11 (2) 2025 75-90 ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
Register: *Jurnal Ilmiah Teknologi Sistem Informasi*

Contents lists available at www.journal.unipdu.ac.id

Register

Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

Spatial Semantic Analysis and Origin-Destination Prediction Based on Extensive GPS Trajectory in Jakarta

Humasak Tommy Argo Simanjuntak [a,*], *Agnes Hutauruk* [b], *Haryati Situmorang* [c], *Yoshua Silitonga* [d]

[a,b,c,d] *Department of Information Systems, Institut Teknologi Del, Jl. Sisingamangaraja Sitoluama-Laguboti, Toba Samosir 22381, Indonesia*
email: [a,*]humasak@del.ac.id, [b]agnesabigael28@gmail.com, [c]rosalinasitumorang291@gmail.com, [d]masiyoshua25@gmail.com
** Correspondence*

ARTICLE INFO

Article history:
Received 7 September 2021
Revised 21 October 2021
Accepted 31 December 2021
Available online 07 April 2022

Keywords:
Annotates GPS Trajectory
Spatial Semantic Analysis
Spatial Temporal Clustering
Origin Destination Prediction
Spatial Temporal Regression

Please cite this article in IEEE style as:
H. T. A. Simanjunyak, A. Hutauruk, H. Situmorang, and Y. Silitonga, "Spatial Semantic Analysis and Origin-Destination Prediction Based on Extensive GPS Trajectory in Jakarta," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 11, no. 2, pp. 75-90, 2025.

ABSTRACT

The rapid growth of mobility data from GPS trajectories offers unprecedented opportunities to gain deep insights into human mobility behavior, with significant implications for urban planning, traffic management, public transportation optimization, emergency response, and smart city development. However, a key challenge lies in transforming raw GPS trajectory data, consisting of sequences of coordinates and timestamps, into meaningful, context-rich information that can support analysis and decision making. This study proposes a semi-supervised framework to enhance the contextual and semantic understanding of journeys, using Grab Jakarta GPS trajectory data as a case study. The framework involves extracting origin-destination pairs, augmenting the data with temporal (day, time) and spatial (postal code, land use) contexts through public datasets, assigning cluster labels to characterize groups of journeys, analyzing mobility patterns, and ultimately predicting trip destinations. Origin-destination clustering, performed using the DBSCAN algorithm, identified five meaningful clusters, achieving the highest silhouette score of 0.56 with epsilon = 7.0 and min_samples = 5. Subsequently, a regression-based prediction model was developed, employing nine algorithms, including three deep learning approaches. The LSTM model demonstrated the best performance, yielding a mean squared error of 0.0053 and a coefficient of determination (R^2) of 86.20% in predicting trip destinations. These findings highlight the potential of integrating spatial-temporal enrichment and machine learning to derive actionable insights from GPS trajectory data.

1. Introduction

Many technologies and mobile applications have been developed to monitor the movement of humans, vehicles, animals, and even natural phenomena. These applications utilize location-based services (LBS) to gather geodata, data collected in real-time using one or more location-tracking technologies. In transportation, location-based services (LBS) have also become important, especially for online taxis that use navigation systems in motorbikes or cars. Online taxi services such as Uber, Gojek, Maxim and Grab have become the primary means of transportation for many urban residents. These services employ applications utilizing GPS (Global Positioning System) technology to serve taxi calls (taxi on call) and track the trajectories. The use of this technology produces massive spatiotemporal data (latitude, longitude, time), depending on the number of taxi trips and the GPS sampling rate of the technology used [1].

Analyzing spatiotemporal data, primarily generated from online taxis, is highly useful for understanding customer travel behavior and preferences (people's decision-making processes during travel regarding the choice of travel mode, route, departure time, and destination). It also helps taxi providers identify areas with high concentrations of potential passengers, such as commuters travelling

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

to and from work or visiting popular attractions. These large-scale datasets are mined to recognize patterns, gain insights into lifestyles (daily routines and habits), and build user profiles, providing an understanding of where people live, work, and travel. Furthermore, from a global view or population level, this understanding is valuable for analyzing the functional regions of a city, predicting the density levels of specific areas, urban planning, traffic management, public transport optimization, emergency and disaster management, and even smart city development.

Learning and understanding spatiotemporal data generated from the mobility of online taxis is challenging. The challenges relate to the raw data's accuracy, sparsity, and interpretability (semantic representation). Spatiotemporal data, represented as a combination of location and time (denoted as {<lat1, lon1, t1>, <lat2, lon2, t2>, …, <latN, lonN, tN>}, is complex to understand without a strong semantic representation. Without sufficient context, the data does not provide a "narrative" that explains what, why, and how a trip was undertaken. In other words, it is difficult to understand the trip's context and reason. Therefore, rich spatial and temporal context is needed to gain a more complete understanding of the meaning behind human journeys.

Research on spatiotemporal data analysis still primarily focuses on moving object databases and statistical analysis, mainly focusing on trajectory modelling [2], [3], [4], trajectory management [5], [6], and trajectory mining [7], [8], [9], [10], [11], [12], [13], [14], [15], as well as, more recently, trajectory generation using generative AI [16], [17], [18]. Related research also still focuses on raw trajectories consisting of spatiotemporal records, often overlooking contextual information such as land use and geographic objects, which can significantly contribute to the semantic knowledge of mobility. As a result, obtaining a holistic interpretation of movement or mobility that includes contextual data remain a challenge.

Semantic trajectories help fill the gap in research related to trajectory data by enriching spatiotemporal points with geographic and external contextual data. In the preprocessing step, the trajectory is enriched by snapping each point to the street utilizing road network data through a map-matching algorithm [19], [20], [21], [22]. Several studies have also carried out semantic analysis on GPS trajectory data by providing annotations based on trajectory segments to provide context. Automatic semantic segmentation of trajectories has been proposed employing ontology-based knowledge, collections of Linked Open Data (LOD), and POI annotations for each trajectory's start and endpoint [23], [24]. At the same time, another research has extracted movement episodes (stop and move), where a stop is detected as a stay point (a place meaningful to users because it indicates where users conduct their activities) from tracking trajectory data to identify meaningful user locations and build semantic episodes based on them [25], [26], [27]. The meaning of a stay point is further enriched by adding contextual data, such as the Points of Interest (POI). However, those previous studies had limited features in the data and used only small datasets to enrich trajectories, leading to partial understanding of mobility based on low-dimensional data analysis. We must enrich large-scale trajectories with rich data from collective, ubiquitous, and social sensing to gain a complete view of mobility. Another challenge is that most studies employ supervised methods, which require labelled data. It is therefore crucial to develop a largely semi-supervised model, considering the massive amount of unlabeled mobility data generated today, especially within the Indonesia dataset. Our research aims to bridge these gaps by presenting a new framework for understanding travel behavior through spatial semantic analysis.

Furthermore, research employing machine learning methods to discover mobility patterns (significantly in predicting destination) at the city or individual level based on specific regularities has also been conducted [28], [29], [30]. At the city level, the WhereNext method was proposed to predict the next location of moving objects using T-pattern decision trees [31]. A deep-learning-based approach, ST-ResNet, has also been proposed to collectively forecast crowd in-flow and outflow in every region of a city [32]. Specifically for taxi data, deep learning has been used to learn patterns in destination prediction by introducing efficient data embedding for time-related features [33], and a multi-view Deep Learning Framework (MDLF) to exploit crowd travel preferences and regional location contexts [34]. At the individual level, an advanced method based on a probabilistic model (Hidden Markov Model) has learned user regularities and automatically predict the next location a user visits [35]. Some studies have also determined the trip purpose by learning from user trajectories and individual

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

characteristics, adding POI or land use type into the data [36], [37]. All of these research results show the importance of spatial and temporal context in improving trip understanding and increasing the accuracy of destination prediction. Therefore, it is essential to enrich trajectories collectively and use various data sources to better understand overall mobility data.

