# Traffic Speed Prediction From GPS Data of Taxi Trip Using Support Vector Regression

Dwina Satrinia
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
dwinasatrinia@gmail.com

G.A. Putri Saptawati
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
putri@informatika.org

*Abstract*— **Traffic congestion prediction is one of the solution to overcome congestion problem. In this paper, we propose a development of system that can predict traffic speed with help of GPS data from history of taxi trip in Bandung city. GPS data from taxi trip in Bandung city does not have data speed and sometimes the location detected from GPS device is less accurate so additional steps required in data preprocessing phase. We proposed using Map Matching with topological information method in pre-processing phase. Map Matching will produce a new trajectory that has corresponded to the road. Then, from that new trajectories we calculate speed for each road segment. To predict traffic speed in the future we utilize Support Vector Regression (SVR) method. The results of this study indicate that Map Matching can help to obtain more accurate traffic speed and SVR has good performance to predict the traffic speed.**

***Keywords— GPS data; Map Matching; SVR;***

## I. INTRODUCTION

Traffic congestion is often occurred in urban areas and need to be solved. One of solution to overcome this problem is traffic state prediction. Speed, degree of saturation and traffic density are good parameter to represent traffic state condition. Data source that can be used to describe traffic condition are traffic sensors and floating car data (FCD) such as GPS data from GPS device or smartphone used while driving. Nowadays, taxi companies in urban areas added GPS device on their fleet so they can see insight of it. Taxi fleet in Bandung city also has installed GPS device. GPS data from taxi trip are easy to collect and has attribute such as GPS ID, spatial information (longitude, latitude) and temporal information (time taken). From GPS data that has spatial and temporal information, we can determine or even predict traffic congestion. We can determine traffic speed by calculating mileage divide by travel time. The mileage from each taxi trip can be calculated by summing the distance between each points from the location point of GPS data. But location detected from GPS device often inaccurate and not lie on the road. This can lead to miscalculation of mileage and speed. Additional process is required to overcome the problem so that the mileage and speed calculation becomes more appropriate.

White, et al. [1] proposed Map Matching algorithm for matching *inaccurate* GPS data with road network. They added route search algorithm in Map Matching process. Map Matching will produce new trajectory which location point of vehicle were exactly lie on the road. This method is suitable to overcome the problem that existed in GPS data of taxi trip that inaccurate to determined vehicle location point.

To predict traffic congestion, we need to predict one of its parameter that is traffic speed. In previous research, Asif et al. [2] used clustering method for grouping road characteristics, to predict traffic speed in every cluster of roads, they used Support Vector Regression (SVR) method. They compare the prediction result between SVR method, Exponential Smoothing, and Neural Network (ANN). In that study, Asif used traffic sensor data. The result of their study is SVR has good accuracy to predict traffic speed beyond ANN and exponential smoothing in some roads cluster.

Another work about traffic state prediction is done by Kong et al. [3] and Yang et al. [4] which predict traffic congestion using GPS data from taxi trip. They use Support Vector Machine (SVM) to predict future traffic speed and volume then utilize Particle Swarm Optimization (PSO) to optimize punish coefficient C. Then they use Fuzzy Congestion Evaluation to change parameter congestion to level of congestion.

This study is motivated from the previous work to predict traffic speed using SVR method by utilize data GPS from taxi trip in Bandung City. To improve the quality of the data so it will produce mileage and speed calculation becomes more appropriate, we proposed Map Matching method in pre-processing phase before calculating traffic speed. Map matching method will produce new trajectory that has corresponded to the road and will reduce inaccurate position obtained by GPS device. Furthermore, this research can be adopted by government to predict future traffic speed in Bandung City.

This paper is structured as follows. In Section II, we briefly describe related works about prediction of traffic congestion. In Section III, we describe about Map Matching and SVR method that use in this paper and in Section IV, we explained our proposed architecture. Finally, in Section V, we summarize our contributions, and we suggest topics for future work.

978-1-5386-1449-5/17/$31.00 ©2017 IEEE

## II. Related Works

In this section, we present an overview of urban traffic congestion estimation and prediction methods based on the recent literature.

