V. S.,^{1*} Ryanda, D.,¹ Wiratna, L.¹ and Pertiwi, A. P.²

¹Center for Data Science and Artificial Intelligence System, Informatics Department, Parahyangan Catholic University, Bandung 40141, Indonesia

E-mail: moertini@unpar.ac.id,* dittoryanda@gmail.com, lionov@unpar.ac.id

²The Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 12489 Berlin, Germany
E-mail: avi.pertiwi@dlr.de

*Corresponding Author

DOI: <https://doi.org/10.52939/ijg.v22i9.5190>

Abstract

This study proposes a framework to categorize toll gates based on the development level of their surrounding areas and localized economic activity profiles. The methodology integrates multi-source satellite imagery with geospatial features extracted from the Google Maps API. The localized economic context of 419 toll gates was analysed within a 1-km circular buffer zone. Within this radius, Sentinel-2 spectral bands (B2, B3, B4 and B8) extracted from Google Earth Engine (GEE) were classified into built-up and non-built-up pixels using a Random Forest algorithm to quantify the built-up area percentage. Gates with a built-up density of at least 40% were categorized as Class-1, representing infrastructure in highly developed regions. For the remaining gates, Sentinel-1 VV polarization, Sentinel-5P tropospheric NO₂ column density, and NOAA-20 VIIRS nighttime lights data were extracted via GEE, alongside point-of-interest (POI) counts from Google Maps. Following feature engineering across five initial variables, the backscatter coefficient (VV) and POI density were identified as the optimal features for clustering. Using these features, a k-Means algorithm grouped the remaining gates into Class-2 and Class-3 clusters, whose final socioeconomic profiles were interpreted using all variables. The classification yielded 297 Class-1 gates (highly developed areas with dense human and commercial activities), 51 Class-2 gates (moderate activity zones), and 71 Class-3 gates (low-to-moderate activity zones). To assess the long-term impacts of toll roads on regional economic growth, Class-3 gates warrant continuous observation, as their surrounding economic activity remains suboptimal. This research contributes to the fields of applied remote sensing and transportation infrastructure mapping by providing a cost-effective, automated methodology to periodically monitor spatiotemporal economic shifts around critical transport infrastructure.

Keywords: Classification, Clustering, Economic Activity, Monitor Toll Gates, Multi-Source Satellite Imagery, Spatial Objects Analysis, Spatiotemporal Economic

1. Introduction

In the last decade, Indonesia has constructed toll roads on a massive scale. By the first half of 2024, 2,839 kilometres of toll roads were built spanning Java, Bali, Sumatra, Kalimantan, and Sulawesi [1]. Because road infrastructure is widely recognized for its ability to boost economic growth [2] and [3], central and regional governments, alongside local communities, expect these benefits to materialize. Therefore, it is essential to assess the impact of these toll roads, especially at a localized level, to determine whether the growth benefits are being realized, as previously investigated in studies such as [4] and [5]. One way to determine the economic impact of toll

roads on the regions they traverse is to assess the areas surrounding entry and exit gates. The Head of Regional Infrastructure Development Agency (BPIW) tasked with formulating technical policies and encouraging integration between regional development and the development of infrastructure nationally under the Ministry of Public Works and Public Housing of Indonesia (PUPR) stated that new economic centres (such as restaurants, stores, and residential areas) have emerged around Trans-Java toll entries and exits, subsequently increasing regional income [6]. As of August 2024, there were 592 toll gates across Indonesia's toll road network

[7]. Considering that these gates are geographically dispersed across several islands, assessing their economic impact using traditional data collection methods, such as surveys and interviews, would be both time-consuming and costly.

To address this challenge, this study proposes a system that categorizes toll gates based on the level of surrounding area development and other economic activity profile. The system utilizes satellite imagery to assess whether the areas surrounding toll gates have shown significant development. This offers a scalable and cost-efficient alternative for monitoring economic impact across dispersed toll road network in Indonesia.

Various big datasets produced by satellite Earth observation systems are widely available in the cloud. Google Earth Engine (GEE), a prominent cloud-based platform, provides extensive geospatial datasets as well as functions to process them [8]. The datasets include those produced by the Sentinel-1, Sentinel-2, Sentinel-5P as well as NOAA-20 satellite systems. Sentinel-1, with its C-band Synthetic Aperture Radar (SAR), is an active system that penetrates cloud cover and operates independently of solar illumination in generating time-series data [9]. Consequently, Sentinel-1 data can be analysed to monitor regional development resulting from physical infrastructure changes [9] and [10]. Meanwhile, data from the Sentinel-2 satellite [11] has been widely used for monitoring land-use and land-cover changes caused by infrastructure development [12] and [13].

Additionally, the Sentinel-5 Precursor (S-5P) system generates NO₂ concentration data related to industrial pollution and fossil fuels burning by vehicles [14] and [15]. This data has been successfully analysed to measure localized economic activities [16]. Finally, NOAA-20 satellites carry the Visible Infrared Imaging Radiometer Suite (VIIRS), which includes a specialized Day/Night Band (DNB)

designed for global night-time light (NTL) observation [17]. VIIRS data has been used for tracking the economic development in specific areas [18] and [19]. Google Maps users have geolocated numerous Points of Interest (POIs) on the map, such as restaurants, cafes, stores, and gyms. Based on our observations, the presence of these POIs around toll gates (Figure 1) directly relates to local economic activities where a greater number of POI objects indicates a higher level of economic activity. Using the Google Places API [20], these can be queried, processed and analysed for research purposes. From an urban and regional planning perspective, transportation hubs (such as airports, harbours, train stations, and bus terminals) are indispensable pillars of economic infrastructure. They do not merely exist as transit points; rather, they function as economic engines that support diverse commercial ecosystems [21] and [22].

The previously discussed satellite data, Google Maps objects, and transportation hub data can be collected, processed, and integrated to categorize toll gates based on the economic activities surrounding them. Using machine learning algorithms, an automated or semi-automated system can be built to analyse this combined data and generate gate classifications, where each class represents a specific economic condition or profile. Considering that satellites data are time series and added periodically in the GEE system, Google Map API can be used to obtain the most current objects, and the current transportation hubs can be obtained from the government website, the system can be used periodically or whenever needed to categorize all toll gates in Indonesia. Over time, these toll gates can be observed to determine whether class changes occur, signalling shifts in surrounding economic activity. Since the required data are cloud-based and mostly open-access, this monitoring system can be executed rapidly and at a low cost.

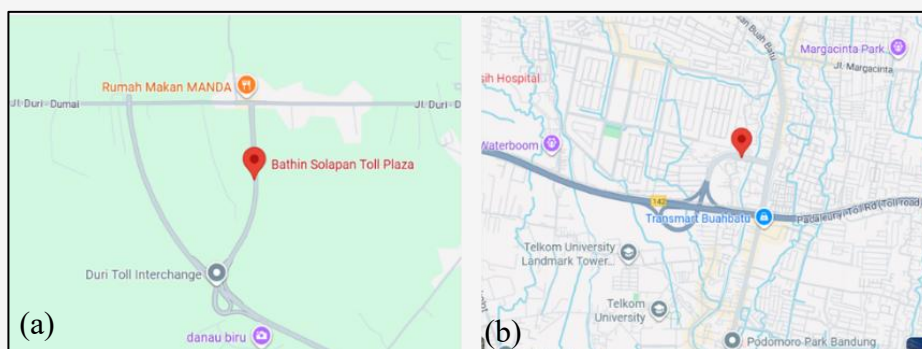


Figure 1: Gate locations: (a) Batin Solapan gate with few objects, and (b) Buah Batu gate with several objects

The focus of this research is to design a smart system architecture employing classification and clustering models, which includes developing the models and experimenting with a one-time collected dataset of Sentinel-1, Sentinel-2, Sentinel-5P, NOAA-20 VIIRS, Google Maps objects, and transportation hubs. The results of the categorization namely, the toll gate classes are then interpreted and analysed. The case study is set in Indonesia, where the majority of toll gates are located in developed areas near big cities or regencies with dense populations. However, the gates that most critically require monitoring are those located in rural or peri-urban areas. Thus, in designing the system, these spatial conditions are taken into account to ensure system efficiency. To the best of our knowledge, such system does not exist yet. A review of existing literature reveals that system development related to toll roads focuses mainly on: (a) the automation of payments and gate management [23][24] and [25]; (b) traffic prediction models [26]; and (c) user acceptance studies regarding toll gate technology [27]. Our research aims to address this gap by designing a system that can be used by central or local governments, as well as institutions that need to monitor the impact of toll roads on regional economic development.

2. Data and Methodology

2.1 Study Area

The study was conducted in Indonesia (Figure 2) during the 2024–2025 period, at which time the country's toll road network spanned the islands of Sumatra, Java (Jawa), Bali, Kalimantan, and Sulawesi (labelled in purple in Figure 2). Consequently, data for this study were collected from these islands and analysed using the proposed method. The numbers in the circle show the number of toll gates on those areas.

2.2 Google Earth Engine (GEE)

GEE is a cloud-based platform that makes it easy to access high-performance computing resources for processing very large geospatial datasets [8]. Earth Engine provides a multi-petabyte analysis-ready data catalogue co-located with a high-performance parallel computation. The catalogue is accessed through an application programming interface (API) and a web-based interactive development environment (IDE). The Earth Engine public data catalogue is a multi-petabyte curated collection of widely used geospatial datasets, including Sentinel-1, Sentinel-2, Sentinel-5P, and NOAA-20. The catalogue is continuously updated at a rate of nearly 6000 scenes per day from active missions. Related images are grouped together and presented as a “collection” [28]. Collections provide fast filtering (querying) and sorting capabilities that make it easy to search through millions of individual images to select data that meets specific criteria. For example, a user can easily select images from Sentinel-2 in the last year, at a specific area, with less than 10% cloud cover.

2.3 Multi-Spectral Sentinel-2 Data

Sentinel-2 (S-2) satellite is a polar-orbiting, optical mission for land and coastal region monitoring and emergency services. The mission is dedicated to monitor agriculture, forests, land-use change, land-cover change, coastal and inland waters, and so on. Frequent revisits of five days at the equator require two identical Sentinel-2 satellites (Sentinel-2A and Sentinel-2B units). Each one measures reflected radiance in 13 spectral bands spanning from the Visible and Near Infrared (VNIR) to the Short Wave Infra-Red (SWIR) spectral range, with high spatial resolution (10 to 60 m) optical imagery. The 13 bands are described on Table 1 [11].

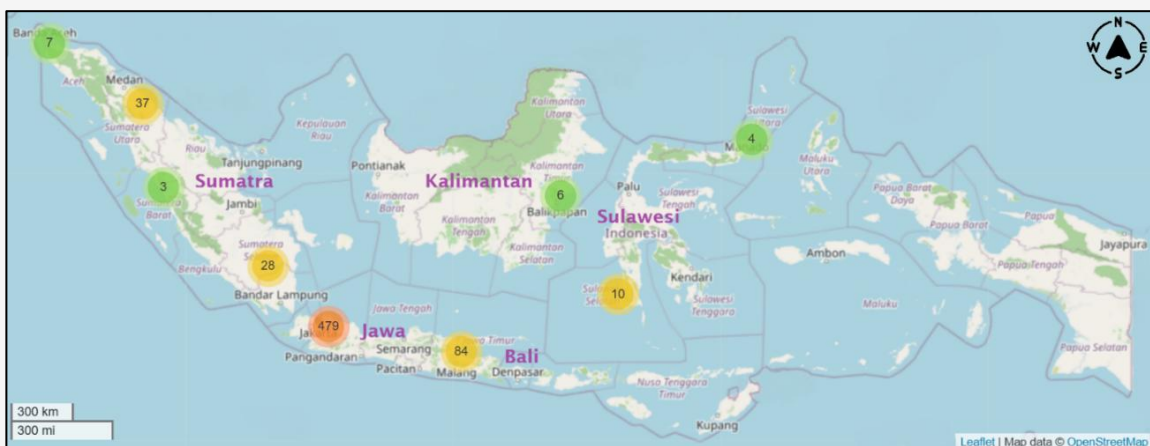


Figure 2: The five Indonesian islands in the study

Table 1: Sentinel-2 spectral bands and resolution

Band	Resolution (m)	Central Wavelength (nm)	Description	Band	Resolution (m)	Central Wavelength (nm)	Description
B1	60	443	Ultra Blue (Coastal and Aerosol)	B8	10	842	VNIR
B2	10	490	Blue	B8a	20	865	VNIR
B3	10	560	Green	B9	60	940	SWIR
B4	10	665	Red	B10	60	1375	SWIR
B5	20	705	VNIR	B11	20	1610	SWIR
B6	20	740	VNIR	B12	20	2190	SWIR
B7	20	783	VNIR				

Sentinel-2 data has been widely used in various fields, including forest and wetlands, water, agriculture, and urban areas [12][29] and [30]. Machine learning with its classification algorithms have been used to analyse Sentinel-2 data bands [12] and [31]. In this study, the classification of built-up and non-built-up areas was conducted using a subset of Sentinel-2 multispectral bands, specifically B2 (blue), B3 (green), B4 (red), and B8 (near-infrared). The selection of these four bands is based on four considerations:

First, the preservation of native spatial resolution. As shown on Table 1, Sentinel-2 bands are distributed across three spatial resolutions (10 m, 20 m, and 60 m). B2, B3, B4, and B8 are the only bands captured at a native 10 m spatial resolution, which is critical for identifying fragmented urban features, narrow road networks, and small-scale built-up structures.

Second, spectral relevance for the binary classification. B2, B3, B4, and B8 are selected to differentiate between built-up and non-built-up area with the following reasons [12][13] and [32]: (1) Due to its sensitivity to water or moisture content, B8 can differentiate between vegetation, water bodies, and urban with built-up areas; (2) B2 combined with B8 is useful to identify water surface; (3) B3 supports in assessing surface with vegetation; (4) B4 combined with B8 is commonly used for vegetation analysis.

Third, computational efficiency. Reducing the input dimensionality from 13 to 4 features significantly diminishes the computational overhead during the training (with sufficiently good results) and inference phases. This reduction is particularly beneficial for the classification algorithms, such as Random Forest and Support Vector Machine (SVM) where the training complexity scales quadratically to cubically

with the number of samples and features [33]. Using bands with 20 to 60 m resolution (such as SWIR B1, needed in Normalized Difference Built-up Index computation), along with 10 m bands will require spatial resampling or pansharpening processes that increases computation complexity.