This study investigated the effects of a semi-supervised framework that annotates semantic context to GPS trajectory data in predicting travel destinations. While earlier studies have explored the impact of spatial and temporal context on improving the understanding of trips, they included limited features, relied on small datasets, and have not explicitly addressed its influence on predicting travel destinations. Therefore, to answer these challenges, this paper presents a valuable contribution to the field of semantic analysis and destination prediction of Grab GPS Trajectory data by delivering the following contributions. First, proposing a semi-supervised framework that annotates semantic context to GPS trajectory data. The framework extracts Origin–Destination GPS trajectory data, derives temporal attributes, enriches postal code and land use context by utilizing spatial and external data, and provides cluster context to the data, thus providing a more meaningful semantic context. Second, the proposed framework is developed using Python, and an extensive experiment was conducted on Grab-Posisi data (the high-resolution GPS data trajectory in Jakarta, Indonesia) [35] to evaluate its effectiveness. Our study is the first to demonstrate a semi-supervised framework for GPS Trajectory in Indonesia. Third, building a regression model using a deep learning architecture (Recurrent Neural Network) based on GPS trajectory data equipped with semantic context, and comparing the prediction results with those obtained using general methods.

The remainder of this paper is organized as follows: Section 2 provides a detailed discussion of the data and the proposed semi-supervised method, while Section 3 presents the results and discussions based on experiments using the Grab-Posisi trajectory dataset. Finally, Section 4 provides some concluding remarks and discusses future work from this research.

2. Materials and Methods

2.1. Data Description

This research uses a collection of Grab taxi trajectory GPS datasets for the Indonesian area. This dataset is sample data obtained from the trajectories of Gab taxi drivers in Jakarta [19], [38]. Data was collected over two weeks, from 08 April 2019 to 21 April 2019, with 6000 trajectories collected daily. The data comes from both Android and iOS devices, and the quality of each trajectory varies. Each category includes 1000 trajectories daily, resulting in a total of 56000 trajectories over two-weeks, comprising 14000 trajectories for each mode (car, motorcycle) and device type (iOS, android). Although the dataset used in this study was collected in 2019, it remains highly valuable for analysing fundamental human mobility patterns. The choice of this dataset, drawn from the widely used Grab-posisi system at the time, is primarily driven by its availability, completeness, and high spatio-temporal granularity, factors critical for developing and validating the proposed framework. Furthermore, using this dataset provides a robust benchmark that aligns seamlessly with many existing urban mobility studies relying on comparable big-city trajectory data. It also offers a clean baseline, enabling the research to focus on methodological advancements in trajectory enrichment, clustering, and prediction.

Each GPS point represents the value of trajectory_ID, latitude, longitude, timestamp (UTC), accuracy level, bearing, and speed. The sampling rate is 1 second, the highest among existing GPS trajectory datasets. Using the Grab-Posisi dataset provides several advantages, including its high sampling rate and the inclusion of contextual information (accuracy, speed, bearing). This dataset is Indonesia's first GPS trajectory dataset and the most recent dataset for the mobility research, so it is still very relevant. Data collection over two weeks reflects community mobility patterns during both weekdays and weekends. Table 1 shows the meta data for Grab-Posisi dataset

Table 1. Grab-Posisi trajectory feature

Attribute	Data Type	Description
Trajectory_ID	string	identifier trajectory
Latitude	float	WGS84
Longitude	float	WGS84
Timestamp	bigint	UTC
Level Accuracy	float	circle radius, meter
Bearing	float	degree, relative to true north
Speed	float	meter/second

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

2.2. Methods

2.2.1. Proposed Semi-Supervised Learning Framework

The proposed semi-supervised framework involves a series of steps, adapted from the Cross-Industry Standard Process for Data Mining (CRISP-DM) for exploratory data science in GPS trajectories. These steps include business understanding, data understanding, data preparation, model building, evaluation, and deployment [39]. The proposed framework can be seen in Figure 1.

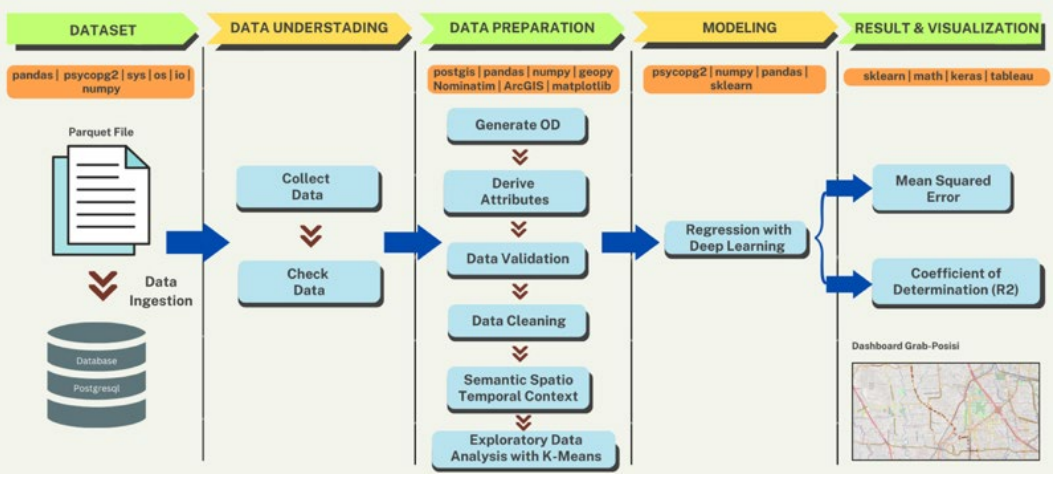


Fig. 1. Proposed Semi-Supervised Framework to Analysis GPS Trajectories

The following are the detailed steps proposed in the framework:

Data Ingestion

Grab-Posisi data is available in parquet file format. Therefore, the first stage is to carry out the data ingestion process into a PostgreSQL database. Data ingestion takes data from an external source, such as a parquet file, into the PostgreSQL database. The goal is to enable access, manage data efficiently (implementing a spatial index), and support analyze through the database.

Data Understanding

This phase aims to obtain an initial understanding of the spatiotemporal data by checking the data types and sizes, and by visualizing the data using QGIS software. The visualization shows all GPS trajectories on a map and helps researchers carry out an initial analysis of the data from both in population (aggregate) and individual views.

Data Preparation

The goal of data preparation is to improve data quality. We propose some steps to improve data quality, especially those related to travel behaviour analysis. Generating OD (Origin-Destination) involves transforming raw data into more meaningful and relevant data to obtain pairs of starting points (origin) and endpoints (destination) for each trajectory. Next, attribute derivation is based on spatial and temporal features to provide semantic context. One derived attribute is the distance travelled, which is calculated from the origin to the destination using the haversine formula, as shown in equations (1), (2), and (3) below

$$a = sin^2\left(\frac{\Delta\varphi}{2}\right) + cos\varphi_1 \cdot cos\varphi_2 \cdot sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (1)$$

$$c = 2 \cdot atan2\left(\sqrt{a}, \sqrt{(1-a)}\right) \quad (2)$$

$$d = R \cdot c \quad (3)$$

where:

φ: latitude

λ: longitude

R: earth's radius (mean radius = 6,371 km)

We also conduct data validation to ensure accuracy and integrity, along with data cleaning to remove invalid values. A data-checking stage is conducted to ensure that the data is valid, accurate, and of good quality. There are two essential things to assess the validity of the data:

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

Travel speed

Calculating the average speed based on distance and time for each location point sometimes produces negative values such as -1 km/h, which do not represent valid trips. Five trips were found with a negative average speed. The speed is then verified by recalculating it based on a comparison between the time and the straight-line distance between the origin and destination. In this calculation, no trip has a negative speed, indicating inconsistencies in speed-related data (travel time and distance).

Distance, time, and number of travel location points

Based on data analysis and visualization, it was found that twelve trips had a duration of 0 minutes, three trips covered 0 km, and three trips had only one location. Usually, a valid trip involve a minimum of two points: the starting and the destination points and some intermediate points. If there is only one point, the data must provide more information to analyze the trip, or in this case, we assume it is not a valid trip.

Based on the analysis above, we concluded that some data are inconsistent and do not represent actual trip, as indicated by negative average speed values, a concise trip durations, and a minimal distances travelled, probably because of the road conditions (in the tunnel) or the accuracy of the GPS. Therefore, action is needed to verify and eliminate invalid data to perform accurate travel-related analysis.

Researchers used percentiles to determine the minimum and maximum data limits. A percentile is a statistical measure used to divide sorted data or group data into several parts. In the context of this research, the 2.5% percentile was selected as the minimum limit and the 97.5% percentile as the maximum limit. This decision is based on the consideration that values between the 2.5% and 97.5% percentiles cover most of the data and minimize the possibility of errors in data evaluation. Using the 2.5% and 97.5% percentiles, researchers hope to produce a more objective and accurate analysis, providing more valid conclusions about the grab trajectory data studied.