Lu et al [5] develop system to evaluate traffic state condition using new index named level of congestion (LOC). In their system, they used Fuzzy Inference with factor set are speed and density. Yoon et al. [6] identify traffic state using Quadrant clustering and success to characterized traffic pattern with 90% of accuracy. Y.C Zhang [7] presented work that detect congestion from main road in China using GPS data from taxi trip. Zhang use Map Matching method before extracting traffic speed to obtain accurate data speed. Another work about traffic prediction is done by Kong et al [3] and Yang et al. [2]. They developed system that predict traffic congestion using GPS data taxi and use the combination of Support Vector Machine (SVM) and Particle Swarm Optimization (PSO) to predict traffic congestion parameter. PSO module will optimized punish coefficient that used in SVM to predict speed and traffic volume. Then another module named Congestion State Fuzzy Division (CSFD) that utilized Fuzzy Congestion Evaluation (FCE) method, used to change value of traffic congestion parameter such as speed, density and degree of saturation to citizen's cognitive congestion state. Another work about traffic prediction is done by Asif et al. [2]. They classify the road segment so that the pattern of spatiotemporal data can be seen. Then use SVR to predict traffic speed for each cluster. The result of this research is good accuracy performed by SVR and outperformed ANN and Exponential Smoothing.

## III. Proposed Method

In this section, we describe Map Matching method to refined trajectory of taxi trip and SVR method for predicting traffic speed. Finally, the whole process of our proposed method is described.

### *A. Map Matching*

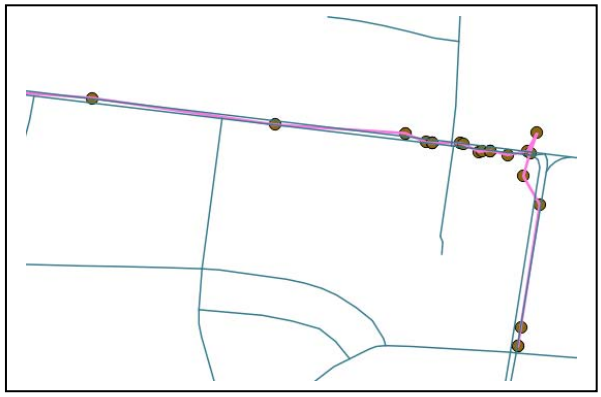

Fig. 1. Trajectory with inaccurate position

Map Matching is simple search problem, which the idea is to find the closest point P to the matched road network. This approach is good to refined points from trajectories that has inaccurate position from GPS data. Figure 1 and 2 show trajectory before and after map matching respectively.

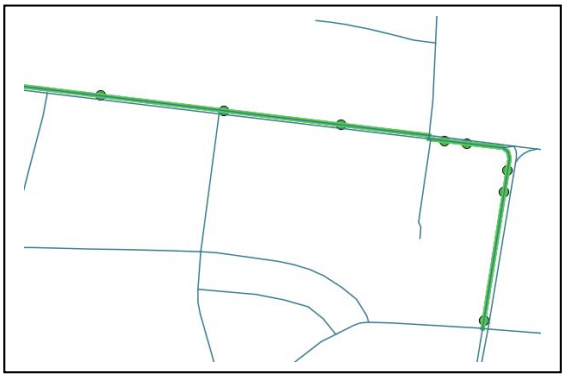

Fig. 2. Trajectory after Map matching process

There are many approaches in map matching problem and we will discuss three common approach such as point-to-curve, curve-to-curve approach and map matching with topological information.

Point-to-curve approach is the simplest approach in Map Matching. It will find the projection point from candidates of selected road that has minimum distance from the original point. To select candidates of selected road, the distance between the original point and the road must still be within a certain range. This approach is simple, but has drawback that it doesn't estimate the projection point from history of another point so the chosen projection point can be misplaced.

Curve-to-curve approach is developed based on point-to-curve approach. This approach take into account the previous point from history. Just like point-to-curve approach, road candidate are selected from road that has the distance between original point must still be within a certain range. It will find projection point from every candidate roads, then make new trajectory from every projection point and calculate the distance between this new trajectory and the old trajectory from original points of GPS data. The new trajectory that has minimum distance from old trajectory will be chosen.

Another approach is map matching with topological information. This approach is proposed by C. White et al. [1], not only use range query to find road candidate, but also it will use topology of road network to find the next matched road. The road candidates are chosen by topology of road network that are reachable from current node or road segment. A route search algorithm is needed in the process of finding the road candidate in this approach. If the chosen road from route search algorithm are not "good", then the road segment will choose from range query as on point-to-curve approach.

In this paper, we'll choose map matching approach with topological information as it is applicable and has good performance based on result of research work C. White et al. [1].