Fourth, low cost on production stage. Classification models created with those 4 features can be executed efficiently on cloud platforms, such as GEE, thereby reducing or even avoiding computational costs. Therefore, the models can be used by broader users.

Cloud Coverage Issue: In tropical regions, such as Indonesia, cloud coverage is the most significant obstacle to optical remote sensing because these areas are characterized by high humidity and intense solar heating [34]. Obtaining a single cloud-free "wall-to-wall" mosaic of Sentinel-2 for the entire country within a calendar year is difficult. To overcome the persistent cloud coverage in tropical regions, this study employed a pixel-based temporal cloud masking strategy within the GEE environment, by applying a median reducer to the Sentinel-2 stacks. When `ee.Reducer.median()` in Google Earth Engine (GEE) is called, a pixel-wise temporal aggregation is performed. This process transforms a 3D "stack" of images (spanning several months or a year) into a single 2D "mosaic" by selecting the middle value for every individual pixel location. The median computation is excerpted as follows [8]: GEE aligns all images in the selected ImageCollection that intersect with the study area (in this research, circles area with toll gate coordinate as their centres and with radius of 1-km). If Sentinel-2 passes N times over a period, GEE creates a "stack" where each pixel coordinate (x,y) has N different values for each spectral band. For each pixel in (x,y) :

- a) GEE collects all valid (unmasked) values for that pixel, for instance: $V = \{2100, 2500, \dots, 2200, 4500, 1800\}$
- b) The values of V are sorted in ascending order, for example: $V_{sorted} = \{1800, 2100, \dots, 2200, \dots, 2500, 4500\}$.
- c) The median is selected from V_{sorted} , for instance, the result is 2200.

The median is a robust estimator as it ignores the extremes value (as long as more than 50% of the images are relatively clear).

2.4 Sentinel-1 C-band Synthetic Aperture Radar (SAR) Data

Sentinel-1, part of the European Space Agency's (ESA) Copernicus program, provides a paradigm shift in Earth observation through its C-band Synthetic Aperture Radar (SAR), which operate with frequency of 5.405 GHz. Unlike optical sensors, Sentinel-1 is an active system that penetrates cloud cover and operates independently of solar illumination, ensuring consistent time-series data even in tropical or high-latitude regions. The product of Ground Range Detected (GRD) includes polarization data of VV, VH, HH, HV, HH, and VV ('V' denotes vertical and 'H' is horizontal) that capture the physical and dielectric properties of the Earth's surface. Those data can be used for monitoring land-cover changes, industrial expansion, and infrastructural development [9]. GEE has prepared ready data within its catalogue. Sentinel-1 GRD data in the catalogue are retained in linear scale (S1_GRD_FLOAT). Data can be selected in single polarization i.e., VH or VV or dual polarization mode i.e., VV and VH [35].

In the context of area development, VV and VH can be used to monitor urban environment [9] and [10]. VV polarization represents signals transmitted and received in the same vertical orientation, often providing strong backscatter in smooth surfaces like water bodies and urban areas. On the other hand, VH polarization involves cross-polarized signals, which are more sensitive to volume scattering from vegetation and complex surface structures [36]. VV and VH can also be integrated with Sentinel-2 data to build classification model for land cover mapping [37].

In GEE, the polarization data is provided in decibel (dB) unit with resolution of 10, 25, or 40 meters. The data provides measures whether the radiated terrain scatters the incident microwave radiation preferentially away from the SAR sensor ($\text{dB} < 0$) or towards the SAR sensor ($\text{dB} > 0$).

This scattering behaviour depends on the physical characteristics of the terrain, primarily the geometry of the terrain elements and their electromagnetic characteristics [38]. For instance, the VV value of -5 to -8 dB indicates strong backscatter (dominated by "double-bounce" scattering off vertical walls and concrete or metallic structures), whereas VV value of -19 to -25 dB indicates very low backscatter (where the signal bounces away from the sensor due to the flat surface) [35].

2.5 Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) Data

The Suomi National Polar-orbiting Partnership (SNPP) and NOAA-20 satellites carry the Visible Infrared Imaging Radiometer Suite (VIIRS), which includes a specialized Day/Night Band (DNB) designed for global night-time light (NTL) observation [17]. The VIIRS DNB is designed to produce imagery of clouds continuously for the day, night and terminator scenes. It has a spatial resolution of about 750 m over the entire 3060 km wide swath, which provides global coverage with 12 hours revisit time for cloud imagery under both solar and lunar illumination. GEE provides VIIRS DNB data composited monthly, annually, and daily raw data [39][40] and [41]. In GEE, the monthly average radiance is in the unit of $nW.cm^2.sr^{-1}$ (nanowatts per square centimetre per steradian). The low value (such as < 1) indicates dark areas whereas high value (such as > 15) indicates area with high illumination, such as metropolitan area.

Night time light data has been used as an economic variable. Researchers have found positive correlations across countries between night-time lights data and several socioeconomic variables. The data has positive correlations with GDP, non-agricultural GDP, manufacturing GDP or capital stocks [42]. In the case of Afghanistan, civilian nighttime lights were analysed to understand the new economic reality in the country post-Taliban takeover and revealed economic change [18]. Along with other nighttime lights, time series data generated from VIIRS has also been used to track the long-term development in Southeast Asia [19].

2.6 Sentinel-5P Nitrogen Dioxide Concentrations Data

The Sentinel-5 Precursor (S-5P) mission, launched as part of the European Copernicus program, carries the Tropospheric Monitoring Instrument (TROPOMI), a state-of-the-art spectrometer designed to monitor atmospheric composition. The S-5P mission supports the global monitoring of the atmosphere and air quality using TROPOMI.

TROPOMI provides daily global coverage of various trace gases, most notably Nitrogen Dioxide (NO_2) at a spatial resolution of 3.5×5.5 km [43]. NO_2 Tropospheric Vertical Column Density (NO_2 TVCD) is generated by the Royal Netherlands Meteorological Institute (KNMI) by designing and applying specific algorithms [14] and [15]. On Earth's surface, NO_2 is primarily a byproduct of high-temperature combustion processes, making it suitable for measuring industrial pollution and fossil fuel burning. Because it has a relatively short atmospheric lifetime, NO_2 concentrations remain localized near their sources, allowing researchers to pinpoint emissions from specific infrastructure such as industrial complexes, power plants, and dense transportation networks. The satellite NO_2 column number density is a different physical quantity than the ground concentration. However, it is reported that satellite NO_2 TVCD correlates significantly with surface concentrations [14][15] and [44].

GEE provides offline high-resolution imagery of NO_2 TVCD concentrations as part of its image collections [45]. While NO_2 TVCD concentrations that are less than 0.000015 mol/m^2 represent areas with little fossil fuel combustion such as rural areas, forested regions, and coastal zones values greater than 0.000050 mol/m^2 represent areas of major industrial activity, power plants, or megacity cores with high-density traffic [44]. Other than night time lights, NO_2 concentrations data collected by satellites has also been used to measure economic activities. A global analysis of cyclical economic activity using NO_2 has revealed that this data may be used to compute measures of economic activity in comparison with GDP. The research results in [16] also shows that NO_2 can be used to observe annual economic fluctuations in addition to both lights and all of the most prominent ground level signals available.

2.7 Machine Learning Algorithms

2.7.1 Clustering algorithms for small tabular datasets

Cluster analysis towards a dataset aims to group objects based on their similarity. Objects in one cluster are similar to each other, and less similar to objects in other clusters [33]. Because datasets have different shapes as well as densities, clustering algorithms are generally divided into five main categories, which are centroid based (partitioning), density-based, hierarchical, distribution-based, grid-based. Among partitioning approach algorithms, k-Means is the most popular. Meanwhile, Agglomerative is a well-known hierarchical algorithm. K-Means and Agglomerative clustering are both suitable for small datasets, although each

offer different advantages as excerpted below [33] [46] and [47]:

(1) k-Means is suited for datasets where the number of clusters (k) is known, offering fast convergence with linear complexity. However, it is sensitive to seed placement and outliers, which can distort cluster centres.

(2) Agglomerative hierarchical clustering requires no prior specification of k and produces a dendrogram that reveals the entire relationship structure of the data, allowing the researcher to determine an appropriate cut-off distance for final cluster formation. While its running time complexity is higher, it remains acceptable for small datasets. It supports varied linkage criteria (e.g., Ward's method), which makes it more flexible in capturing varied cluster shapes.

Clusters Evaluation: One method for evaluating the quality of clustering results is by computing Silhouette coefficient [33] and [47]. The coefficient is used to measure how well each data point is clustered with other data points within its assigned cluster, compared to points in other nearby clusters. The coefficient for each point, $s(i)$ falls in the range $-1 \leq s(i) \leq 1$, where $s(i) = 1$ means the point is well-clustered, $s(i) = 0$ means the point lies on or near the boundary between two clusters and $s(i) = -1$ means the point is misclassified and assigned to the wrong cluster. The overall Silhouette coefficient is the average of $s(i)$ of all data points being clustered. If this value is > 0.5 , it usually indicates that there are clusters with good cohesion and separation [33].

2.7.2 Classification algorithms for Sentinel-2 data

Classification is a method for fitting objects to their categories or class labels which best suits each object's characteristics, which are represented by their features. It is a two-step process. First, a model is built from a training dataset. Secondly, the model is evaluated using test dataset [33] and [48]. Classification algorithms have been widely used for land cover classification, specifically for processing Sentinel-2 data. The adoption of Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB) and Decision Tree (DT) classification algorithms for processing Sentinel-2 can be excerpted as [49] and [50]:

- Random Forest (RF) and Support Vector Machine (SVM) consistently outperform Naive Bayes (NB) and Decision Tree (DT). RF is often preferred for Sentinel-2 because it naturally

handles the multi-spectral bands (13 bands) without requiring extensive feature scaling.

- Built-up areas often have similar spectral signatures to bare soil. While NB might struggle to distinguish these, SVM is particularly adept at finding the optimal hyperplane to separate these two closely related classes in a higher-dimensional space.
- NB is rarely used as a primary classifier for modern satellite missions like Sentinel-2 because it cannot capture the complex non-linear relationships between bands that define urban materials (such as concrete, asphalt, and roofing).

Classification Model Evaluation: The trained classification models are usually evaluated using test dataset and four widely known metrics of precision, recall, accuracy, and F-measure (F1) [33] and [48].

2.8 Big Data Analysis Stages

Big data is characterized by 4 V, which are volume (large quantity of generated and stored data), variety (multi-type and multi-source of data), velocity (high speed at which the data are generated and processed), veracity (the accuracy, reliability, and trustworthiness of data) [51] and [52]. Having those characteristics, handling and processing big data raise challenges such as scalability of data processing or computation, dealing with a variety of formats from multiple sources, ensuring that the various data is valid, etc. Compared to general big data, the special thing of geographical big data is spatiotemporal data analysis, which can be roughly categorized into geometrical measurement analysis, human geographical data analysis, and physical geographical analysis [51].

Depending on its characteristics and organizations need, big data can be analysed as data streams (in near real time), batch (periodically), or whenever needed. In general, the last one is performed with the following stages [53] and [54]: (1) Defining what the problems to be solved and the objectives; (2) Identifying, selecting and gathering data, possibly from several sources; (3) Data understanding and exploration; (4) Data preparation, integration and feature engineering, which can be a complex task; (5) Data analysis (modelling and evaluation); (6) Deployment, launching the results.

Stage 2-5 may be performed iteratively to find the best results that meet the objectives before launching. On Stage 2, selecting the data that must be collected (to meet the objectives) is a big issue and must be resolved [51]. On another aspect, selecting the needed parts of big data provided by providers in cloud (such as GEE and Google Map) can be viewed as big data reduction. The results are smaller size of datasets that can be processed with limited computing resources.

2.9 Google API Places

The Google Places API allows developers to access location-based data and "place objects" within Google Maps. As a service that processes HTTP requests, it provides formatted data and imagery for business establishments, geographic coordinates, and prominent points of interest (POIs) [20]. One of its core features is the ability to query objects based on specific spatial criteria, such as a Nearby Search within a defined radius. The POIs can be selected based on their type (store, café, laundry, bank, etc.). For instance, Figure 3 shows some part of Python codes that utilizes the Places API to retrieve all objects within a 1 km radius of a toll gate for several types stored in a list (namely, place_type) by firstly defining its coordinate. The pagination line codes are needed with the following reasons: By default, Google limits 20 results for each request. If there are more than 20 results available, Google generates a unique, temporary string of text and includes it at the root of the JSON response as next_page_token, which acts like a secure digital bookmark. If this token is detected, the next 20 results will be retrieved. The execution returns data of POIs in JSON format.