Based on the origin-destination location pairs, context annotation for postal codes, road networks, and land use is carried out using external data sources such as OpenStreetMap, Nominatim, and ArcGIS library. The OD data, equipped with semantic context, is analyzed using the Density-Based Spatial Clustering (DBSCAN) algorithm to obtain origin-destination location clusters and uncover similarities between trajectories. We determine the best parameters (epsilon and minPoints) for the DBSCAN algorithm to obtain clusters with the best Silhouette coefficient values. Epsilon refers to the radius of the circle around each data point to inspect the density, while minPoints is the minimum number of data points required within that circle for the point to be classified as a core point. The Silhouette coefficient is a metric used to measure how closely related objects are within a cluster (cohesion) and how far apart a cluster is from others (separation), as shown in the following equation (4) and (5).

$$\mathrm{s(i)} = \frac{b(i) - a(i)}{\max\{\mathrm{a(i)}, \mathrm{b(i)}\}}, if\ |C_I| > 1 \quad (4)$$

$$\mathrm{s(i)} = 0, if\ |C_I| = 1 \quad (5)$$

The following algorithm in Figure 2 is a learning algorithm specifically designed for data preparation, with a focus on generating semantic Origin-Destination (OD) trajectories. Therefore, the output of the data preparation stage is a set of complete OD trajectories with 19 features.

Build Regression Modelling

In the modelling stage, we build a deep learning model based on the Recurrent Neural Network architecture to extract spatiotemporal features for predicting travel destinations. During the model-building process, the data is divided into training and test data. We use three deep learning models, Simple RNN, GRU, and LSTM, to find the best accurate prediction model. We also compare the results with general regression algorithms such as Linear Regression, Random Forest, Gradient Boosting, K Nearest Neighbour, Ridge Regression, and Decision Tree. Figure 3 shows the proposed LSTM architecture for feature extraction and prediction using origin-destination and contextual features. The architecture consists of three LSTM layers (50, 150, and 50 units); each LSTM layer is followed by a dropout layer to prevent overfitting and speed up the learning process.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

```
Input: S – Set of GPS trajectories (grouped by trj_id)
Output: C – Set of Origin-Destination clusters

#Preprocessing:
for each OD in S do;
    origin, destination = extract origin and destination (OD);
end for
#Derives Spatial Temporal features:
for each OD in S do;
     dayofweek, days, arriving_time, departure_time = extract temporal features (pingtimestamp);
     trj_duration = arriving_time – departure_time;
     distance = calculate haversine distance (origin, destination);
     speed_avg_coord = calculate average speed for each coordinate;
     speed_feature = calculate speed based on distance / duration;
     total_points = count total points(OD);
     add features to OD(dayofweek, days, arriving_time, departure_time, distance, duration, speed_avg, speed_feature, total_points);
end for
#Validation and Cleaning:
for each OD in S do;
    validate and clean(OD);
end for

#Semantic Enrichment with postcode and landuse:
for each OD in S do;
    load geopy, nominatim, ArcGIS, postgis (openstreetmap)
    postcode = get postcode(OD);
    origin_landuse=spatial_join(openstreetmap, origin, distance)
    destination_landuse=spatial_join(openstreetmap, destination, distance)
end for

#Clustering:
min_sample, epsilon = Grid Search Best Silhoutte Coefficient(OD)
Cluster = DBSCAN(min_sample, epsilon);
C <- Clustering and Silhoutte Coefficient;

# Visualization:
visualize clusters with map(OD, C);
visualize clusters PCA(C);
return C;
```

Fig. 2. Unsupervised Learning algorithm based on Origin-Destination

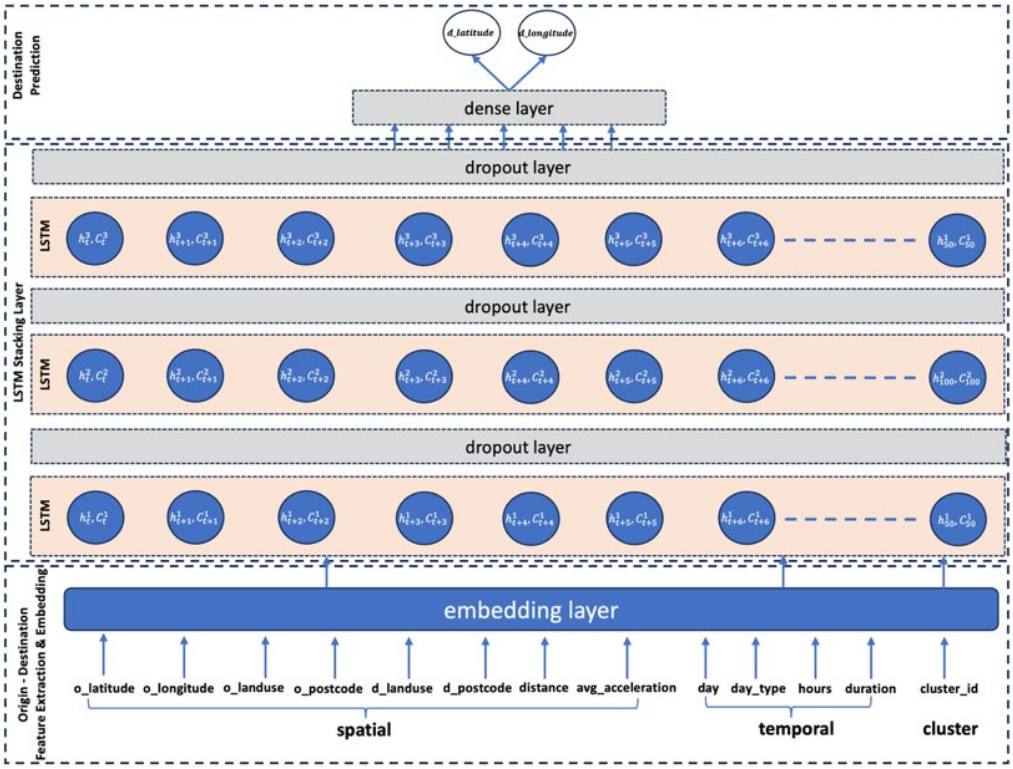


Fig. 3. LSTM Architecture for Origin-Destination Feature Extraction & Prediction

Evaluation (Result and Visualization)

In the final stage, the model's performance in predicting trip destinations is evaluated using Mean Square Error (MSE) and the Coefficient of Determination value (R2), as shown in equations (6) and (7). Mean Squared Error (MSE) is a standard loss function for regression cases that show the difference between the model's predictions and the actual data, square it, and get an average across the entire dataset.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (6)$$

http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

where:
MSE = mean squared error
N = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

The Coefficient of Determination measures how well the model predicts the dependent variable. The better a model makes predictions, the closer the R^2 score is to 1 (R^2 is [0,1]).

$$R^2 = 1 - \frac{RSS}{TSS} \quad (7)$$

where:
R^2 = coefficient of determination
RSS = sum of squares of residuals
TSS = total sum of squares

Model visualization through the dashboard built with Tableau is also used to verify the prediction results and distribution of origin-destination taxi trajectories.

2.2.2. System Architecture

We designed an architecture that supports the process of providing semantic annotations to GPS trajectory data. Figure 4 shows the architecture of system.