### *B. Support Vector Regression (SVR)*

The data traffic speed that can be taken from GPS data of taxi trip, can be in the form of time series. To predict time

series data required time series analysis such as regression, moving average, etc. In the case of time series data prediction using regression as a model, Support Vector Regression (SVR) used with the same concept with SVM. Muller et al. [8] explains that the idea of SVR is to map data into the high-dimensional feature space using mapping functions called kernel and perform regression in space. Equation 1 show equation used for SVR.

$$f(x) = \sum_{i=1}^{l} (a_i - a_i^*) K(x, x_i) + b \quad (1)$$

Where K(x, $x_i$) is kernel function.

Kernel function that commonly used are linear, Radial Basis Function (RBF), and polynomial, equation 2, 3, and 4 show these kernel function consecutively.

$$K(x,y) = x.y \quad (2)$$

$$K(x,y) = exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \quad (3)$$

$$K(x,y) = (x^T y + c)^d \quad (4)$$

## IV. PROPOSED ARCHITECTURE

### A. Data

Data used in this study are GPS data from taxi trip, where the data is semi structure file .csv that has many column. Figure 3 shows the GPS data obtained from taxi fleet in Bandung City. The marked columns are the attributes used in this study (GPS ID, time data taken, longitude and latitude). On this GPS data there is no speed data, so it takes a process to calculate the speed of traffic based on existing data in order to be able to predict the speed in the future.

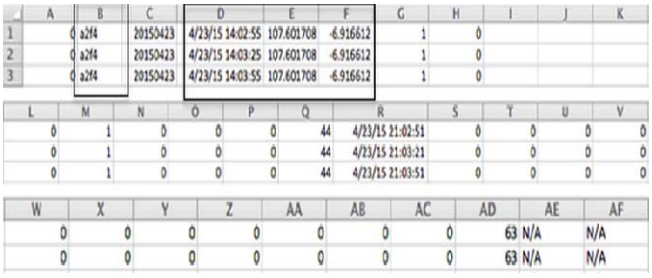


Fig. 3. Selected attribute of GPS Data from taxi trip

Other data needed for this study is the roadmap of Bandung city in shapefile format (.shp) that can be obtained from Open Street Map. The Shapefile format does not have topology information. So it can not be used to find something that requires information related to topology such as finding a connected path, finding a route, etc. Beside that, attribute of each road segment not yet equipped with long distance road segment. We need to create topology from road network and define distance of every each road segment.

### B. Proposed Architecture

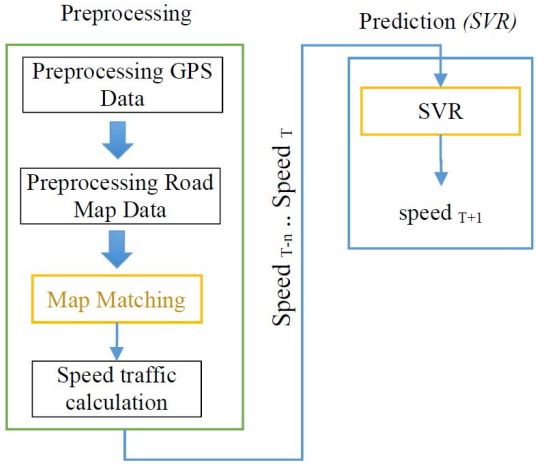


Fig. 4. Proposed Architecture

Figure 4 is the proposed architecture in this study. There are 5 main steps: preprocessing GPS data, preprocessing road map data, Map Matching, speed calculation and speed prediction using SVR method.

### C. Preprocessing GPS Data

This phase produces new structure of GPS data which can be used for Map matching process and prediction using SVR. It consists of four main steps as shown in figure 5 as follow:

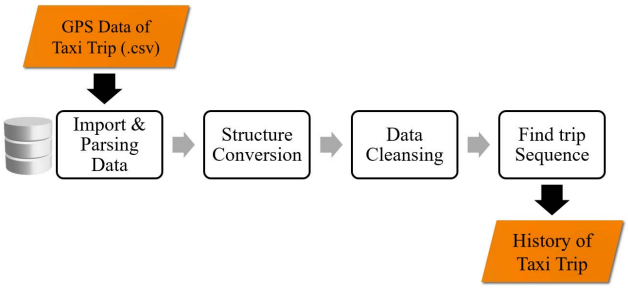


Fig. 5. Preprocessing GPS Data

#### 1) Import & Parsing Data

This step conducts selection of attributes on semi structure .csv formatted file of GPS data taxi trip. The selected attribute are the attribute that has temporal properties such as time of data taken, and has spatial properties such as latitude and longitude. In this study, the selected attributes are shown in table 1. After select the attribute, the next step is import data to the structured database PostgreSQL so it easy to collect.