2.10 Proposed System Architecture

By leveraging the previously discussed data to derive insights into economic development, we propose a system architecture for categorizing toll gates, as shown in Figure 4. The system incorporates four types of input data source: (1) the geographic coordinates of toll gates; (2) Sentinel-2 satellite imagery, data of Sentinel-1 of VV polarization, NOA VIIRS and NO₂ TVCD provided by Sentinel-5P are acquired and pre-processed in the Google Earth Engine (GEE); (3) the locations of transportation hubs, including airports, harbours, obtained from official government websites; and (4) the spatial distribution of business entities (i.e. restaurants, stores, banks) near toll gates, extracted from Google Maps.

```

1 lon = 109.4335838140001
2 lat = -6.91442475299935
3 base_url = "https://maps.googleapis.com/maps/api/place/nearbysearch/json"
4 radius=1000 # 1 km
5 params = {"location": f"{lat},{lon}", "radius": radius,
6           "type": place_type, "key": API_KEY} # place_type contains list of types
7 if next_page_token: #Do pagination
8     params["pagetoken"] = next_page_token
9 response = requests.get(base_url, params=params) # Send API request to get POIs
10 data = response.json() # Store the objects in JSON format

```

Figure 3: Obtaining Google Map objects within radius of 1 km of a toll gate

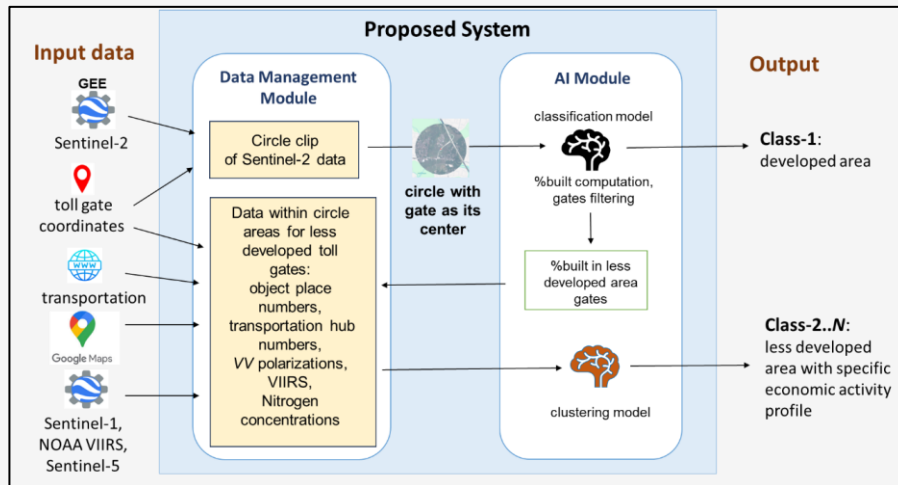


Figure 4: The proposed system architecture

As shown in Figure 4, the retrieval and processing of Sentinel-1, Sentinel-5P, transportation hub, and Google Maps object data are exclusively performed during the clustering stage to categorize toll gates in less developed areas. This approach optimizes resource and computational efficiency. Because Sentinel-2 is widely recognized for its high performance in land-cover classification including urban and rural detection it was used here to identify toll gates in developing areas. Given that the majority of toll gates in Indonesia are located within developing regions, filtering them early significantly reduces the remaining data pool. Consequently, the three additional datasets required for the final stage are smaller in size and can be processed much more efficiently.

The method for processing toll gates is as follows:

Firstly, for every entry-exit toll gate, the imagery data of Sentinel-2 around all toll gates are clipped as circle with radius of 1 km. The selected bands are fed into classification model, which classify each pixel as built-up or non-built-up area. The classification results are used to compute the built-up area percentage. If the built-up percentage in the circle area is greater or equal to 40%, the gate is classified as Class-1, which means that the gate is located in the

developed area, which already support its surrounding economic activities. The use of 1 km as radius is based on: (a) [55] finding, where typical road-effect zones have a 100–1000 m range; (b) research in [56], with the objective of discovering changes in land value zones due to the construction of toll roads, within radius of 1 km from each toll gate. The built-up area percentage threshold of 40% is based on the finding reported in [57]. In the regency areas of West Java, an area has been classified as “developed” when the built-up area is around 40%.

Secondly, the coordinates of gates having built-up percentage less than 40% are used to obtain other data needed for clustering. Within radius of 1 km from each gate: (a) object places (Point of Interests) are collected from Google Map, processed and calculated; (b) transportation hubs are search and count; (c) median value of VV and VH polarization, VIIRS and NO₂ concentration are collected and processed from GEE. The preprocesses data of built-up percentage, number of object places, transportation hubs, value of VV, VIIRS and NO₂ concentration are clustered and used to determine the class of gates located in less developed area.

Both the classification and clustering models must be developed before being integrated into the system. The following is the method for developing the models.

2.11 Classification model development

By adopting the stages defined in [48][53] and [54] as discussed in Subsection 2.8, the classification model is developed through the following steps:

Step-1: Toll gates data collection and cleaning: Toll roads and gates data were obtained from [58]. The gates data is in the GeoJSON FeatureCollection consisting of many features, such as the gate coordinates, names, toll road section names, status, etc. The geometry type for toll gate data is Point. The raw dataset contains 658 rows/records with 38 columns/attributes but with rows duplication. After the duplicates were removed, there are 592 toll gates. By observing the 592 gates, we found that there were gates that connect among toll road sections or located in the middle of toll roads (see Figure 5). As we are processing the entry-exit toll gates only, the non-entry-exit gates were excluded manually. (By observing non versus real exit-entry gates, we conclude that complex algorithms involving image analysis will be needed to distinguish between those two types. In this research, we have not developed the algorithm.) The remaining is 419 entry-exit toll gates spread in 16 provinces (see Table 2). As can be seen, most gates are located in Java Island, which has the longest toll roads and where currently inhabited by approximately 56% of total population [59].

Step-2: Exploring toll gates: After studying the gates on Google Maps, it was found that the environment around the gates varied. There are gates in developed area with large area of buildings or infrastructures, less developed area with less infrastructures and undeveloped area with few infrastructures only. Around toll gates, the built-up area with man-made

infrastructure around toll gates include road or runway (with asphalt as well as cement surfaces), buildings (houses, factories or warehouses) with variety of roof colour (dark and light brown, grey, blue, cream, and light grey or white), parking lots (asphalt or cement), and cement casting. The non-built-up areas are garden or agriculture areas, forest, grasses, bare lands, beach or seaside, river, lake and sea.

Step-3 and Step-4: Dataset preparation and feature selection: The dataset for classification generated from Sentinel-2 bands were prepared on GEE platform manually. Based on the results of Step-2, the class labels are defined as built-up and non-built-up. In GEE, lots of small polygons covering built-up area (road, building, etc) are marked with *built* label, while polygons covering non-built-up area (agriculture areas, forest, bare lands, etc) are labelled with *non-built-up*. As toll gates surroundings across Indonesia vary, the polygons were created around sample of toll gates located in: (a) Semarang, Central of Java; (b) Probolinggo, East of Java; (c) Kahuripan, West of Java; (d) Denpasar, Bali; (e) Bakuheni, Lampung, Sumatera; (f) Medan, North Sumatera; (g) Balikpapan, West Kalimantan; (h) Makassar, South of Sulawesi. Example of polygons on GEE are provided on Figure 6. Figure 7 and Figure 8 show some part of the JavaScript codes used to collect and preprocessing Sentinel-2 data from GEE. The data containing 4 bands and class labels was saved as CSV dataset. There are 30.798 records (points), with 18.853 records labelled as non-built-up and 11.945 as built. Therefore, the records are unbalanced between the two classes. To reduce biased in the model, the records with non-built-up class are randomly down sampled to 12.035. As a result, the prepared dataset has 23.980 records/rows with 12.035 labelled as '0' (*non-built-up*) and 11.945 as '1' (*built*).



Figure 5: Example of non-entry-exit gates on Google Map: (a) Kalihurip Utama 1 connecting Cipularang and Cikampek toll road, (b) Kalikangkung, which is located in KM 414 of Trans Java toll road

Table 2: Distribution of the 419 entry-exit toll gates in 16 provinces of Indonesia

No	Province	Island	Number of gates	No	Province	Island	Number of gates
1	West Java	Jawa	107	9	South Sulawesi	Sulawesi	8
2	Jakarta	Jawa	104	10	South Sumatera	Sulawesi	6
3	East Java	Jawa	61	11	East Kalimantan	Kalimantan	6
4	Banten	Jawa	39	12	Aceh	Sumatra	5
5	Central Java	Jawa	30	13	Bali	Bali	3
6	North	Sumatra	22	14	Bengkulu	Sumatra	2
7	Lampung	Sumatra	16	15	North Sulawesi	Sulawesi	1
8	Riau	Sumatra	8	16	West Sumatera	Sulawesi	1

**Figure 6:** Example of polygons of built-up (red) and non-built-up (green) areas

```

1 #The S2 image with its selection method from GEE ImageCollection
2 var image = ee.ImageCollection('COPERNICUS/S2_SR_HARMONIZED')
3   .filterDate('2025-06-01', '2025-12-30') # filter by this range date
4   .filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 20)) # max cloud: 20%
5   .map(maskS2clouds) # map with the masking function
6   .median() # get the median value
7   .select(['B2', 'B3', 'B4', 'B8']); # select 4 bands only

```

Figure 7: Script for Sentinel-2 image collection and preprocessing

```

1 # Masking cloud for Sentinel-2
2 ~function maskS2clouds(image) {
3   var qa = image.select('QA60'); # get "bitmask" band where each bit represents
4   #cloud, cirrus and other flag ('1': cloud/cirrus,
5   # '0': no cloud/cirrus
6   var cloudBitMask = 1 << 10; # left-shift, take the 10th bit
7   var cirrusBitMask = 1 << 11; # left-shift, take the 11th bit
8   var mask = qa.bitwiseAnd(cloudBitMask).eq(0)
9   .and(qa.bitwiseAnd(cirrusBitMask).eq(0)); # take pixels with no clouds
10  # AND no cirrus
11  return image.updateMask(mask);
12 }

```

Figure 8: Cloud masking function for Sentinel-2 imagery

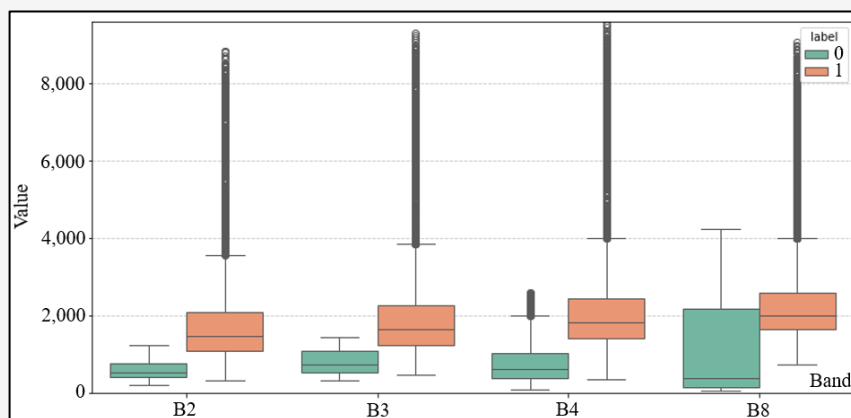


Figure 9: Band distribution for each class, where '0' represents non-built-up and '1' is built-up

Step-5: Dataset Exploration: The prepared dataset with its 23,980 records has the following statistic aggregate: B2 minimum is 207, maximum is 8832, mean is 1163.40; B3 minimum is 308, maximum is 9288, mean is 1339.04; B4 minimum is 95, maximum is 9544, mean is 1389.65; B8 minimum is 40, maximum is 9064, mean is 1640.39. The distribution comparison among the 4 bands is shown on Figure 9. The most B2, B3, B4 values (within the IQR values) are distinctive. Only the values greater than upper whisker (outliers) are intersected. As for the B8 values, some part of the IQR intersects. The minimum until almost half of the IQR of class 1 intersects with the IQR of class 0. Interestingly, while class 0 value has no outlier, the class 1 has many outliers. By learning the comparison between the two bands distribution for each class, it is concluded that a classification model can be built with the potency having good precision, recall, accuracy, F1 scores.

Step-6: Classification Model Fitting and Evaluation: As the model will be used on the GEE platform to predict Sentinel-2 data around toll gates, model fitting is conducted on GEE using JavaScript. The dataset is split into 80% for training and 20% for testing the model. The classification algorithms adopted are Naïve Bayes, Decision Tree, SVM and Smile Random Forest (RF) with 50 trees which are implemented with parallel computation on GEE distributed system. Except the number of trees in RF, RBF kernel type and decision procedure in SVM all of the parameters used are default parameters in GEE, which are:

- Naïve Bayes: no parameter;
- Decision Tree: *maxNodes* (maximum number of nodes) = null (no limit), *minLeafPopulation* (minimum training samples per leaf node) = 1;

- Random Forest: *numberOfTrees* (number of trees) = 50, *variablesPerSplit* (number of random predictors chosen at each split) = null (the default is square root of the number of features), *minLeafPopulation* (minimum training samples per leaf node) = 1, *bagFraction* (proportion of training data used to train each individual tree) = 0.5, *maxNodes* (maximum number of nodes) = null (no limit), *seed* = 0 (when the codes rerun, the random data remains the same);
- SVM: *kernelType* = RBF, *decisionProcedure* = 'Voting', *svmType* = C_SVC (Standard SVM, use C parameter to decide how strict in penalizing training errors), *cost* = 1 (the value of C parameter), *gamma* = null (the default is 1 divided by the number of features), *degree* (polynomial degree) = 3, *terminationEpsilon* = 0.001, *shrinking* = true (use heuristic to temporarily remove/shrink points far away from boundary for faster computation).

The selection of RBF kernel for SVM is based on the findings in the literature that points in bands of Sentinel-2 are not linearly separable. The RBF creates circular or spherical boundaries around data points, which is better in separating the hyperplanes in SVM. RBF is better in handling complex textures and mixed pixels around toll gates compared to linear kernel. The voting decision is selected to ensure a robust decision is made based on the probability of the pixel belonging to one side of the "margin" (of class 0 or 1) versus the other (of class 1 or 0). The execution time for fitting and evaluating (using the specified train/test dataset) the model for each algorithm on GEE was extremely fast (less than 1 second). Thus, this is not used in comparing the model.

Using the model metrics, the results of models' evaluation are provided on Table 3. It can be seen that RF and SVM with RBF kernel models perform the best. This is in line with the results found in [60] and [61]. The RF model is selected and saved as an asset in GEE, such that it can be used later to predict circle area around all entry-exit toll gates.

2.12 Computing percentage of built-up area and filtering class-1 gates method

Using bands of B2, B3, B4 and B8, each pixel of the clipped area is predicted its class label. The saved rf_model saved is used to classify the Sentinel-2 pixel on GEE, with computations performed on cloud server. The prediction results of all pixels are then used to compute the percentage of built-up in the defined circle area by dividing the number of pixels classified as built with total number of pixels. The RF model is used to predict each pixel class with in each circle area, with the centre is the gate coordinate and the radius is 1 km. After all of the pixels' class are predicted, then the percentage of built-up area is computed. Gates having built-up percentages equal or greater than 40% is classified as Class-1. There are 297 toll gates grouped into this Class-1. Note that 6 toll gates were excluded entirely from the classification as the RF model could not produce valid predictions for those locations due to incomplete data. This issue is discussed further in Section 3.1 and Subsection 3.3.3.