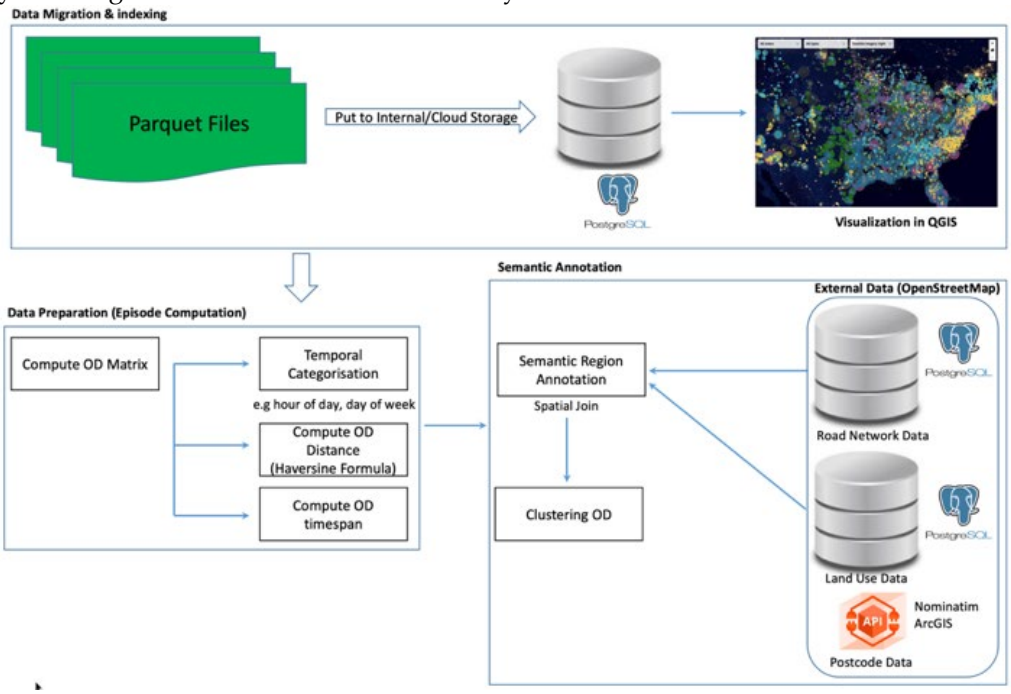


Fig. 4. System architecture semantic annotation on Grab-Posisi Dataset.

The process includes the extracting trajectory data from data sources (parquet files) provided by Grab. The data is then stored in a PostgreSQL database. Postgresql supports the implementation of spatial indexes and PostGIS functions, so that indexing, querying, and spatial operations to be carried out more quickly. Visualizing the trajectories through the QGIS software helps achieve an initial understanding of the data.

The most critical stage in this architecture is providing semantic annotation, which includes both spatial and temporal meaning. The Origin-Destination matrix and temporal features are generated. Then, the distance and travel time are also calculated based on the resulting origin and destination points. The postal code and land use context annotation process is carried out using external data. Postal code data are extracted using the nominal library and ArcGIS. Land use annotation is obtained through road networks and OpenStreetMap data. Annotation uses spatial join operations, namely ST_Within, ST_Within and ST_DWithin, by considering the maximum distance between the origin or destination to the land use polygon features as <= 1000 m.

http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

3. Results and Discussion

3.1. Data Preparation Result

The travel context in the grab data is obtained based on business understanding, data understanding, and data preparation, including Generating OD, Deriving spatiotemporal attributes, Data cleaning, and Context annotation for land use and postal codes). The following figures present some visualizations of the extracted context.

Figure 5 shows the origin-destination heatmap of Grab Taxi activity in the Jakarta region. Visualizing origin-destination data using a heatmap allows us to represent the intensity or density of trip across geographical locations. This approach is effective when dealing large datasets like Grab-Posisi. It provides valuable insights by identifying Grab mobility patterns and trends based on the aggregation of origin-destination data by neighbourhoods or regions. Hotspots (areas with higher intensity) and cold spots (areas with lower intensity) reveal the spatial context and valuable insights into Grab-Posisi activity. The departure locations are spread out, while the destination locations are concentrated towards the city centre. This shows that most trips go to downtown Jakarta.

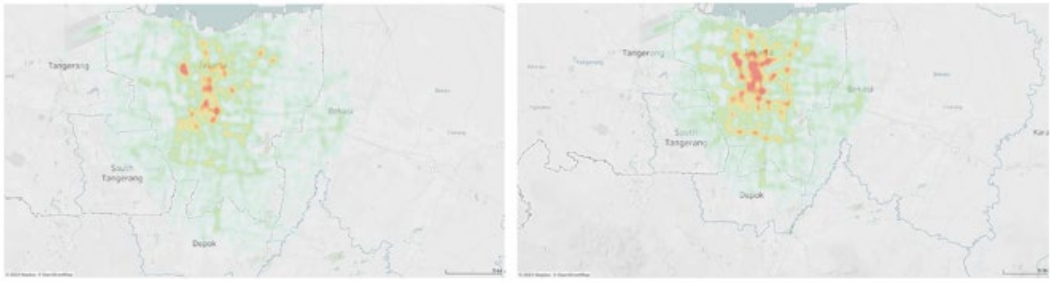
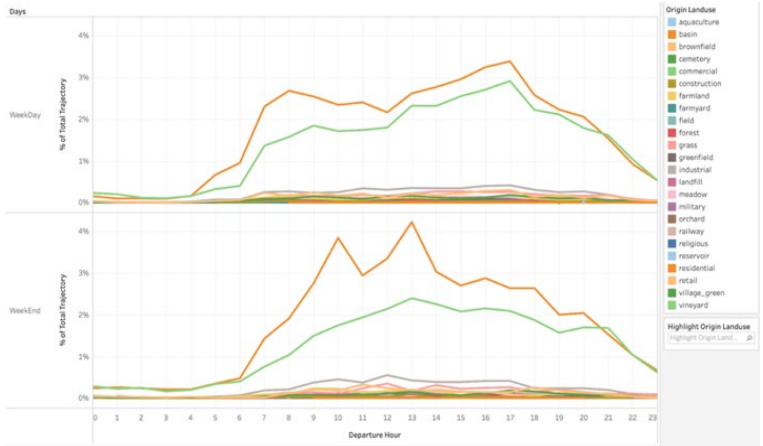


Fig. 5. Origin – Destination Heatmap

Figure 6 shows the temporal mobility pattern based on land use location popularity over a 24 hour period for weekdays and weekends, with origin locations in Figure 6(a) and destination locations in Figure 6(b). We extract the departure and arrival times to relate them to the purpose of the trip, referencing the land use features from OpenStreetMap annotated to the spatial origin and destination points. The temporal patterns reveal time-varying popularity, indicating that certain areas may interest individuals. The departure and arrival times show an identical trend, providing insight that the journeys are short, under one hour. From a spatial context, trips are dominated by trips departing from and heading to residential, commercial, and industrial areas.

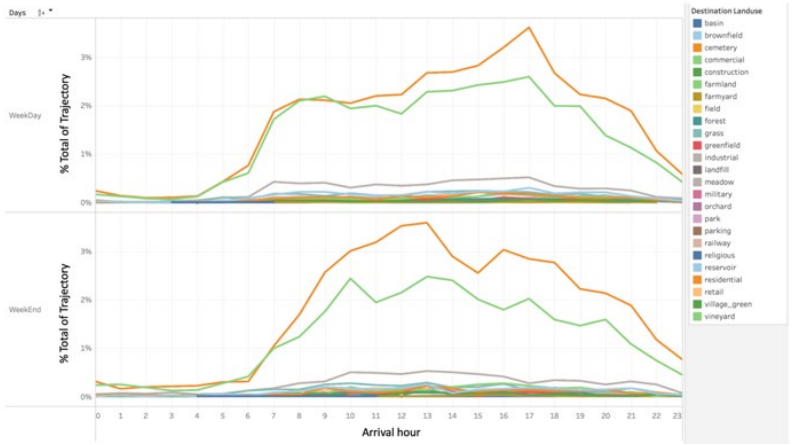
On weekdays, mobility increases significantly starting at 06.00 and 07.00, the morning activity to start work. There are peaks at 08.00 and 17.00, likely to be activities for going to work and returning home, respectively. Between these hours, the amount of mobility slowly increases. This pattern is consistent across major residential, commercial, and industrial areas. Then, after 17.00, the number of trips tends to decrease. On weekends, mobility from residential areas increases significantly starting at 08.00, with peaks observed at 10 and 13.00. This differs with commercial areas, where mobility rises slowly, peaks at 13.00, and decreases after 16.00.