TABLE I. THE SELECTED ATTRIBUTE

| Attribute | Description |
|---|---|
| GPS_ID | ID of GPS, every taxi fleet has unique GPS ID |
| Datetime | Time when data is taken |
| Latitude | Latitude coordinate |
| Longitude | Longitude coordinate |

*2) Structure Conversion*

In conversion structure process, attribute longitude and latitude from GPS data will change into spatial data type named geometry point in the database. In this step, the date format also change into the specified format so it has same format. This step is required so spatial query can be done from it.

*3) Data cleansing*

This step will remove the noise data which point from taxi trip that has zero or "null" value in attribute longitude and latitude.

*4) Finding set of points in one trip*

The purpose of this steps are looking for a set of points that exist in one trip and labeled them. First we looking for time-based connectivity and taxi IDs so as to find a series of points that include into a single trip. The points that found in one trip will get same label.

## D. Preprocessing Roadmap Data

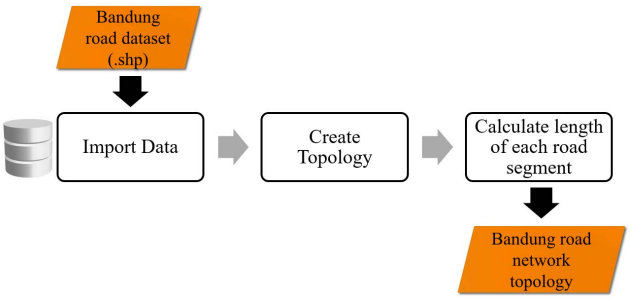


Fig. 6. Preprocessing roadmap data

Preprocessing phase also done to roadmap data. The figure 6 shows step that conduct in this phase.

*1) Import Data*

Roadmap data of Bandung city was in shapefile format. This file needed to be imported to structure database PostgreSQL using tools **PostGIS 2.0 Shapefile and DBF Loader Exporter**. Figure 7 shows example of roadmap data after import data process.

| gid integer | id double p | osm_id numeric | type character | name character varying (48) | tunnel integer | bridge integer | oneway integer | ref char |
|---|---|---|---|---|---|---|---|---|
| 225 | 225 | 4625566 | residen... | Jalan Terusan Sutami 2 | 0 | 0 | 0 | [null |
| 226 | 226 | 4625573 | residen... | Jalan Terusan Sutami 1 | 0 | 0 | 0 | [null |
| 227 | 227 | 4625575 | residen... | Jalan Terusan Sutami | 0 | 0 | 0 | [null |

| z_order double pr | access characte | service characte | class character | geom geometry |
|---|---|---|---|---|
| 3 | [null] | [null] | highway | 0102000020E6100000030000006 |
| 3 | [null] | [null] | highway | 0102000020E6100000050000004 |
| 3 | [null] | [null] | highway | 0102000020E6100000070000007 |

Fig. 7. Roadmap data after imported to database

*2) Create topology*

To perform map matching using map matching with topology information approach, we need to create topology that define the connected road or road network. This step will create topology from Bandung city roadmap that will added node/point source and target from every road segment and create table vertices that contain list of point that become node at roadmap network. To create topology, Postgresql database has provide **pgRouting** extension that has procedure named createTopology.

*3) Calculate length of each road segment*

In this step, the length of each road segment is calculated using spatial query that provide by PostGIS extension. The result will be saved to table in column cost_len with meter as a unit.

## E. Map Matching Phase

We do Map matching to get the correct position of GPS data, so when we calculate the distance of taxi trip, it will produce a more precise distance. The step that existed in map matching with topological information approach shown in figure 8 and the descriptions are as follow:

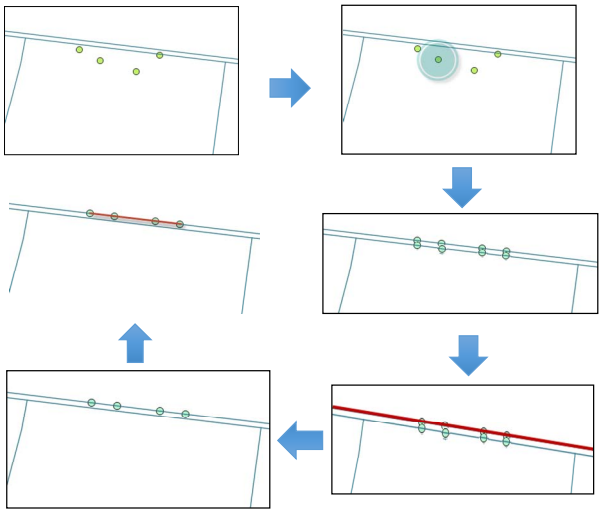

Fig. 8. Map Matching Process

- Find road candidate from every point in one trip within radius ± 50 meter.
- Calculate the distance between point and candidate road segment.
- Find projection point from every point to candidate road segment.
- Find start and end point from every trip along with finding start and end time of trip.
- Find the node of road segment from candidate road segment that has minimum distance to start and end point.

- Find route that has shortest path from start node to end node using Dijkstra algorithm. This process use topological information to find the reachable road segment. The result of this process is list of road segment that may be passed by taxi trip points.
- Locate the projection point that passed by the route chosen by the Dijkstra algorithm, then calculate the error. If there are many point could be located, then the process stop and the result are projection point that lie in chosen path by Dijkstra algorithm, but if vice versa, the projection point that has minimum distance from points will be chosen.
- The original point from GPS data will be replaced by the chosen projection point from the earlier step.

### F. *Calculating Traffic Speed*

After done map matching process, the location form GPS data has been accurate. To predict the traffic speed in the future, we need to calculate traffic speed in road segment from GPS data of taxi trip history that has been preprocessed with map matching. The step to calculate the traffic speed per road segment are as follows:

- Calculate average speed from each road segment that has at least 2 GPS data points on the same road segment through calculate total mileage of the point divide by the delta time between the start and end time in the same road segment. This average speed become the traffic speed in that road segment.
- If on the road segment $R_{t1}$ that include in the chosen route there is no point GPS data at all but there is data point in next interval time $P_{t2}$, the calculation speed on that road segment is as follows:
  - Get time and location of next data point ($P_{t2}$) after the road segment that doesn't have GPS data
  - Get time and location of data point ($P_{t0}$) before the road segment that doesn't have GPS data
  - Calculate the distance between point $P_{t0}$ and $P_{t2}$ and divide it with the delta time between point $P_{t0}$ and $P_{t2}$. The result of this calculation become the traffic speed on road segment $R_{t1}$ that doesn't have point GPS data.
- If on the road segment $R_{t1}$ there is only one point GPS data, the calculation speed on that road segment is as follows:
  - Get data point ($P_{t1}$) that lie on the road segment that only have one point GPS data
  - Get data point ($P_{t0}$) before the road segment that only have one point GPS data
  - Calculate the distance between point $P_{t1}$ and $P_{t0}$ and divide it with the delta time between point $P_{t1}$ and $P_{t0}$. The result of this calculation become the traffic speed on road segment $R_{t1}$ that only have one point GPS data

### G. *Speed Prediction*

The next step after calculation of the traffic speed are built regression model from traffic speed history then predict the value of traffic speed in the next hour for each of road segment. To meet this goal, data time series of traffic speed from taxi trip history must be built. We need to aggregate the traffic speed data to build time series data. If we want to predict traffic speed data in every next hour, then we need to aggregate data up to an hour.

The method used for prediction in this study is SVR. In order to predict traffic speed, the regression model of the data history needs to be established. Data are divided into two kind, training and testing. Each of this kind of data were divided into two category, first category from weekday traffic, and second category from weekend traffic. The regression model from both weekday and weekend traffic of data training were built using linear and RBF kernel. Then, we fit the data testing into regression model and calculate the Mean Squared Error (MSE) and Mean Absolute Error (MAE).