2.13 Clustering model development

The gates that are not classified as Class-1, which are 116 gates, are further processed using clustering algorithms. The data needed for this process is the gates' coordinates with their built-up percentage around them. The coordinates of the non-Class-1 gates are used to extract VV and VH backscatter

coefficients from Sentinel-1, tropospheric NO₂ column densities from Sentinel-5P, VIIRS nighttime lights from NOAA, and points of interest (POIs) from Google Maps. Each of the datasets is then explored and cleaned if needed, then merged into one dataset. Features are selected and/or generated such that the dataset will support clustering model with good quality. The dataset is then used to train k-Means and Agglomerative model. Both models are compared and the better one is selected. The clustering model development is presented in Figure 10.

Step-1(a): Collecting Sentinel-1, Sentinel-5, NOAA VIIRS data: The data of VV, VH, VIIRS (monthly average) and NO₂ are collected from GEE using their median during the year of 2025, see Figure 11 for part of python code used.

Step-1(b): Get and count transportation hubs: The Indonesia government has provided data of transportation hubs, which are airports, harbours and train stations, including their coordinates. Figure 12 shows the algorithm for counting the number of each within radius of 1 km of a toll gate. Figure 13 shows the python code to implement line 3 in the algorithm in Figure 12.

Step-1(c): Get object places from Google Maps: Object places (POIs) around a circle area with radius of 1 km were obtained using API provided by Google Map. The types of objects used to collect the data are places that are common around toll gates that represent the existing of small to medium business activities, such are café, stores, lodging, etc. The algorithm to count such objects is shown in Figure 14 and the total time needed to execute the algorithm was 471 seconds (7.85 minutes).

Table 3: The model evaluation results

Algorithm	Accuracy	Class	Precision	Recall	F1-score
Naïve Bayes	0.57	non-built-up	0.63	0.33	0.33
		built-up	0.54	0.80	0.80
Decision Tree	0.98	non-built-up	0.98	0.98	0.98
		built-up	0.98	0.98	0.98
SVM (RBF kernel)	0.99	non-built-up	0.99	0.99	0.99
		built-up	0.99	0.99	0.99
Random Forest (50 trees)	0.99	non-built-up	0.99	0.99	0.99
		built-up	0.99	0.99	0.99

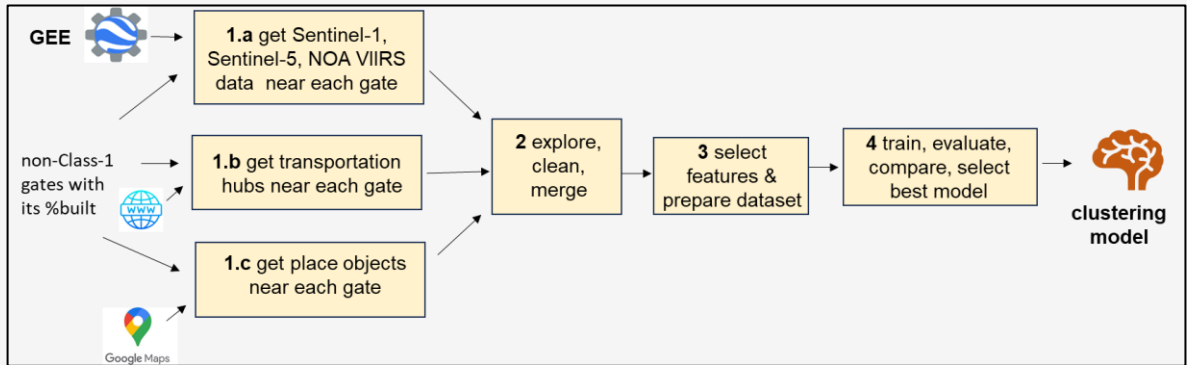


Figure 10: Steps of clustering model development

```

1 #Collect Sentinel-1:
2 var s1 = ee.ImageCollection('COPERNICUS/S1_GRD')
3   .filterDate(startDate, endDate)
4   .filter(ee.Filter.eq('instrumentMode', 'IW'))
5   .filter(ee.Filter.listContains('transmitterReceiverPolarisation', 'VV'))
6   .select(['VV', 'VH']);
7
8 # Collect monthly average of VIIRS:
9 var viirs = ee.ImageCollection("NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG")
10  .filterDate(startDate, endDate)
11  .select('avg_rad');
12
13 # Collect Sentinel-5P NO2 TVCD:
14 var s5p = ee.ImageCollection("COPERNICUS/S5P/OFFL/L3_NO2")
15  .filterDate(startDate, endDate)
16  .select('tropospheric_NO2_column_number_density');
17
18 # Get the median value from the collected data:
19 var s1Med = s1.median();
20 var viirsMed = viirs.median();
21 var s5pMed = s5p.median();
  
```

Figure 11: Script for the collection and median computation of three datasets: Sentinel-1 (VV and VH), VIIRS (average radiance), and Sentinel-5 (NO₂)

Algorithm : Count_transportation_hubs_around_gates
Input : gates of non Class-1, df_non_class1, df_airport, df_harbor, df_station
Output : result_count_transport_hubs.csv
Steps :

- Using Geopandas library: Convert Coordinate Reference System (CRS) of geometry column in df_non Class 1, df_non_class1, df_airport, df_harbor, df_station, which are in degree (EPSG:4326) to meters (EPSG:3857), then store the results in geometry column of new dataframe, gdf_gates, gdf_airport, gdf_harbor, gdf_station
- Create buffer of a circle with 1 km radius for each gate and store the coordinate boundaries as polygon in gdf_gates['buffer_1km']
- Use buffer_1km to find point of airports, harbors, stations within the polygon then count them and store the results in ct_airport, ct_harbor, ct_station column in gdf_gates
- Save gdf_gates into CSV file of result_count_transport_hubs.csv

Figure 12: Algorithm for counting transportation facility objects

```

1 # Function to count transportation object facilities within a circle of a gate
2 def count_in_buffer(gate_gdf, facility_gdf, col_name):
3     # Spatial join: Find transport facility in the buffer, using buffer_1km
4     temp_gdf = gate_gdf.set_geometry('buffer_1km')
5     joined = gpd.sjoin(temp_gdf, facility_gdf, predicate='intersects', how='left')
6     # Count number of objects found per 'objectid'
7     counts = joined.groupby(joined.index['objectid']).count()
8     return counts
  
```

Figure 13: Python function for counting transportation facility objects within a 1 km buffer of each toll gate

```

Algorithm : Collect_object_places_around_gates
Input      : gates of non Class 1, df_gates, object type to be searched, types_to_search,
              radius
Output     : result_count_transport_hubs.csv
Steps      :
1.   for each record in df_gates
2.       for each type in types_to_search
3.           get objects within its circle area created with radius, do pagination if more
              than 20 objects found
4.       merge with other cols in the record append to final_data
5.       save final_data to gates_with_objects.csv

```

Figure 14: Algorithm for collecting object places around gates

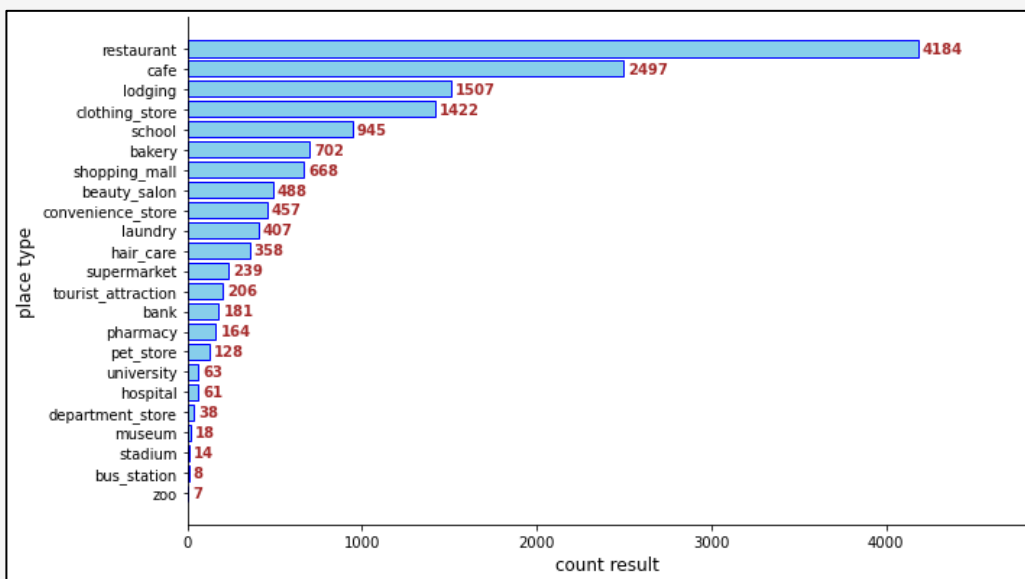


Figure 15: Number of objects for each place around toll gates

Step-2: Explore, Clean and Merge. This step is performed as follows:

(a) Exploring object places data. The results can be excerpted as follows:

- The CSV file has 14,941 rows, which means that as many as 14,941 place objects around 116 toll gates were obtained from Google Map.
- There are 112 unique gate IDs. As there are 116 gates of non-Class-1, this means that there are 4 toll gates with no object places within the specified circle area.
- Among the 14,941 place objects, there are 13,269 unique value of 'place_id'. Thus, there are $14941 - 13269 = 1672$ records with duplicates of place_id. Further investigation revealed that there are 147 objects having 2-3 or more place_type or place_name (with the similar meaning), such as café and restaurant, supermarket and convenience_store.

(b) Cleaning object places data: As each object that will be processed must be unique, the duplicate records identified by pair of id_gate and place_id were then dropped using Python code `df_unique = df_total.drop_duplicates(subset=['id_gate','place_id'])`

As shown on the above code, the results are stored as a new dataframe, df_unique. The cleaned dataset has 14,762 rows. The number of objects for each place type is shown on Figure 15. It can be seen that the top four is restaurant (4,184), café (2,497), lodging (1,507) and clothing store (1,422). Using the cleaned dataset, the number of objects around each toll gate is then calculated. The results are placed in no.of object (number of object) column. The dataframe df_nonClass1_gate is then left join with df_countObj, the value of no.of object column for 4 gates that do not have any place object is 0.

(c) Exploring data of object places number around gates. The minimum value is 0, average is 127.26, median is 112.5, and maximum is 417. The gate with the top objects number (417) is Cibubur 2, located in West Java, near Jakarta, whereas the gate having just 1 place object is Indralaya, in South Sumatera province. The gates with no object around them are Binjai and Kualanamu gates in Deli Serdang regency in North Sumatera province, Menggala gate in Tulangbawang regency in Lampung province, and Pamulutan gate in Ogan Ilir regency in South Sumatera province.

(d) Exploring transportation objects data. The results of the transportation objects revealed that among the 116 gates, no airport and harbour is located within 1 km radius of those gates. This means that all gates near those major transportation hubs are included in Gate 1 class gate. There are only 4 gates that are near transportation (train station) facility, which are gate of Adi Sumarmo, Entrance Cerme, Exit Cerme, and Jatiluhur.

(e) Exploring VV and VH data. The Sentinel-1 VV values around toll gates is characterized with minimum is -20.44, average is -8.90, median is -8.69 and maximum is -6.56. The VH is characterized with min -29.19, mean -15.45, median is -15.06, and maximum is -13.38. To check whether the two values are correlated, the Pearson correlation coefficient between VV and VH is computed. The result is 0.96, which indicates that VV and VH columns have strong positive linear relationship. In the case of selecting feature from a dataset, the two columns are providing almost identical information. Therefore, for clustering purpose, only VV data will be selected.

(f) Exploring VIIRS data. The NOAA VIIRS values around toll gates are characterized with minimum is 0.21, average is 7.51, median is 6.88, and maximum is 20.05. The gate with the minimum VIIRS, which almost dark in the night, is Padang gate in West Sumatera, whereas the brightest gate is Cibubur 1 and 2, near Jakarta city.

(g) Exploring Sentinel-5P tropospheric NO₂ column density data. The NO₂ around toll gates is characterized with minimum is 0.000003, average is 0.000031, median is 0.000028, and maximum is 0.000080. The gate with the least NO₂ concentration is Air Madidi in North Minahasa regency, in North Sulawesi province, while the gate with the most NO₂ concentration is Karanggan Off and On Ramp in Bogor regency, West Java province.

To summarize, the distribution of 5 datasets is given on Figure 16. It can be seen that the distribution of each is not uniform, skew left or right, which indicate that the integration of those datasets is potentially clustered using the specified clustering algorithms.

Step-3: Feature Selection and Dataset Preparation. This step is performed as follows:

(a) Normalization: Based on the exploration and cleaning results, dataset for clustering was prepared with 5 attributes, which are %built, no.of objects, VIIRS, VV, and NO₂. Because the range of attribute values in the dataset is different, normalization is necessary so that no attribute will be dominant for distance computation among records or objects. Min-max normalization using Sklearn library of Python was applied. As the results, the distribution of each attribute is visualized using boxplot and shown on Figure 17. The prepared dataset now has the minimum value is 0 and maximum is 1. There are outliers on attribute VIIRS, VV and NO₂, which indicate that among the 116 toll gates, there exists gates surrounding with extreme condition with regard to night light, buildings and pollution coming from large vehicles or industries.

(b) Selecting the best features: In the first experiment, clustering the prepared dataset using the five attributes with various number of clusters did not deliver any model having Silhouette score greater than 0.5. Therefore, selecting the best features to increase to quality of cluster models is necessary. The algorithm for selecting the best features using k-Means algorithm is presented in Figure 18. The results of the experiment are presented on Table 4, it is found that the best features are no.of objects and VV, with Silhouette score of 0.5716.

Step 4: Train, Evaluate and Compare Clustering Models. Experiments were performed to find the best cluster number for k-Means using the normalized dataset with best features ('no.of objects', 'VV'). The parameters used in experiments are: distance = Euclidean, random_state = 42 (use the same starting random points on every execution), n_init = 10 (run 10 times to find the initial centroid). The evaluated cluster number, k, is 2 to 10. We find that the larger the number of k, the smaller the Silhouette score. For instance, with k = 3, the score is 0.5211, while with k = 10, the score is 0.3768. Therefore, using k-Means, the best cluster number is 2.