(a)

http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

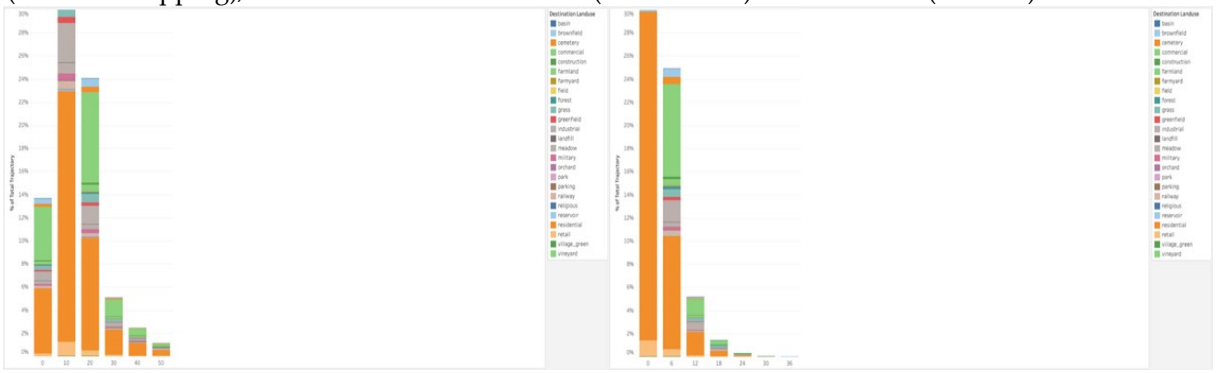


(b)

Fig. 6. Landuse popularity at origin (a) and destination (b) locations over 24 hours for weekends and weekday

At the destination location, on weekdays, trips to residential and commercial areas peak at 17.00. This shows a tendency for people to return home and visit commercial areas such as malls, shopping centres, and entertainment places. Meanwhile, on weekends, trips to residential places peak at 13.00, while trips to commercial areas peaks at 10.00 and 13.00. Visits to residential, commercial, and industrial areas tend to decrease after 17.00.

In data analysis, we also examined the association between trip distance or duration and the destination context based on land use features. Figure 7 shows how long and far it takes to reach specific destination areas, with trip duration in Figure 7(a) and trip distance in figure 7(b). As explained in the previous figure, travel destinations are dominated by residential, commercial, and industrial areas. We divided the trip duration into six bins of equal width, each spanning 10 minutes. The results show that more than 70% of trips are under 30 minutes, with most going to residential, commercial, and industrial places. This duration is directly proportional to travel distance, with around 61% of travel distances to the same areas being under 18 km. Trips to these locations, typically home, work and entertainment (malls or shopping), tend to be short in both time (<30 minutes) and distances (< 18 km).



(a) Duration (b) Distance

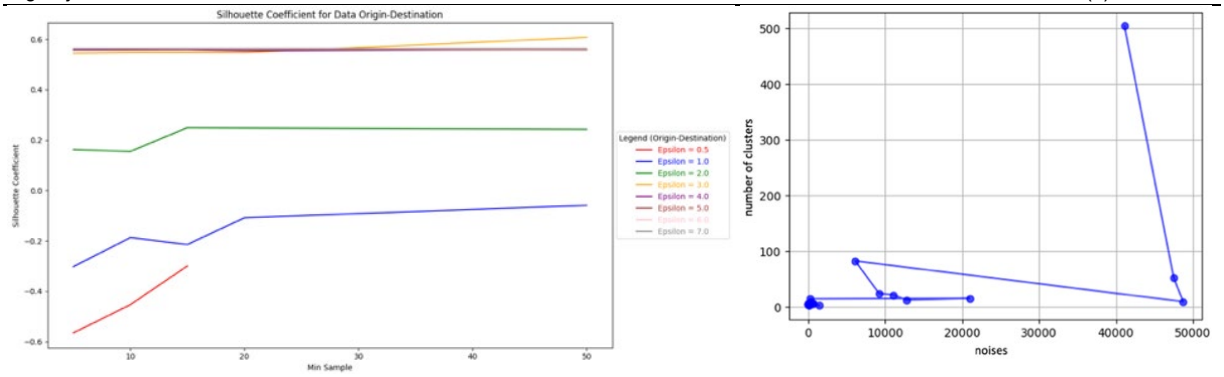
Fig. 7. Trip duration (a) and distance (b) to destination land use

3.2. Clustering Result

The experiments are carried out to find the best DBSCAN parameters in obtaining the maximum silhouette coefficient for origin-destination clustering. We use a range of epsilon values [0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0] and min_samples [5, 10, 15, 20, 50]. The experimental results can be seen in the Figure 7. Through Figure 8(a), we can see the visualization of the relationship between the epsilon, min_samples and silhouette coefficient values in origin-destination clustering. A low epsilon value usually leads to poor clustering, whereas raising the epsilon value typically enhances clustering quality. However, having the minimum number of samples for cluster formation does not notably improve the silhouette coefficient value. The improvement in clustering quality has its limits, as demonstrated by epsilon values between 3.0 and 7.0, where the silhouette coefficient remains relatively constant. The selection of appropriate epsilon and min_samples values is essential for obtaining optimal clustering results and may depend on the characteristics and context of the observed data.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90



(a) Silhouette coefficient value for pair epsilon and min_samples (b) Relationship between number of clusters with noise

Fig. 8. Silhouette coefficient value and noise generated for Origin-Destination Clustering

Figure 8(b) shows the relationship between the number of clusters and the amount of noise produced when clustering Origin-Destination data using the epsilon and min_samples parameter values as mentioned above. When fewer clusters are formed, the amount of noise tends to decrease significantly. This indicates that clustering provides more optimal results with a smaller number of clusters, where the amount of data excluded from clusters is also lower. In addition, there is a relationship between the silhouette coefficient and a smaller number of clusters. At high silhouette coefficient values (0.55-0.60), there is a tendency to form three or more clusters with a relatively low noise. This shows that the number of clusters formed in optimal clustering does not have to be large but provides good results by reducing the amount of noise. Connecting to the previous image, an epsilon value between 3.0 - 7.0 delivers a constant silhouette coefficient and a low level of noise.

Based on the experiments, the best parameters for clustering were found to be epsilon = 7.0 and min_sample = 5. These parameters produced a Silhouette coefficient value of 0.56, the highest obtained in this experiment, resulting in five clusters and four noise points. The number of clusters formed and the low noise indicate that these parameters deliver better clustering results compared to the other tested parameters. The following Figure 9 shows the best clustering of origin-destination data on the map.

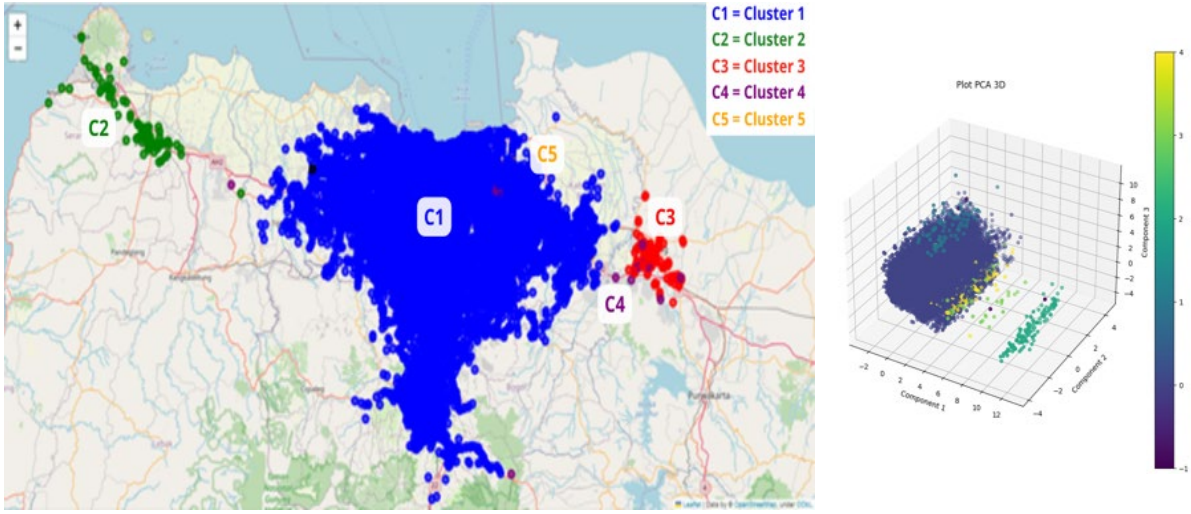


Fig. 9. Results of Clustering Origin-Destination Data on the map and 3D

The results of the formed clusters concluded that the clustering process successfully grouped origin-destination trips into five different groups. These results help to a better understanding of patterns and relationships within origin-destination data. A brief explanation of each cluster, covering distance, travel time, postal code, and land use is as follows:

Cluster 1 is referred to as the Travel cluster, with an average distance of 5.67 km and a duration of 19.15 minutes between districts or cities within the DKI Jakarta province. Most trips in cluster 1, 72.8% (35,461), occur on weekdays. This shows a mass movement to workplaces, which can be seen from the

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

departure time data showing an extreme increase at 6-7, 12-13, and 15-16. Cluster 1 also shows that trips are dominated by short, as indicated by arrival times that closely match departure times. All trips occur in areas with postal codes starting with 1XXXX. The first two digits of the postal code show that 77.32% (37,624) of trips are between different districts or cities (origin and destination are in other districts/cities. In more detail, the last two digits, representing the sub-district code, show that 91.11% (44,336) of trips have different origin and destination sub-districts.