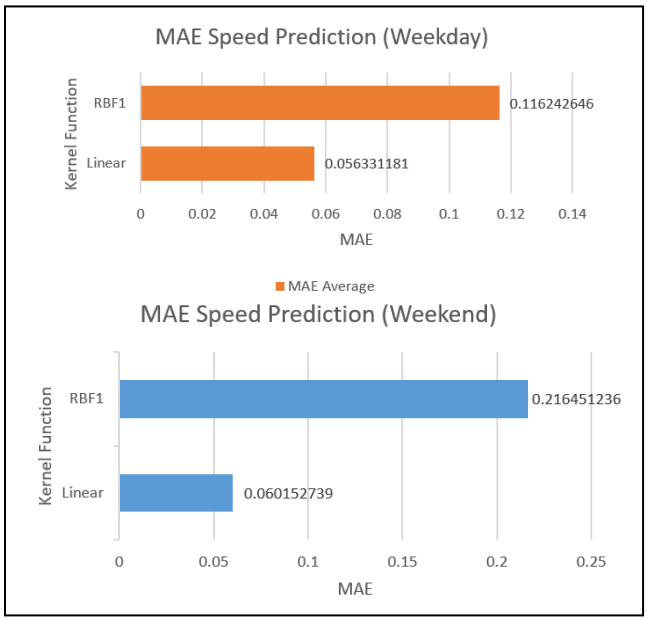


Fig. 9. MAE result from speed prediction

Figure 9 and figure 10 shows the result of speed prediction using SVR with linear and RBF kernel function. Both MAE and MSE value were below 1. It indicates that SVR has good performance to predict traffic speed.

## V. CONCLUSION & FUTURE WORK

From experiment, the map matching method will refined the GPS data so the calculation of traffic speed data will be more precise. From experiments result, SVR has good performance to predict traffic speed data which has MAE and MSE value not more than 1.

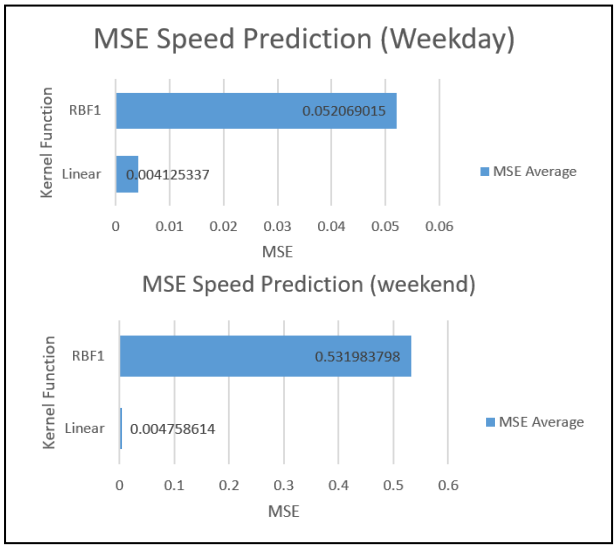


Fig. 10. MSE result from speed predicition

This paper is framework or guide how to predict traffic speed using data GPS taxi by utilizing Support Vector Regression method. Furthermore, we need test this proposed architecture so we can see if aggregation of speed data into specific interval time will impact the accuracy of prediction or divide the data into peak time and non-peak time to improve the accuracy of prediction. To get the best result of SVR, we need to find the best parameter of kernel function, so test the data into several parameter option of kernel function in SVR method also could be future work.

## REFERENCES

[1] C. E. White, D. Bernstein, dan A. L. Kornhauser, "Some map matching algorithms for personal navigation assistants," vol. 8, 2000.

[2] M. T. Asif *dkk.*, "Spatiotemporal Patterns in Large-Scale Traffic Speed Prediction," vol. 15, no. 2, hal. 794–804, 2014.

[3] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, dan B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *FGCS*, vol. 61, hal. 97–107, 2015.

[4] Q. Yang, J. Wang, X. Song,w33 X. Kong, Z. Xu, dan B. Zhang, "Urban Traffic Congestion Prediction Using Floating Car Trajectory Data," *Proc. 11th Int. Conf. Algorithms Archit. Parallel Process.*, vol. 9529, hal. 18–30, 2015.

[5] J. Lu dan L. Cao, "Congestion Evaluation from Traffic Flow Information based on Fuzzy Logic," hal. 50–53, 2003.

[6] J. Yoon, B. Noble, dan M. Liu, "Surface Street Traffic Estimation," *MobiSys 2007*, hal. 220–232, 2007.

[7] Y. C. Zhang, X. Q. Zuo, L. T. Zhang, dan Z. T. Chen, "Traffic congestion detection based on GPS floating-car data," *Procedia Eng.*, vol. 15, hal. 5541–5546, 2011.

[8] K. Muller, A. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, dan V. Vapnik, "Predicting time series with support vector machines," *Artif. Neural Networks—ICANN'97*, vol. 1327, no. x, hal. 999–1004, 1997.