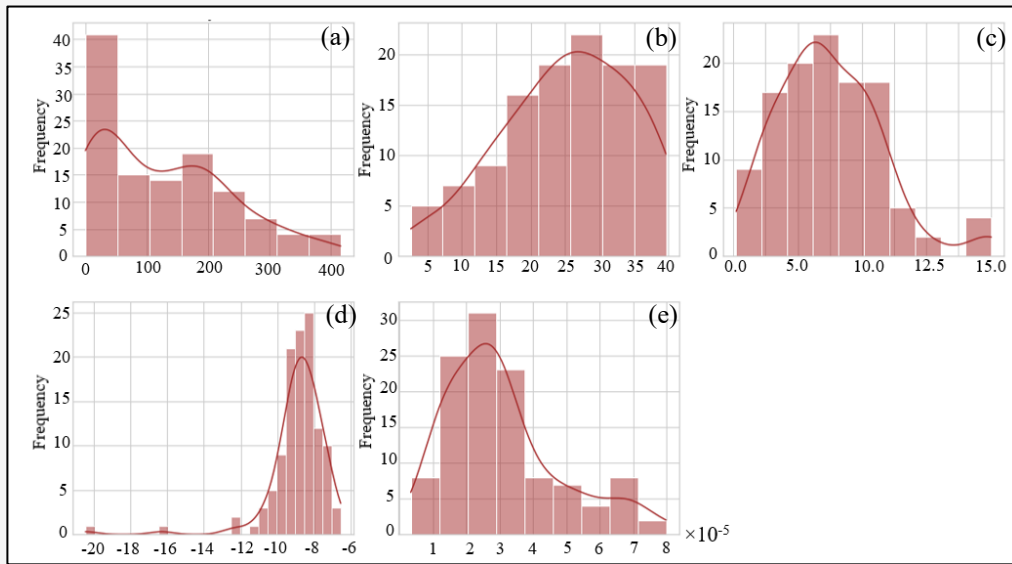


Figure 16: Distribution of the five dataset: (a) objects, (b) %built, (c) VIIRS, (d) VV, and (e) NO₂

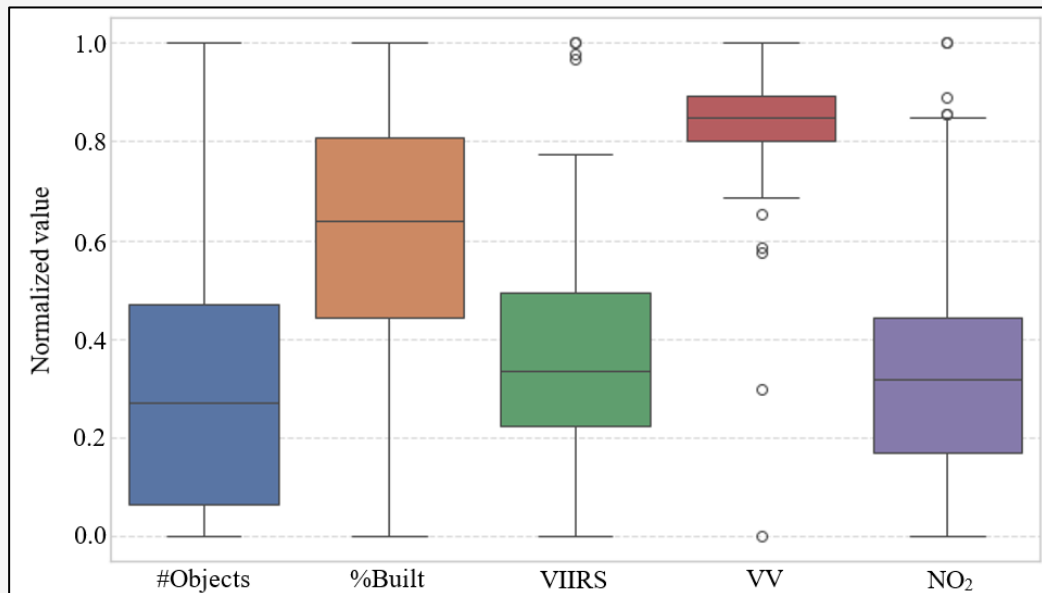


Figure 17: Boxplots of normalized features

Table 4: Results of experiments for finding the best features

Ncls (k)	Best feature - Score	Ncls (k)	Best feature - Score
2	no.of objects, VV - 0.5716	7	VV, NO2 - 0.4800
3	no.of objects, VV - 0.5211	8	VV, NO2 - 0.4809
4	no.of objects, VV - 0.5507	9	VV, NO2 - 0.4669
5	no.of objects, NO2 - 0.5081	10	VV, NO2 - 0.4548
6	no.of objects, NO2 - 0.5048		

```

Algorithm : Finding_best_features
Input      : df_cluster
Output     : {Ncls, best_feature, best_score}
Steps      :
1.  Attributes = columns of df_cluster  ##built, #objects, VIIRS, VV, NO2
2.  for Ncls = 2 to 10  #number of cluster (k)
3.      best_score = -1
4.      best_feature = None
5.      for r = 2 to 5  #size of attribute combinations
6.          for r attributes combination from Attributes
7.              create subset from combination of attributes
8.              train k-Means model using Ncls, subset columns values from df_cluster and
                cluster the subdataset
9.              score = Silhouette score from clustered dataset
10.             if score > best_score
11.                 best_score = score
12.                 best_feature = subset
13.             output Ncls, best_feature, best_score

```

Figure 18: Algorithm for finding best features

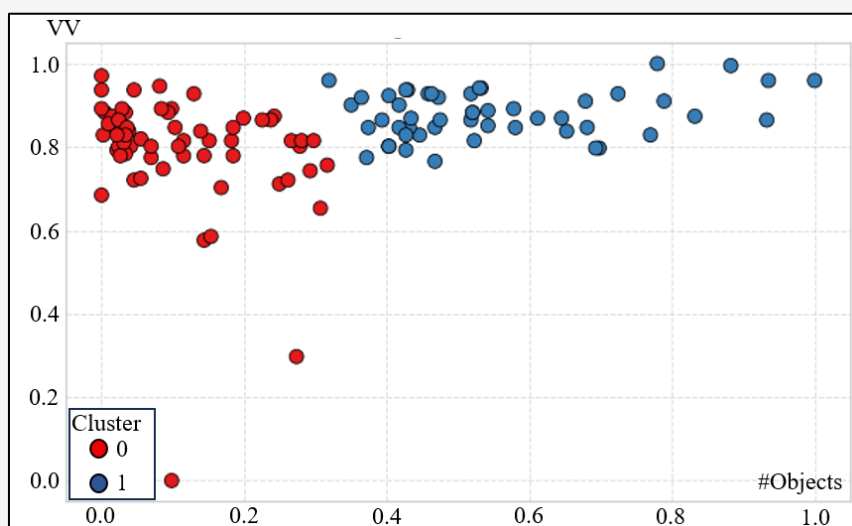


Figure 19: Cluster 0 and 1 points separation

As Agglomerative algorithm is also suitable for small dataset with numeric columns, and there are linkages that can be selected to obtain model with better Silhouette score, the next experiment was performed for clustering with $k = 2$ and ward, average and complete linkage. The distance used is Euclidean. The score results are: ward is 0.5716, average is 0.5743 and complete is 0.4763. As average linkage gives better performance (than k-Means with $k = 2$), the clusters are checked. It turns out, the clusters' member is very unbalance, cluster 0 has 114 members, while cluster 1 has only 2 members. Therefore, for processing this dataset, Agglomerative average linkage is more suitable to find outliers. Then, the clusters as the result of clustering using ward linkage were also investigated. It is found that the cluster members are exactly the same with the members of k-Means (with $k = 2$), which are 51 and 65 gate points. Therefore, the best cluster model is k-

Means model with $k = 2$ or as Agglomerative model with ward linkage with $k = 2$. As the dataset is small (2 columns with 116 rows), both models can be created quickly. Clustering the dataset using the best model produces clusters with the following member and centroid:

- Cluster 1 (51 gates): no.of objects is 0.553910, VV is 0.878665
- Cluster 0 (65 gates): no.of objects is 0.110164, VV is 0.794872

The separation of 116 toll gate points is depicted on Figure 19. It can be seen that points are quite well separated (this is inline with the Silhouette score, which is 0.5716) and there is no point that is miss-clustered.

3. Results and Discussion

3.1 Predicting Classes and Percentage Built-up Area Computation

Processing on GEE which involved predicting class labels using a Random Forest (RF) model for each cloud-free Sentinel-2 pixel across the four selected bands within a defined circular area around the 419 entry-exit gates and subsequently computing the built-up area percentage was highly efficient, taking only 188.11 seconds. Due to efficient execution, this computation is feasible to be launched on the designed system. Among the 419 gates, the built-up area percentage was successfully computed for 413 gates, while the calculation failed for the remaining 6 gates. The reason for this failure was that the RF model could not predict the Sentinel-2 pixels around those specific gates. The suspected cause of the failure is the complete absence of cloud-free pixels within the circular area, likely due to persistent cloud cover in Riau and Aceh during the 2025 data collection period. Consequently, only the 413 successfully processed gates were retained for further analysis with clustering algorithms, while the 6 gates are handled separately.

3.2 Results of Stage-1: Class-1 Toll Gates

There are 297 toll gates having built-up percentage at least 40%, these gates are grouped into Class-1. Among those gates, 211 gates are mostly located in the big cities and 86 gates are in regencies. The cities outside of Java Island are Medan in Sumatera (5 gates), Makassar in Sulawesi (7 gates), and Balikpapan in Kalimantan Island (3 gates). The rest of the cities are in Java (with 195 gates). The regencies are in Kalimantan (1 gate), Sumatera (13 gates) and the rest are in Java (72 gates). Thus, among all 297 gates, 267 gates are located in Java

Island. This is in line with the fact that Java is the most developed island with the longest toll road, where toll ways have been existed since 1980s. It is also found that toll gates intended to support fast access to major airports and harbours included in this class.

The distribution of built-up percentage around circle area is minimum is 40.73%, mean is 69.58% with deviation of 14.65%, median is 69.90%, and maximum is 99.59%. The gates with 99% (or more) of built-up area around them are Kayu Besar 1 (106.73014386700004, -6.129488934999927) and Kayu Besar 2 (106.73023222100004, -6.129935370999931) in West Jakarta city. Example of three gates, Kayu Besar 1 - West Jakarta, Tanjung Mulia - Medan city (72%), Bira Timur - Makassar city (59%) with their circle area created with Google Earth Pro is provided on Figure 20. Toll gates located in cities with a high percentage of built-up area (above 90%) were subsequently examined in Google Earth Pro at a high resolution for greater clarity. As shown in Figure 20(a), there appear to be green spaces around the gate resembling parks or vegetated areas. However, upon closer inspection, most of these green patches turned out to be paved surfaces such as concrete that had become overgrown with moss. Consequently, the model correctly predicted the pixel classes in these areas as built.

By examining the toll gates that fall into this class, where their locations are within or on the outskirts of large cities or districts near large cities, some of which also support access to airports and ports, it can be concluded that the classification model and calculation of the percentage of built-up areas have been successful, and the determination of the percentage limit of 40% is also appropriate.

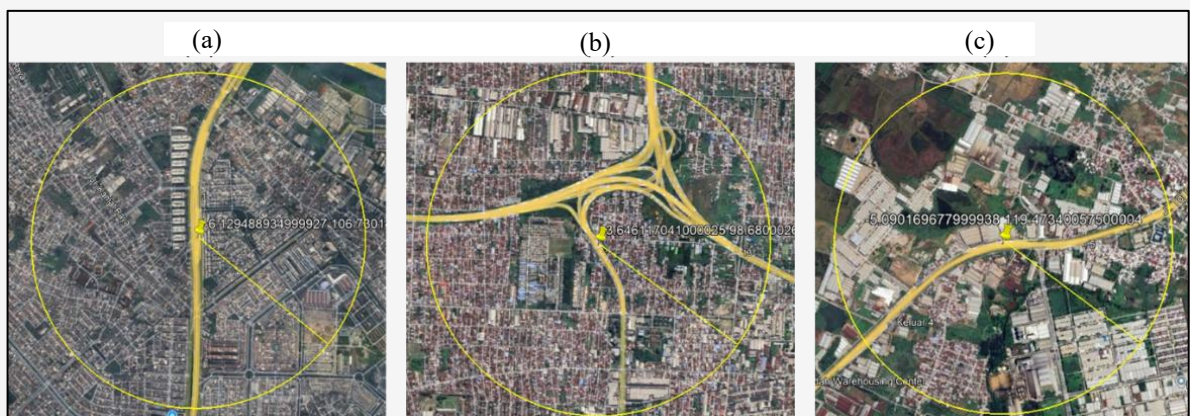


Figure 20: Example of circle area around gate: (a) Kayu Besar 1, (b) Tanjung Mulia, (c) Bira Timur

3.3 Results of Stage-2: Class-2 and Class-3

To interpret the results, the labelled dataset obtained from the clustering process was joined back with the original dataset, which contains the five columns with their original values. Each cluster, namely the Class-2 and Class-3 gates, is then interpreted and analysed in the following sections. Because the clustering dataset relies primarily on the number of object places and VV polarization, the primary interpretation of each cluster profile is derived from these two features. However, since the statistical values of the other attributes (percentage of built-up area, VIIRS, and NO₂) vary distinctly between Class-2 and Class-3, they are also informative. Therefore, the cluster profiles incorporate these additional attributes to provide deeper contextual information regarding the socio-economic and environmental landscape of each group.

3.3.1 Class-2 toll gates

This class comprises 51 gates, with 46 located in regencies and 5 in cities. The specific cities include Depok, Jakarta Timur, and Pasuruan, while the regencies are primarily distributed across Java (44), Sulawesi (1), and Sumatra (1). In Indonesia, regencies generally encompass administrative areas on the outskirts of major cities, developing suburban zones, or rural areas. The detailed profile of these toll gates and their surrounding environments is discussed below. By examining the statistic values depicted on Table 5, it can be interpreted that this cluster gates serve as transit points and act as the “economic engines” of the roads network. The interpretation comes from cluster features as discussed below:

- a. Object places: With a median of 216 and a maximum of 417, these gates are surrounded by intense human or economic activity. The gates are mostly embedded in a rich ecosystem of business entities (eateries, stores, schools, etc.) and possibly have supported the economic development around them.
- b. Sentinel-1 VV: The median backscatter of -8.37 dB is very high (closer to zero). In line with the value of object places number, built-up area and VIIRS, this value confirms the presence of

massive man-made structures around the gates. The very low standard deviation (0.77) shows high level of uniformity and stability in the radar signal. In the context of SAR analysis, this means the scattering properties of the area are not changing much from pixel to pixel or over time, indicating a consistent, homogeneous surface cover [9] and [10].