Cluster 2 is summarized as a Trip cluster, with an average distance of 5.76 km and an average duration of 17.21 minutes between districts in the West Java province (origin postal code 4XXXX). For most trips, 84.27% (75) occurred within the same district or city (the first two digits of the postal code had the exact origin and destination). Furthermore, 70.8% (63) of trips took place on working days (weekdays), supported by departure time data showing a drastic increase between 6-8 and 18-19, the periods usually leaving and returning from work. The trips are generally short and brief, with 61.8% (55) falling below the average trip distance and 48.3% (43) below the average trip duration.

Like Cluster 2, Cluster 3 represents a journey with an average distance of 5.57 km and an average travel time of 17.82 minutes within the Karawang district, West Java sub-district. Analysis of the last two digits of postal codes shows that 80.51% (95) of trips occurred between different origin and destination sub-districts. Out of 118 trips in cluster 3; 71.2% (84) took place on weekdays, with departure and arrival times experiencing significant increases at 6-7, 12-13, and 16-17, times which are generally departure times, lunchtime, and back home from work. The data also shows that the journeys were relatively short and fast, with 55.93% (66) below the average distance and 52.54% (62) below the average duration. All trips (100%) originated from postal code 413XX, with 80.51% (95) involving different origins and destinations, while 19.49% (23) had the exact origin and destination.

Cluster 4 can be defined as a cluster with an average travel distance of 6.92 km and an average travel time of 20.40 minutes, typically made from West Java province to Jakarta on weekdays. There are 58.67% (44) trips below the average distance and duration, indicating that most trips are relatively short in both distance and short duration. Apart from that, the cluster contains 75 trips, with 76% (57) occurring on weekdays. Peak departure and arrival times at 6-8 and 15-16 strengthen this indication.

Cluster 5 is a travel cluster with an average distance and duration of 6.54 km and 19.06 minutes, representing trips from Jakarta to West Java province, indicated by postal codes starting with 1XXXX to 4XXXX. There are 67 trips in this cluster, with 64.2% occurring on weekdays. Peak departure and arrival times are observed at 7-8, 12-13, and 15-17, which are usually the times of leaving for work, lunch, and back home from work. There are 67.16% (45) trips below the average travel distance, and 61.19% (41) are below the average duration. This fact shows that most trips involve relatively short in both distance and duration.

In general, each cluster exhibits a unique pattern, even though they may be close together in geographic space. Cluster differences are influenced by spatiotemporal aspects such as travel days, travel times, postcodes, land use, and origin-destination points. For example, although clusters 1 and 3 include work trips on Tuesday afternoons, their points of origin and destination differ, making them clusters. These clustering result also complete the results obtained in previous explanatory data analysis with land use semantic annotation, indicating that the journeys are typically short in both time and distance.

3.3. Prediction Model Result

The data preparation and analysis stages of the proposed framework produce 16 features used to build a regression-based destination prediction model. A deep learning model based on RNN architecture is proposed to predict the travel destination coordinates (latitude, longitude) and is compared with baseline methods. The following Figure 10, Figure 11, Figure 12 are the results of experiments using deep learning methods carried out to estimate the next destination of a trip.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

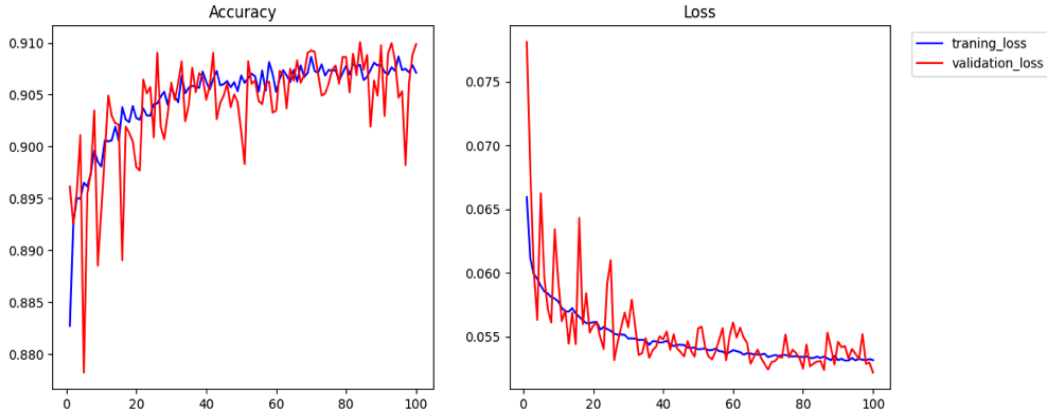


Fig. 10. Training result in destination prediction using SimpleRNN.

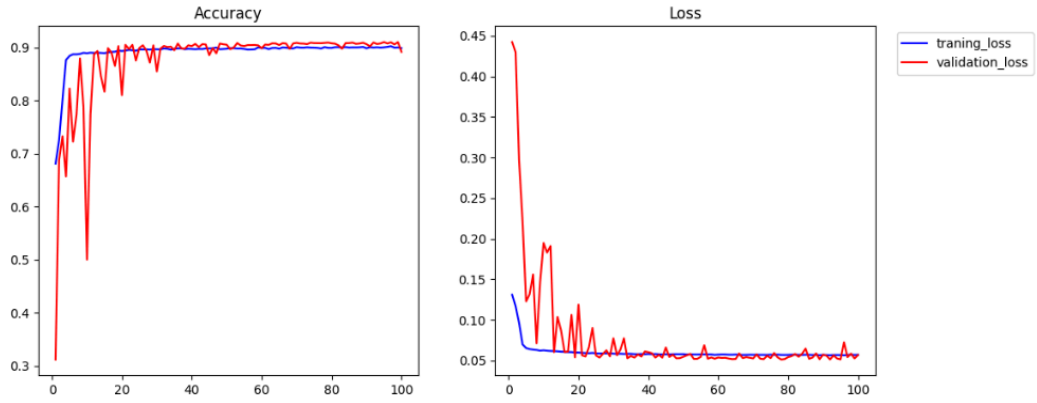


Fig. 11. Training result in destination prediction using GRU.

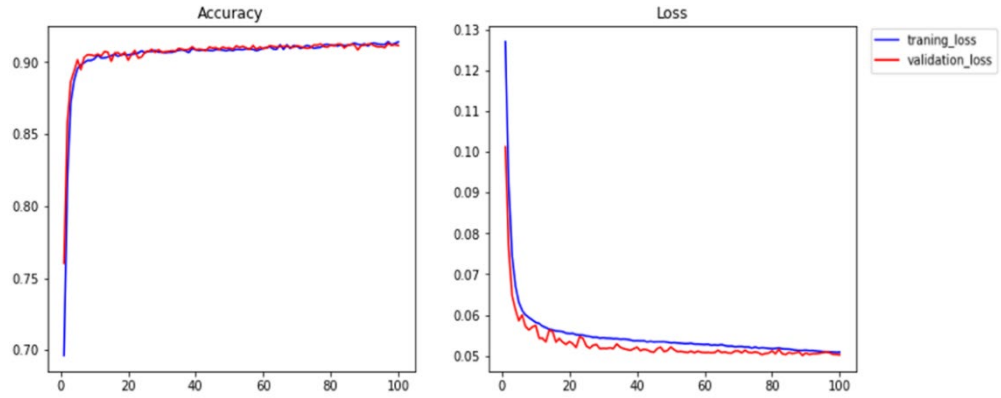


Fig. 12. Training result in destination prediction using LSTM.