Additional contextual information:

- a. Built-up area: A median built-up area of 32.03% indicates a “semi-urban” or “peri-urban” environment. While the gates are not like the Class-1 gates (where mostly located near or in city centres), we consider a built-up ratio around 30% might show an active transition to an urban area. The area around these gates may have experienced rapid land-use conversion, where agricultural land is being replaced by industrial estates, small business buildings, or housing complexes triggered by toll road access.
- b. NOAA VIIRS: The median value of 9.58 and mean of 9.96 indicate that most of the gates surrounding are highly illuminated during night time. The residential areas lit up at night, some of the business places may operate 24/7, or there is heavy night-time traffic.
- c. Sentinel-5P NO₂: The median NO₂ value (0.000032) is significantly high, which means that the gates in this cluster are pollution hotspots (as the high NO₂ levels are a direct result of heavy fossil fuel combustion). Given the high VIIRS and object place counts, this is clearly caused by high volumes of heavy-duty vehicles (trucks/logistics) and idling traffic at or around the gates. This finding is inline with the fact previously discussed (in 4 items) which indicate that the gates highly support economic engines activities around them.

Outlier of Class-2 Gates: As shown in Table 5, several gates exhibit extreme minimum and maximum values across certain metrics, indicating that they are outliers within this group. A closer examination of these specific gates revealed several interesting findings, which are discussed below:

Table 5: The profile of class 2 gates

Statistical Measures	No.of objects	VV (dB)	%Built	VIIRS	NO ₂ × 10 ⁻⁶ (mol/m ²)
Mean	230.98	-8.25	30.03	9.97	38
Deviation	70.68	0.78	7.03	3.55	17
Minimum	133.00	-9.81	15.58	3.09	3
25% quartile	178.00	-8.69	24.51	7.44	28
Median	216.00	-8.38	32.03	9.58	32
75% quartile	277.50	-7.60	35.31	11.44	46
Maximum	417.00	-6.56	39.66	20.05	80

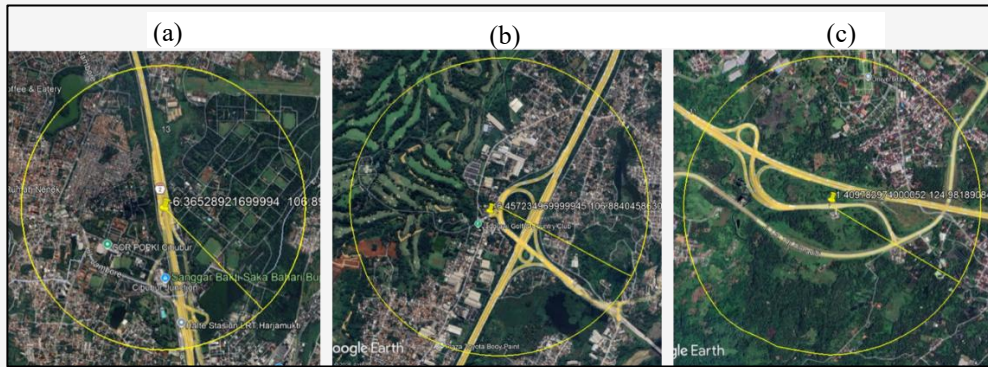


Figure 21: Circle area of gate: (a) Cibubur 2 – West Java; (b) Karangan Off Ramp – West Java; (c) Air Madidi – North Sulawesi

Table 6: The profile of class 3 gates

Statistical Measures	No.of objects	VV	%Built	VIIRS (dB)	NO ₂ × 10 ⁻⁶ (mol/m ²)
Mean	45.94	-9.41	21.50	5.58	25
Deviation	40.52	1.96	9.26	3.33	15
Minimum	0	-20.44	2.61	0.21	9
25% quartile	12	-9.63	14.31	3.27	14
Median	35	-9.12	22.74	5.41	22
75% quartile	76	-8.44	27.67	6.87	29
Maximum	132	-6.97	39.64	19.34	71

Firstly, the gate with object counts of 417 is Cibubur 2 (106.89423404000009, -6.36528921699 994), located in East of Jakarta city (see Figure 21(a)). Its built-up area percentage is 39.09%, VIIRS value is 20.04 (which indicate that the area is highly illuminated during night time), VV value is -7.12 (this value may come from the be combination of urban structure, densely vegetated crops, and rough oil) and median value of NO₂ is 67 × 10⁻⁶ (which is moderately high to high air pollution levels). Closer examination of the gate and its surrounding reveal that these values seem justified, on the west side of the toll way is densely built-up area, indicating highly developed area.

Secondly, there are 2 gates having maximum of NO₂ (80 × 10⁻⁶ mol/m²). Those are Karangan Off Ramp (106.88404586300004, -6.45723496999994 5) and Karangan on Ramp (106.884019642 00006, -6.457186606999926). Both gates are close by and located in Bogor regency, West Java. They have almost 40% of built-up area (39.5 and 39.3), VV value is -7.37, VIIRS value is 10.33, object places counts are 222 and 221. After further examination the gates and their surrounding (see Figure 21(b)), these values seem justified. On the north, south and east side of the toll gates is densely built-up area, indicating highly developed area.

Thirdly, the gate with the minimum value of NO₂ (30 × 10⁻⁶) is Air Madidi (124.98189084000012, 1.409782974000052) located in North Minahasa regency, North Sulawesi. Its built-up percentage is 15.58, no.of objects is 168, VIIRS is 5.93, VV is -7.62. The results of inspecting on the gate and its surrounding (see Figure 21(c)) reveals that these values are justified. There are lots of buildings on the north-west side of the toll gates, while most of the rest is agriculture or bare land, indicating moderately developed area.

3.3.2 Class-3 toll gates

This class comprises 65 gates, with 58 located in regencies and 7 in cities. The specific cities include Denpasar (on Bali), Bengkulu and Prabumulih (on Sumatra), and Makassar (on Sulawesi). The regencies are primarily distributed across Sumatra (31 gates) and Java (26 gates), with one located on Bali. The detailed profile of these toll gates and their surrounding environments is discussed below. As indicated by the statistics in Table 6, this cluster profile characterizes toll gates located in suburban to semi-urban transitional areas. These gates are situated neither in remote rural regions nor in highly dense metropolitan areas, but rather in developing zones or areas characterized by low-to-moderate human and business activities. The following section provides a more detailed profile of each cluster, interpreted within the context of economic activities:

- a. The object places number shows high variance (the deviation is 40.5) around gates. While there are gates that have almost no nearby object (the minimum is 0), the median of 35 suggests that most gates serve as small hubs for local businesses (restaurants, cafes, stores, etc).
- b. The Sentinel-1 VV mean (−9.41 dB) and median (−9.12 dB) are relatively high for non-urban area. As in SAR terms, values closer to −9 dB indicate significant “surface roughness” or “volume scattering” [9] and [10], those VV measures are likely caused by the toll infrastructure itself, highway pavements, and nearby buildings (residential, stores, restaurants/cafes, etc.).
- d. The mean of NO₂ value, 0.000025 mol/m², indicates detectable but relatively low nitrogen dioxide concentrations. This level is typical for areas with consistent vehicles flow (highways) but without the massive industrial emissions [44]. In the context of economic activity, this means no or little industrial activities around these gates.

Outlier of Class-3 Gates: As shown in Table 6, several gates exhibit extreme minimum and maximum values, indicating that they are outliers within this group. A closer examination of these specific gates revealed several interesting findings, which are discussed below.

Additional contextual information:

- a. With a mean built-up area of 21.5% (ranging up to ~40%), these toll gates are located in areas where the majority of the 1 km radius is still undeveloped (likely greenery, agriculture, or open space).
- b. A mean value of 5.58 nighttime light (VIIRS) indicates moderate artificial light emission. It is significantly brighter than rural areas (typically < 1) but much dimmer than a city centre (which often exceeds 20–50) [17] and [19]. This suggests that while the toll gates are illuminated, the surrounding area does not have high-density street lighting or massive commercial neon signage or indicate low activities at night time.
- c. The 75th percentile of VIIRS (6.87) versus the maximum value (19.34) suggests that the vast majority of the gates are in quiet areas (with little light at night), while a small number of gates are approaching “urban” light intensity (that relate to vehicles, residences, business or small industries with lights or electricity at night time).

Firstly, there is a gate with median VV value of −20.44 which indicate that around this gate is dominated by a large, flat, and empty area as a result from the condition that Sentinel-1 radar pulse hits a very smooth surface and bounces away from the satellite [9] and [10]. This gate is Nusa Dua (115.20856036900012, −8.778222245999928) on Bali Island which has 6.1% of built area. As shown on Figure 22(a). Nusa Dua is located above sea where within 1 km of radius, only a small area of land is covered. This indicates that in this circle area, economic activities are low.

Secondly, there is one gate with its circle area with median of VIIRS value equal to 0.21, which is Padang gate (100.331367, −0.793467) (Figure 22(b)). With built-up area percentage of 17.57% and object places count around this gate is 40, there are buildings or vehicles that illuminated with small intensity at nights. Therefore, small size of economic activities exists around this gate.

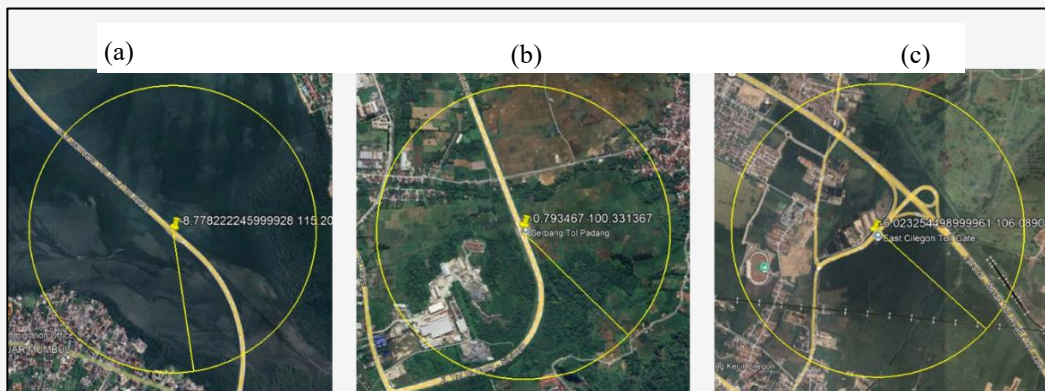


Figure 22: Circle area of gate:

(a) Nusa Dua - Bali; (b) Padang – West Sumatera; (c) East Cilegon– West Java

Thirdly, there is one gate having NO₂ median value of 0.000071, built-up area percentage is 35%, object places number is 63, and VIIRS median value is 11.68 (Figure 22(c)). This high value of NO₂ often occurs in places where there are many vehicles with engine running while they are stationary, or there are many vehicles which are not in motion [44]. This NO₂ value indicates that within this area, there are lots of vehicles that move slowly or traffic jams happen frequently. This gate is Cilegon Timur (106.08904049100003, -6.023254498999961), located in Serang regency, West Java. It is the main exit gate to reach Krakatau Steel, the largest steel factory in Indonesia. Further examination reveals that this gate connects to Jalan Lingkar Selatan (JLS) dan Jalan Raya Anyer, which are often used by large industry vehicles. Thus, this gate supports high economic activities even though it is grouped in this class by k-Means algorithm.

Fourth, there are 4 gates with no place object around them, which are Binjai (98.54002627300008, 3.645315666000045), Kualanamu (98.84483185700005, 3.5887637920000657), Menggala (105.1778883080004, -4.499827602999972), and Pamulutan (104.72240450300006, -3.105694377999953) gate. Binjay and Kualanamu are located in Deli Serdang regency, North Sumatera province. Menggala is in Tulangbawang regency, Lampung province in Sumatra Island. Pamulutan is in Organ Hilir regency, South Sumatra Island. Further observation on the Google Map (as shown in the Figure 23) reveals that the area around the toll gates within a radius of 1 km is mostly agricultural areas, forests or bare land.

The gates are relatively far from residential and business areas. Thus, except for agricultural activities, other kind of economic activities are hardly found.

In summary, this cluster in general is characterized by having circle area with moderate infrastructure development and detectable human activity, but they still retain a significant amount of natural or open land. They are “active” enough to show light and pollution, but they haven't yet reached the saturation of a fully urbanized environment. This means that the gates are located in low to medium economic activities. However, there are 4 outlier gates with no sign of economic activity, and one gate that support high economic activity.

3.3.3 Six gates without built-up area percentage

The six gates for which the built-up area percentage could not be computed are Bathin Solapan, Dumai, Kandis Utara, and Pinggir (parts of the Pekanbaru–Dumai toll road in Riau), as well as Jantho and Seulimeum (parts of the Sigli–Banda Aceh toll road in Aceh). Observations on Google Maps reveal that these gates are located in rural areas, predominantly surrounded by agricultural land, forests, or bare soil. To categorize the gates, data for VIIRS, VV, NO₂, and the number of POIs/objects on Google Maps were compiled separately. The results are presented in Table 7. A comparison between Table 7 and Table 6 (the statistical parameters for Class-3) reveals that all metric values excluding the unavailable built-up area percentage fall within the minimum-to-average range of Class-3. Consequently, these six gates can be appropriately reassigned to Class-3.

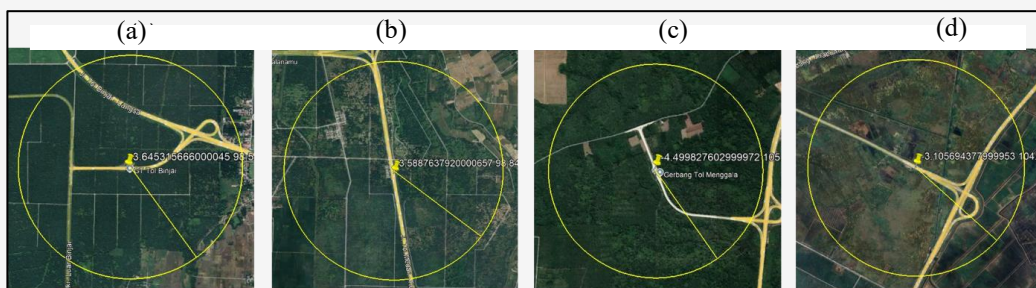


Figure 23: Four toll gates with no place object around them: (a) Binjai, (b) Kualanamu, (c) Menggala and (d) Pamulutan

Table 7: The profile of six gates without built-up percentage area

No	Gate Name	VIIRS	VV (dB)	NO ₂ × 10 ⁻⁵ (mol/m ²)	no.of object
1	Seulimeum	2.495	-8.436	1.098	18
2	Pinggir	3.425	-7.812	1.341	8
3	Kandis Utara	3.190	-7.812	1.263	27
4	Jantho	3.840	-8.312	1.124	2
5	Dumai	4.430	-7.938	1.210	42
6	Bathin Solapan	3.355	-8.124	1.587	26

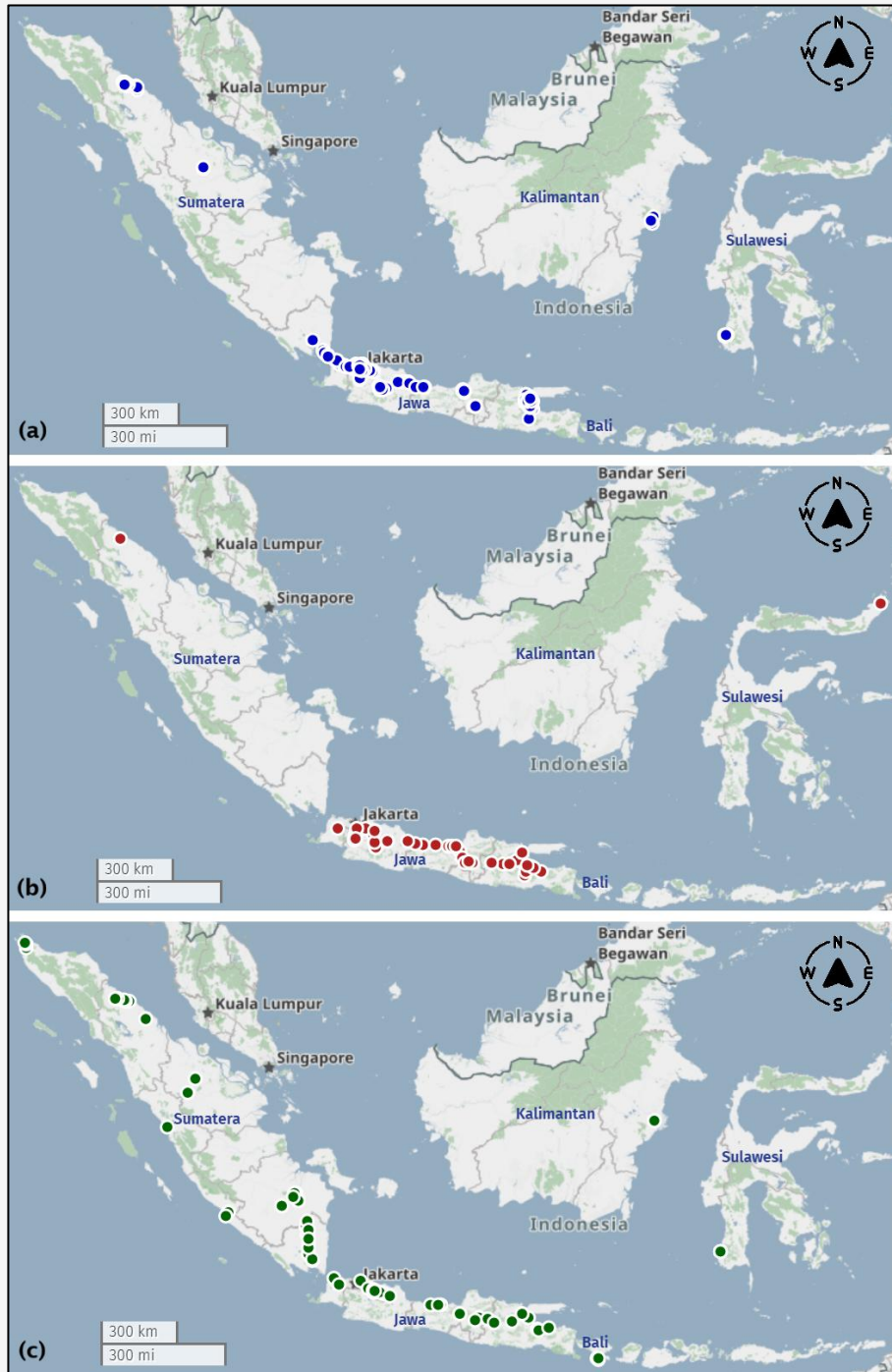


Figure 24: Spatial distribution of 413 toll gates across the five islands under study: Sumatra, Kalimantan, Jawa, Sulawesi, and Bali: (a) Class-1, (b) Class-2, and (c) Class-3. An interactive version of the map is available at <http://u.osmfr.org/m/1441635/> (powered by OpenStreetMap)

3.3.4 Class distribution of all gates

In summary, the distribution of Class-1, Class-2, and Class-3 toll gates across each province in Indonesia having toll roads is presented in Table 8, while their spatial distribution across Indonesia is shown in

Figure 24. Among the Class-1 gates, 266 are located on Jawa Island, which features the country's longest toll road network, while 21 are situated on Sumatra Island. Notably, Class-3 gates are present across all islands.

Table 8: Distribution of the 419 entry-exit toll gates in 16 provinces of Indonesia

No	Island	Province	Number of gates	Class-1	Class-2	Class-3
1	Sumatra	Aceh	5	0	0	5
2		North Sumatera	22	14	1	7
3		West Sumatera	1	0	0	1
4		Riau	8	2	0	6
5		South Sumatera	6	0	0	6
6		Bengkulu	2	0	0	2
7		Lampung	16	5	0	11
8	Java	Banten	39	36	1	2
9		Jakarta	104	102	2	0
10		West Java	107	80	18	9
11		Central Java	30	12	12	6
12		East Java	61	36	16	9
13	Bali	Bali	3	0	0	3
14	Kalimantan	East Kalimantan	6	4	0	2
15	Sulawesi	North Sulawesi	1	0	1	0
16		South Sulawesi	8	6	0	2

4. Conclusion

This research contributes to the fields of remote sensing and transportation infrastructure mapping by developing a novel toll gate classification methodology based on surrounding economic development. This is achieved through the integration of multi-source satellite data specifically Sentinel-1, Sentinel-2, Sentinel-5P, and NOAA-20 with spatial data extracted from Google Maps. The main findings of this study demonstrate that the synergy of these diverse datasets allows the profile of each distinct gate cluster to be comprehensively interpreted, directly linking derived environmental metrics (from the datasets) to localized economic activities. This research successfully demonstrates that Sentinel-2 spectral bands (B2, B3, B4, and B8) are effective in training a Random Forest model to distinguish between built-up and non-built-up pixels within a 1 km radius of toll infrastructure. By quantifying the percentage of built-up areas, the study establishes a primary metric for identifying Class-1 gates, which are located in economically developed regions. Furthermore, clustering algorithms tailored for small datasets specifically k-Means and Agglomerative Hierarchical Clustering prove useful for processing the combined datasets representing toll gates in less-developed regions. This approach generates two easily interpretable clusters, designated as Class-2 and Class-3 gates.

Google Earth Engine (GEE) significantly facilitates the simultaneous ingestion of multi-source geospatial data including Sentinel-1, Sentinel-2, Sentinel-5P, and NOAA-20 satellite systems while streamlining critical preprocessing tasks, such as

cloud masking and temporal median compositing. Particularly for Sentinel-2 data processing, the built-in GEE libraries simplify the training of machine learning classifiers and their application to predict pixel-level land cover classes. Furthermore, GEE enables the efficient quantification of built-up area percentages within a 1 km circular buffer around toll gates. The Google Maps API serves as an effective tool for extracting and compiling localized points of interest (POIs) surrounding the toll gates. Compared to built-up area percentages computation, however, extracting these POIs needed longer time.

Nevertheless, we acknowledge a limitation of our GEE-based methodology: the persistent cloud cover characteristic of tropical regions, which reduce the availability of clean Sentinel-2 imagery, as shown in our experiments. To monitor the impacts of toll roads on local economic growth, the 71 Class-3 toll gates particularly those exhibiting a below-average number of points of interest (POIs) warrant continuous observation, as the surrounding economic activity is indicated to be suboptimal. Notably, this cluster includes four specific gates that completely lack commercial establishments, such as shops and cafes, and remain physically isolated from active economic corridors. The availability of time-series satellite data (as reported in [62]) and POIs are suitable for monitoring the economic growth from time to time. Future research should focus on developing an automated, cloud-based monitoring system to track the dynamic socioeconomic evolution surrounding toll infrastructure, with a particular emphasis on Class-3 gates.

Including the temporal dimension enables continuous tracking on how surrounding area evolves over time. This growth rate information would be valuable for both researchers or policymakers. To improve data collection and analysis, the analytical radius around the toll gates could be expanded beyond 1 km, as spatial observations indicate that certain gates are situated far from active commercial clusters. Integrating new high-resolution datasets, such as real-time human mobility data (e.g. from GPS trajectory data), would significantly enhance the model's fidelity. Furthermore, the proposed monitoring system could incorporate advanced features, including interactive dashboards for real-time statistical visualization and spatial tracking modules designed to identify optimal locations for future toll gate expansions.

Acknowledgements

This study was funded by the *Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi*, Indonesia, for the 2025 fiscal year.

References

- [1] Badan Pengatur Jalan Tol Indonesia (BPJT). (2023). *Laporan Tahunan Badan Pengatur Jalan Tol 2023: peningkatan kualitas konektivitas melalui jalan tol [Annual Report National Toll Road Authority 2023: Improving the Quality of Connectivity via Toll Roads]*. [Online]. Available: <https://pu.go.id/pustaka/biblio/laporan-tahunan-badan-pengatur-jalan-tol-2023-peningkatan-kualitas-konektivitas-melalui-jalan-tol/GJJ1D9>. [Accessed: Feb. 15, 2024].
- [2] Muryani, M. and Amalia, A., (2018). Impact of Road Infrastructure, Education, Health and Foreign Direct Investment Towards Indonesia's Economic Growth: Level of 33 Provinces. *AFEBI Economic and Finance Review*, Vol. 3(2), 35-42. https://repository.unair.ac.id/95639/6/601_Karil_Muryani_Impact-of-Road.pdf.
- [3] Santosa, W., Joewono, T. B., Hadi, P. L. and Wasanta, T., (2025). *Infrastruktur Transportasi sebagai Akselerator Pembangunan Indonesia* (ed). UNPAR PRESS, Bandung.
- [4] Badan Perencanaan Penelitian dan Pengembangan Daerah Kota Surakarta (BPPD Surakarta). (2020). *Penyusunan Kajian Dampak Jalan Tol Terhadap Pertumbuhan Ekonomi Kota Surakarta [Compilation of a Study on the Impact of Toll Roads on the Economic Growth of Surakarta City]*. [Online]. Available: https://ekonomi.surakarta.go.id/assets_frontend/images/file_kajian/bc9fac75b39f190ba923b1449d03263a.pdf. [Accessed: Oct. 10, 2024].
- [5] Juneidi, M. B., (2023). Dampak Pembangunan Jalan Tol Pekanbaru-Dumai Terhadap Sosial Ekonomi Masyarakat di Kecamatan Pinggir Kabupaten Bengkalis [The Impact of the Pekanbaru-Dumai Toll Road Construction on the Socio-Economic Impact of Communities in Pinggir District, Bengkalis Regency]. *Journal of Regional Economics Indonesia*, Vol. 4(2). <https://doi.org/10.26905/jrei.v4i2.11006>.
- [6] The Trans-Java Toll Road Has a Positive Impact on the Community's Economy. *Regional Infrastructure Development Agency of Indonesia*. [Online]. Available: <https://bpiw.pu.go.id/berita/tol-trans-jawa-memberikan-dampak-positif-bagi-ekonomi-masyarakat>. [Accessed: Jan. 26, 2025]
- [7] Master Data Aset Infrastruktur: Gerbang Tol [Infrastructure Asset Master Data: Toll Gates]. *Ministry of Public Works and Public Housing of Indonesia*. [Online]. Available: <https://data.pu.go.id/dataset/master-data-aset-infrastruktur-gerbang-tol>. [Accessed: Aug. 24, 2024].
- [8] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R., (2017). Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sensing of Environment*, Vol. 202, 18-27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- [9] European Space Agency (ESA). (2013). *Sentinel-1 User Handbook*. [Online]. Available: https://ftp.itc.nl/pub/Dragon4_Lecturer_2018/ESA%20EEs%20and%20Sentinels%20Brochures%20pdf/Copernicus%20Sentinels%201,2,3/Sentinel-1%20User%20Handbook.pdf. [Accessed: Sep. 21, 2024].
- [10] Semenzato, A., Pappalardo, S. E., Codato, D., Trivelloni, U., De Zorzi, S., Ferrari, S., De Marchi, M. and Massironi, M., (2020). Mapping and Monitoring Urban Environment Through Sentinel-1 SAR Data: A Case Study in the Veneto Region (Italy). *ISPRS International Journal of Geo-Information*, Vol. 9(6). <https://doi.org/10.3390/ijgi9060375>.
- [11] European Space Agency (ESA). (2015). *Sentinel-2 User Handbook*. [Online]. Available: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook. [Accessed: Oct. 24, 2024].
- [12] Schiller, C., Költzow, J., Schwarz, S., Schiefer, F. and Fassnacht, F. E., (2024). Forest Disturbance Detection in Central Europe Using Transformers and Sentinel-2 Time Series.