Figure 10, Figure 11, and Figure 12 show the training and validation results of trip destination predictions using three variants of the RNN model. We can see that training and validation prediction carried out using the LSTM algorithm, with a number of epochs of 100 provides more stable results compared to SimpleRNN and GRU. The larger number of layers in the LSTM model, along with the inclusion of a dropout layer, enables it to better extract travel patterns from origin-destination, which have been given semantics. We use seven neural network layers in the LSTM model. In detail, the following Figure 13 and Figure 14 compare prediction results obtained using baseline regression algorithms and deep learning methods. The overall average values are MSE = 0.0060 and R2 = 0.841, clearly showing that all algorithms can effectively learn travel patterns through origin-destination data features that have been annotated with context data. In general, LSTM provides stable training and validation performance and consistent MSE and R2 values.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

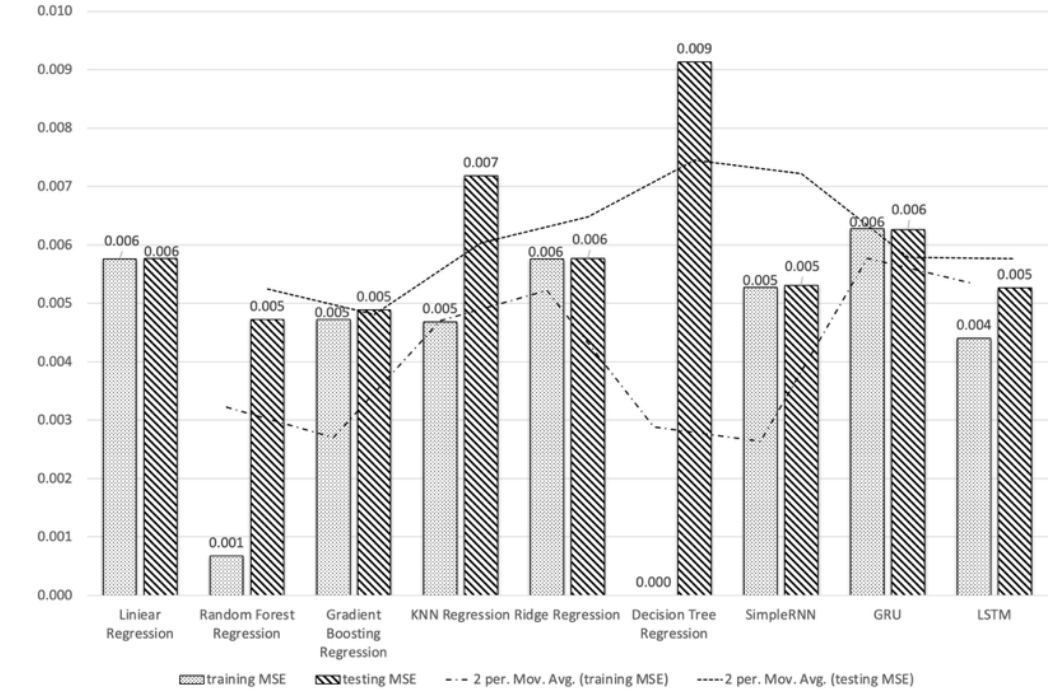


Fig. 13. MSE score regression prediction model on semantic annotated origin-destination data.

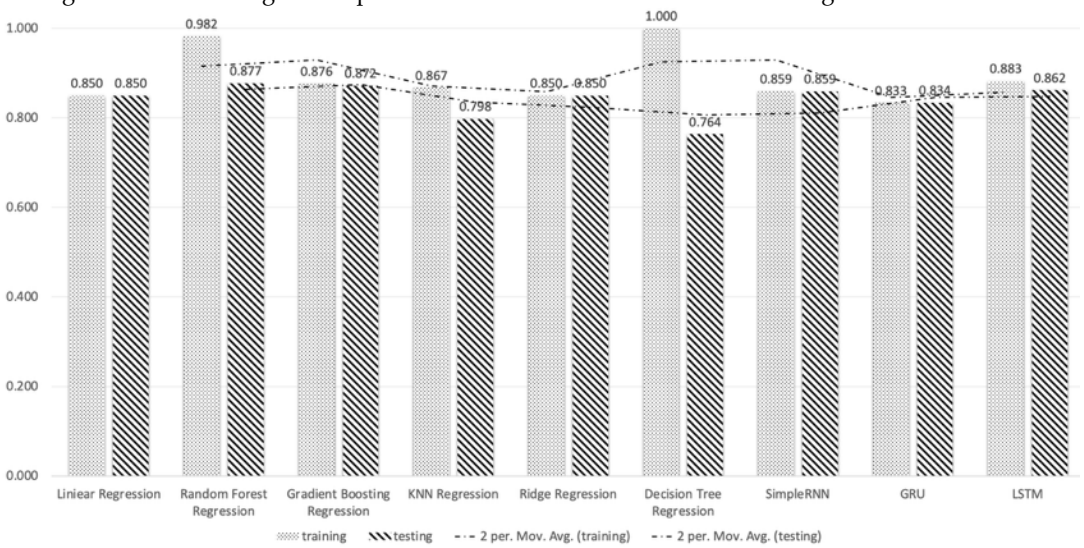


Fig. 14. R2 score regression prediction model on semantic annotated origin-destination data.

Our proposed method, which consists of some CRISP-DM steps, can produce more meaningful Origin-Destination data from trajectory inputs. The annotation of several attributes, spatial, temporal, and clustering, provides contextual information that enables a detailed understanding of taxi trips. We found that it is possible to analyze the attractiveness of residential, commercial, and industrial facilities from the perspective of spatial accessibility. Our study revealed that most trips were short in distance (< 7 km) and duration (< 20 minutes). The data also shows that the increase in taxi trips occurs during typical peak times, namely morning (07.00), midday hours (12.00) and afternoon (16.00). Prediction results across all methods (both baseline and neural network) give good results. All methods achieve MRSE values below 0.009 and R2 scores above 0.764 (methods based on neural methodology get an R2 value above 0.85). These further strengthens the idea that contextual annotation provides an easy method for providing meaning to taxi trips. Our proposed semantic annotation feature can learn and utilize public data interaction information from taxi GPS trajectories data better and significantly. To summarize, our approach is able to learn and better leverage trip interaction information extracted from Grab-Posisi trajectory data significantly.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

4. Conclusion

We proposed a semi-supervised framework for building semantic annotation analysis and destination prediction based on historical GPS taxi trajectories. We evaluated our model using a high-resolution Grab-Posisi dataset in Jakarta, Indonesia. Recent observations suggest that semantic annotation based on spatial and temporal features, exploits external datasets (postcode, OpenStreetMap), and explains the context of the trip in greater detail. Our findings provide conclusive evidence that a well-structured origin-destination cluster (silhouette value of 0.56) can explain the grouping of taxi trips clearly, recognizing natural places and time-varying patterns. Furthermore, the regression-based destination prediction model also performs significantly with annotated data, which gives significant results across both deep learning and baseline methods (average R2 = 0.841 and average MSE = 0.0060), confirming that our proposed framework is better and more applicable to the semantic annotation data. Our study demonstrates no substantial difference between deep learning and baseline methods in predicting trip destinations. Future work may extend this study by exploring alternative approaches for generating semantic context and trajectory embeddings, and utilizing datasets spanning multiple years and encompassing lower-resolution GPS data across broader regions of Indonesia. Additionally, subsequent research will consider integrating other valuable data sources to enhance trip context, such as weather information from BMKG, event schedules, and satellite-derived remote sensing data to capture land cover and urban dynamics, thereby enriching the analytical framework and enhancing interpretability of mobility patterns.

Author Contributions

H.T.A. Simanjunyak: Conceptualization, data curation, formal analysis, validation, experimentation, funding acquisition, investigation, methodology, project administration, resources, supervision, and writing - review & editing. A. Hutauruk: data curation, investigation, experimentation, visualization, and writing - original draft. H. Situmorang: data curation, investigation, experimentation, visualization, and writing - original draft. Y. Silitonga: data curation, investigation, experimentation, visualization, and writing - original draft.

Acknowledgment

The authors would like to thank to Grabtaxi Holdings Pte Ltd. for providing the Grab-Posisi Dataset for this research. Furthermore, we also gratefully acknowledge the support of LPPM Institut Teknologi Del through its internal funding schemes.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] P. Jittrapirom, V. Caiati, A.-M. Feneri, S. Ebrahimigharehbaghi, M. J. A. González and J. Narayan, "Mobility as a Service: A Critical Review of Definitions, Assessments of Schemes, and Key Challenges," Urban Planning, vol. 2, no. 2, pp. 13-25, 29 June 2017.

[2] V. Bogorny, C. Renso, A. R. de Aquino, F. d. L. Siqueira and L. O. Alvares, "CONSTAnT – A Conceptual Data Model for Semantic Trajectories of Moving Objects," Transaction in GIS, vol. 18, no. 1, p. 66–88, February 2014.

[3] M. Bevis, J. Bedford and D. J. Caccamise II, "The Art and Science of Trajectory Modelling," Geodetic Time Series Analysis in Earth Sciences, pp. 1-27, August 2020.

[4] H. Noureddine, C. Ray and C. Claramunt, "Semantic Trajectory Modelling in Indoor and Outdoor Spaces," in 21st IEEE International Conference on Mobile Data Management (MDM), France, 2020.

[5] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in 1984 ACM SIGMOD international conference on Management of data, 1984.

[6] S. Shang, R. Ding, B. Yuan, K. Xie, K. Zheng and P. Kalnis, "User oriented trajectory search for trip recommendation," in 15th international conference on extending database technology, 2012.

[7] N. Andrienko, G. Andrienko, N. Pelekis and S. Spaccapietra, "Basic Concepts of Movement Data," in Mobility, Data Mining, and Privacy Geographic Knowledge Discovery, Springer Berlin Heidelberg, 2008, p. 15–38.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

[8] D. Kumar, H. Wu, Y. Lu, S. Krishnaswamy and M. Palaniswami, "Understanding Urban Mobility via Taxi Trip Clustering," in 17th IEEE International Conference on Mobile Data Management (MDM), 2016.

[9] D. Zhang, K. Lee and I. Lee, "Mining hierarchical semantic periodic patterns from GPS-collected spatio-temporal trajectories," Expert Systems with Applications, vol. 122, pp. 85-101, 15 May 2019.

[10] S. Wang, . G. Mei and S. Cuomo, "A generic paradigm for mining human mobility patterns based on the GPS trajectory data using complex network analysis," Concurrency and Computation: Practice and Experience, vol. 33, no. 4, 25 February 2021.

[11] S. Dutta, A. Das and B. K. Patra, "CLUSTMOSA: Clustering for GPS trajectory data based on multi-objective simulated annealing to develop mobility application," Applied Soft Computing, vol. 130, no. 109655, November 2022.

[12] S. Wang, X. Niu, P. Fournier-Viger, D. Zhou and F. Min, "A graph based approach for mining significant places in trajectory data," Information Sciences, vol. 609, pp. 172-194, September 2022.

[13] W. Li, H. Zhang, R. Shibasaki, J. Chen and H. H. Kobayashi, "Chapter two - Mining individual significant places from historical trajectory data," Handbook of Mobility Data Mining, vol. 2, pp. 15-26, 2023.

[14] J. Wang, W. Jiang and J. Jiang, "LibCity-Dataset: A Standardized and Comprehensive Dataset for Urban Spatial-temporal Data Mining," Intelligent Transportation Infrastructure, liad021, 7 November 2023.

[15] W. Tu, H. Ye, K. Mai, M. Zhou, J. Jiang, T. Zhao, S. Yi and Q. Li, "Deep online recommendations for connected E-taxis by coupling trajectory mining and reinforcement learning," International Journal of Geographical Information Science, vol. 38, no. 2, pp. 216-242, 2024.

[16] C. Chu, H. Zhang, P. Wang and F. Lu, "Simulating human mobility with a trajectory generation framework based on diffusion mode," International Journal of Geographical Information Science, vol. 38, no. 5, 06 February 2024.

[17] L. Gong, S. Guo, Y. Lin, Y. Liu, E. Zheng and Y. Shuang, "STCDM: Spatio-Temporal Contrastive Diffusion Model for Check-In Sequence Generation," IEEE Transactions on Knowledge and Data Engineering, pp. 1-14, 10 January 2025.

[18] A. Crivellari and Y. Shi, "Generative adversarial deep learning model for producing location-based synthetic trajectory data," Connection Science, vol. 37, no. 1, p. 2458502, 30 January 2025.

[19] Z. Xu, Y. Yin, C. Dai, X. Huang, R. Kudali, J. Foflia, G. Wang and R. Zimmermann, "Grab-Posisi-L: A Labelled GPS Trajectory Dataset for Map Matching in Southeast Asia," in Proceedings of the 28th International Conference on Advances in Geographic Information Systems, 2020.

[20] W. Bian, G. Cui and X. Wang, "A Trajectory Collaboration Based Map Matching Approach for Low-Sampling-Rate GPS Trajectories," Sensors, vol. 20, no. 7, 6 April 2020.

[21] M. Liu, L. Zhang, J. Ge, Y. Long and W. Che, "Map Matching for Urban High-Sampling-Frequency GPS Trajectories," ISPRS International Journal of Geo-Information, vol. 9, no. 1, p. 31, 5 January 2020.

[22] M. Dogramadzi and A. Khan, "Accelerated Map Matching for GPS Trajectories," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 5, pp. 4593-4602, May 2022.

[23] L. Ruback, M. A. Casanova, A. Raffaetà, C. Renso and V. Vidal, "Enriching Mobility Data with Linked Open Data," in Proceedings of the 20th International Database Engineering & Applications Symposium, 2016.

[24] Y. Gao, L. Huang, J. Feng and X. Wang, "Semantic trajectory segmentation based on change-point detection and ontology," International Journal of Geographical Information Science, vol. 34, no. 12, pp. 2361-2394, 2020.

[25] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra and K. Aberer, "Semantic trajectories: Mobility data computation and annotation," ACM Transactions on Intelligent Systems and Technology, vol. 4, no. 3, p. 1–38, 01 July 2013.

 http://doi.org/10.26594/register.v11i2.5388

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
11 (2) 2025 75-90

[26] S. Hwang, C. Evans and T. Hanke, "Detecting Stop Episodes from GPS Trajectories with Gaps," in Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics, Springer Nature, 2017, p. 427–439.

[27] H. Simanjuntak and F. Ciravegna, "Semantic Understanding of Human Mobility Lifestyle to Support Crisis Management," in Proceedings of the 16th International Association for Information Systems for Crisis Response and Management (ISCRAM) Conference, 2019.

[28] D. Guo, X. Zhu, H. Jin, P. Gao and C. Andris, "Discovering Spatial Patterns in Origin-Destination Mobility Data," Transaction in GIS, vol. 16, no. 3, pp. 411-429, June 2012.

[29] Y. Liang, Z. Zhao and X. Zhang , "Modeling taxi cruising time based on multi-source data: a case study in Shanghai," Transportation, vol. 51, p. 761–790, 2024.

[30] R. F. Jonaghani, M. Wachowicz and T. Hanson, "Matrix Factorization for Globally Consistent Periodic Flow Prediction in Taxi Systems," Transportation Research Record: Journal of the Transportation Research Board, vol. 2678, no. 5, pp. 1-12, May 2024.

[31] A. Monreale, F. Pinelli, R. Trasarti and F. Giannotti, "WhereNext: a location predictor on trajectory pattern mining," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009.

[32] J. Zhang, Y. Zheng and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," in Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[33] X. Zhang, Z. Zhao, Y. Zheng and J. Li, "Prediction of Taxi Destinations Using a Novel Data Embedding Method and Ensemble Learning," IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 1, pp. 68-78, 1 January 2020.

[34] J. Zhao, L. Zhang, J. Ye and C. Xu, "MDLF: A Multi-View-Based Deep Learning Framework for Individual Trip Destination Prediction in Public Transportation Systems," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 13316-13329, 2022.

[35] M.-F. Chiang, E.-P. Lim and J.-W. Low, "On Mining Lifestyles from User Trip Data," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[36] X. Guangnian, Z. Juan and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," Transportation Research Part C: Emerging Technologies, vol. 71, pp. 447-463, October 2016.

[37] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," Transportation Research Part C: Emerging Technologies, vol. 77, pp. 96-112, April 2017.

[38] X. Huang, Y. Yin , S. Lim, G. Wang, B. Hu, J. Varadarajan, S. Zheng, A. Bulusu and R. Zimmermann, "Grab-Posisi: An Extensive Real-Life GPS Trajectory Dataset in Southeast Asia," in Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, 2019.

[39] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana and P. Flach, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048 - 3061, 01 August 2021.