- Remote Sensing of Environment*, Vol. 315. <https://doi.org/10.1016/j.rse.2024.114475>.
- [13] Brandt, J., Ertel, J., Spore, J. and Stolle, F., (2023). Wall-to-wall Mapping of Tree Extent in the Tropics with Sentinel-1 and Sentinel-2. *Remote Sensing of Environment*, Vol. 292. <https://doi.org/10.1016/j.rse.2023.113574>.
- [14] European Space Agency. (2021). *Copernicus Sentinel-5P (processed by ESA), TROPOMI Level 2 Nitrogen Dioxide Total Column Products. Version 02*. European Space Agency. [Online]. Available: <https://www.tropomi.eu/data-products/nitrogen-dioxide>. [Accessed: Mar. 21, 2026].
- [15] Royal Netherlands Meteorological Institute. (2022). *Sentinel-5 precursor/TROPOMI Level 2 Product User Manual: Nitrogen dioxide*. [Online]. Available: <https://sentinels.copernicus.eu/documents/247904/2474726/Sentinel-5P-Level-2-Product-User-Manual-Nitrogen-Dioxide.pdf> [Accessed: May 3, 2026].
- [16] Ezran, I., Morris, S. D., Rama, M. and Riera-Crichton, D., (2023). Measuring Global Economic Activity Using Air pollution. *Policy Research Working Paper 10445, The World Bank*. [Online]. Available: <https://openknowledge.worldbank.org/server/api/core/bitstreams/d4f3bb17-905e-44c2-b7c215794db39cca/content>. [Accessed: Apr 10, 2026].
- [17] National Oceanic and Atmospheric Administration (NOAA). (2013). Visible Infrared Imaging Radiometer Suite (VIIRS) Sensor Data Record (SDR) User's Guide). *NOAA Technical Report NESDIS 142A*. U.S. Department of Commerce. [Online]. Available: <https://nsidc.org/sites/default/files/viirs-sdr-users-guide>. [Accessed: Apr 12, 2026].
- [18] Cabanillas, O. E. B., Kosmidou-Bradley, W., Redaelli, S., Tateishi, E. and Teruggi, I., (2024). Afghanistan's New Economic Landscape: Using Nighttime Lights to Understand the Civilian Economy after 2021. *Policy Research Working Paper Series 10969, The World Bank*. [Online]. Available: <https://documents1.worldbank.org/curated/en/099256511062426629/pdf/IDU-f7011ce8-5ecb-494d-a0e6-042efed1a409.pdf>. [Accessed: Apr 18, 2026].
- [19] Zhao, M., Zhou, Y., Li, X., Zhou, C., Cheng, W., Li, M. and Huang, K., (2020). Building a Series of Consistent Night-Time Light Data (1992–2018) in Southeast Asia by Integrating DMSP-OLS and NPP-VIIRS. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 58(3), 1843–1856. <https://doi.org/10.1109/TGRS.2019.2949797>.
- [20] Places-API, Google Maps Platform. *Google*. [Online]. Available: <https://developers.google.com/maps/documentation/places/web-service/overview>. [Accessed: Jan. 26, 2025].
- [21] Quiroz Villanueva, M. J., Guglielmi, F., De Fabiis, F. and Coppola, P., (2025). Social and Economic Impacts of Transportation Multi-Modal and Multi-Service Hub Development. *Sustainability*, Vol. 17(4). <https://doi.org/10.3390/su17041767>.
- [22] Wibowo, T. A., Hamzah, M. Z. and Sofilda, E., (2024). Economic Impacts of Port Development on Coastal Communities: A Vocational Education Perspective. *International Journal of Economics, Management and Accounting*, Vol. 1(3), 128–138. <https://doi.org/10.61132/ijema.v1i3.159>.
- [23] Naaz, S., Parveen, S. and Ahmed, J., (2021). An Artificial Intelligence Based Toll Collection System. *The 2nd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD)*, February 2021, New Delhi, India. <https://doi.org/10.4108/eai.27-2-2020.2303171>.
- [24] Dorcas, S., (2022). Smart Tolling through AI: Revolutionizing Payment Systems for Modern Transportation Networks. *Iconic Research and Engineering Journals*, Vol. 5(8). <https://www.irejournals.com/paper-details/1703187>.
- [25] Manojkumar, P., Raja Sekaran, N., Dhanapal, M. and Srinivasan, C., (2021). Virtual Tollgate using Machine Learning. *IOP Conference Series: Materials Science and Engineering*, Vol. 1084(1). <https://doi.org/10.1088/1757-899X/1084/1/012040>.
- [26] Zahedian, S., Nohekhan, A. and Sadabadi, K. F., (2021). Dynamic Toll Prediction Using Historical Data on Toll Roads: Case Study of the I-66 Inner Beltway. *Transportation Engineering*, Vol. 5. <https://doi.org/10.1016/j.treng.2021.100084>.
- [27] Hermawan, I., (2023). Technology of Acceptance Systems of Toll Roads Payment: Comparison of E-Toll Payment System and MLFF Technology of Trans Sumatera Toll Road. *International Journal of Engineering Business and Social Science*, Vol. 1(05), 439–466. <https://doi.org/10.58451/ijebss.v1i05.81>.
- [28] Earth Engine Data Catalog. *Google Earth Engine Developers*. [Online]. Available: <https://developers.google.com/earth-engine/datasets> [Accessed: Mar. 24, 2026].
- [29] Fu, Y., Zhu, Z., Liu, L., Zhan, W., He, T., Shen, H., Zhao, J., Liu, Y., Zhang, H., Liu, Z., Xue, Y. and Ao, Z., (2024). Remote Sensing Time Series Analysis: A Review of Data and

- Applications. *Journal of Remote Sensing*, Vol. 4. <https://doi.org/10.34133/remotesensing.0285>
- [30] Pertiwi, A. P., Lee, C. B. and Traganos, D., (2021). Cloud-Native Coastal Turbid Zone Detection Using Multi-Temporal Sentinel-2 Data on Google Earth Engine. *Frontiers in Marine Science*, Vol. 8. <https://doi.org/10.3389/fmars.2021.699055>.
- [31] Traganos, D., Pertiwi, A. P., Lee, C. B., Blume, A., Poursanidis, D. and Shapiro, A., (2022). Earth Observation for Ecosystem Accounting: Spatially Explicit National Seagrass Extent and Carbon Stock in Kenya, Tanzania, Mozambique and Madagascar. *Remote Sensing in Ecology and Conservation*, Vol. 8(6), 778-792. <https://doi.org/10.1002/rse2.287>.
- [32] Sentinel 2 Bands and Combinations. *GISGeography*. [Online]. Available: <https://gisgeography.com/sentinel-2-bands-combinations> [Accessed: 22 Jan 2025].
- [33] Han, J., Kamber, M. and Pei, J., (2012). *Data Mining: Concepts and Techniques (3rd ed.)*. Morgan Kaufmann. San Francisco.
- [34] Dimyati, M., Husna, A., Handayani, P. T. and Annisa, D. N., (2022). Cloud Removal on Satellite Imagery Using Blended Model: Case Study Using Quick Look of High-Resolution Image of Indonesia. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, Vol. 20(2), 373-382. <https://doi.org/10.12928/telkomnika.v20i2.21085>.
- [35] Chuengpituk-u-dom, P., Tripathi, N., and Thammaboribal, P. (2025). A Comparative Study of Flood Monitoring Techniques Using the UN-SPIDER Recommendations and a Generative AI-Based Model (SATGPT): A Case Study in Ayutthaya, Thailand. *International Journal of Geoinformatics*, Vol. 21(10); 116-128. <https://doi.org/10.52939/ijg.v21i10.4535>.
- [36] Ashilah, H. and Lasminto, U., (2025). Enhancing Flood Detection in Surabaya: A Comparative Study of VV and VH Polarizations with Sentinel-1 Data. *Journal of Civil Engineering*, Vol. 40(1), 89-89. <https://journal.its.ac.id/index.php/jce/article/view/7272/6268>.
- [37] De Luca, G., MN Silva, J., Di Fazio, S. and Modica, G., (2022). Integrated use of Sentinel-1 and Sentinel-2 Data and Open-Source Machine Learning Algorithms for Land Cover Mapping in a Mediterranean Region. *European Journal of Remote Sensing*, Vol. 55(1), 52-70. <https://10.1080/22797254.2021.2018667>.
- [38] Sentinel-1 Algorithms, Google Earth Engine. *Google*. [Online]. Available: <https://developers.google.com/earth-engine/guides/sentinel1>. [Accessed: May 3, 2026]
- [39] VIIRS Nighttime Day/Night Band Composites Version 1, Earth Engine Data Catalog. *Google*. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_MONTHLY_V1_VCMCFG [Accessed: May 3, 2026]
- [40] VIIRS Nighttime Day/Night Annual Band Composites V2.2, Earth Engine Data Catalog. *Google*. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_DNB_ANNUAL_V22. [Accessed: May 3, 2026]
- [41] VNP46A1: VIIRS Daily Gridded Day Night Band 500m Linear Lat Lon Grid Night, Earth Engine Data Catalog. *Google*. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/NOAA_VIIRS_001_VNP46A1 [Accessed: May 3, 2026]
- [42] Addison, D. and Stewart, B., (2015) Nighttime Lights Revisited: The Use of Nighttime Lights Data as a Proxy for Economic Variables. *Policy Research Working Paper 7496, The World Bank*. [Online]. Available: <https://openknowledge.worldbank.org/server/api/core/bitstreams/c2ee7772-906b-590d-b8c2-c3626439f4ef/content>. [Accessed: Apr. 18, 2026].
- [43] Nitrogen Dioxide (L2_NO2_). *TROPOMI*. [Online]. Available: <https://www.tropomi.eu/data-products/nitrogen-dioxide>. [Accessed: May 3, 2026].
- [44] Grzybowski, P. T., Markowicz, K. M. and Musiał, J. P., (2023). Estimations of the Ground-Level NO₂ Concentrations Based on the Sentinel-5P NO₂ Tropospheric Column Number Density Product. *Remote Sensing*, Vol. 15(2). <https://doi.org/10.3390/rs15020378>
- [45] Sentinel-5P OFFL NO₂: Offline Nitrogen Dioxide, Earth Engine Data Catalog. *Google*. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2. [Accessed: May 7, 2026]
- [46] Xu, D. and Tian, Y., (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, Vol. 2(2), 165-193. <https://doi.org/10.1007/s40745-015-0040-1>.
- [47] Kumar, J. M., Nanda, R., Rath, R. K. and Rao, G. T., (2020). Image Segmentation using K-means Clustering. *International Journal of Advanced Science and Technology*, Vol. 29(6s): 3700 – 3704.

- [48] Moertini, V. S., Septrianto, M. F. and Venica, L., (2019). Incremental Parallel Classifier for Big Data with Case Study: Naïve Bayes Using Mapreduce Patterns. *Journal of Theoretical and Applied Information Technology*, Vol. 97(11): 3077- 3097.
- [49] Aji, A., Husna, V., and Purnama, S. (2024). Multi-Temporal Data for Land Use Change Analysis Using a Machine Learning Approach (Google Earth Engine). *International Journal of Geoinformatics*, Vol. 20(4); 19–28. <https://doi.org/10.52939/ijg.v20i4.3145>.
- [50] Thanh Noi, P. and Kappas, M., (2018). Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*, Vol. 18(1). <https://doi.org/10.3390/s18010018>.
- [51] Shu, H., (2016). Big Data Analytics: Six Techniques. *Geo-spatial Information Science*, Vol. 19(2), 119-128. <https://doi.org/10.1080/10095020.2016.1182307>.
- [52] Domingue, J., Lasierra, N., Fensel, A., van Kasteren, T., Strohhach, M. and Thalhammer, A., (2016). Big Data Analysis. In *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, 63-86. https://10.0.3.239/978-3-319-21569-3_5
- [53] Marr, B., (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. John Wiley & Sons. United Kingdom.
- [54] Provost, F. and Fawcett, T., (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media Inc, USA.
- [55] Forman, R. T. and Deblinger, R. D., (2000). The Ecological Road - Effect Zone of a Massachusetts (USA) Suburban Highway. *Conservation Biology*, Vol. 14(1), 36-46.
- [56] Nathania, J., Subiyanto, S. and Suprayogi, A., (2017). Analisis Perubahan Lahan dan Zona Nilai Tanah di Kecamatan Ungaran Timur Akibat Pembangunan Jalan Tol Semarang-Solo (Tahun 2008–2017) [Analysis of Land Changes and Land Value Zones in East Ungaran District Due to the Construction of the Semarang-Solo Toll Road (2008–2017)]. *Jurnal Geodesi Undip*, Vol. 6(4), 433-442.
- [57] Salim, W. and Faoziyah, U., (2022). The Effect of Transport Infrastructure on Land-Use Change: The Case of Toll Road and High-Speed Railway Development in West Java. *Journal of Regional and City Planning*, Vol. 33(1), 48-65. <https://doi.org/10.5614/jpwk.2022.33.1.3>.
- [58] Open Data PU. *Ministry of Public Works and Public Housing of Indonesia*. [Online]. Available: <https://data.pu.go.id/>. [Accessed Aug. 12, 2024].
- [59] Aditya, R. B. and Husna, Z., (2022). Identification of Sprawl Development Typologies around Toll Road Gates in Java, Indonesia. *TATALOKA*. Vol. 24(1), 1-14. <https://doi.org/10.14710/tataloka.24.1.1-14>.
- [60] Balqis, H., Hidayati, I., and Susilo, B. (2026). Comparison of Built-Up Expansion Prediction Models for Magelang City and Its Peri Urban Areas: CA-Logistic Regression, CA-SVM, CA-MLP, and CA-RF. *International Journal of Geoinformatics*, Vol. 22(5), 89–112. <https://doi.org/10.52939/ijg.v22i5.4982>.
- [61] Ghayour, L., Neshat, A., Paryani, S., Shahabi, H., Shirzadi, A., Chen, W., Al-Ansari, N., Geertsema, M., Pourmehdi Amiri, M., Gholamnia, M., Dou, J. and Ahmad, A. (2021). Performance Evaluation of Sentinel-2 and Landsat 8 OLI Data for Land Cover/Use Classification Using a Comparison between Machine Learning Algorithms. *Remote Sensing*, Vol. 13(7). <https://doi.org/10.3390/rs13071349>.
- [62] Jayeola, T. and Orlíková, L., (2025). Analyzing Land Cover Changes (2013-2024) using Remote Sensing: A Case Study of Lagos City, Nigeria. *International Journal of Geoinformatics*, Vol. 21(5). <https://doi.org/10.52939/ijg.v21i5.4167>.