Optimization Big Data Real-time Analytics Using Mobile Phone Data in Origin Destination National Transportation (*ATTN*) Survey

Okkie Putriani, and Sigit Priyanto

Abstract— **The ATTN 2018 data collection process is obtained from the data collection of the sample OD-matrix using cellular data, carried out with the aim of obtaining data on the Origin Destination Matrix (movement) of the mobile phone user movement for a given period to get the sample OD-matrix. Data signals from cellular networks can be a means of analysing transportation systems to help formulate transportation models to predict future users. FCD (floating car/cellular data) is based on the collection of localization data, speed, direction of travel information and time from the cellular telephone on the vehicle. The initial challenge in combining the algorithms is: how to scale the algorithm to a big way, how to cleanse the data, how to use both manual queue detection, and how to report it in time when the queue occurs. The approach of these research is to obtain a combination real-time analysis model of movement distribution and simultaneous mode selection.**

***Index Terms*— ATTN-people, big data, real-time analysis, smart phone.**

I. ATTN-People 2018

Transportation has an important role in supporting economic, socio-cultural, political, and defences security development. The accuracy transportation policies and planning are needed so that transportation management is more effective and efficient supported by valid and updated data. One of the data needed is data on the movement of the origin of the destination. Data from a complete transportation destination from a region, namely in the form of a matrix from the destination of transportation, are the main prerequisites so that a transportation policy can be formulated properly. The effort to get the desired movement performance will be easily realized. Data from the origin of the movement objectives are periodically carried out through a national transportation survey that collects data from the destination of the movement of people (*ATTN* Survey). The *ATTN* 2018 data collection process is obtained from the data collection of the sample OD-matrix using cellular data, carried out with the aim of obtaining data on the Origin Destination Matrix (movement) of the mobile phone user movement for a given period to get the sample OD-matrix. Collecting data on the traffic flow of people on the road transportation network is by conducting a survey, carried out with the aim of getting traffic flow data of people on the intermodal transportation network. Collecting data on the network at the transportation node is carried out by the aim of getting data on the daily movement of traffic on the transportation node.

1.1 Roles of ATTN 2018 Data

Data from the destination of transportation describes the demand for movement of people. The data is used as an analytical material to formulate transportation policies in an effort to facilitate the demand for movement of people. This policy is used to improve transportation equipment as a form of intervention from the supply side. Data Origin Destination of Transportation People are needed by various government and non-government parties for the analysis process to formulate transportation policy and the basis of analysis predicts transportation performance (transportation costs, congestion, financial feasibility, economic feasibility) such as: infrastructure investment (Mass Rapid Transit (MRT) development, airport , urban double track tracks, transportation infrastructure investments (increasing mass public transport facilities to reduce congestion), transportation operations (increasing frequency of commuter rail transport in Java), and regulations (*Transjabodetabek* development). Transportation analysis is used by the Ministry of Public Works and Public Housing (PUPR) to formulate policies for the development of the road network, the development, the management of road networks, and others.

1.1.1 Output Survey ATTN – people 2018

The output is generated by national and regional movement patterns of people in the Origin Destination Matrix (OD-matrix) format and the person movement database system and it is used for transportation policies at the National and Regional levels.

1.1.2 ATTN– people Survey Implementation

The ATTN survey includes the continuous stages of socialization, preparation of resources (organizational and institutional), conducting surveys, data processing, monitoring.

O. Putriani is with the Department of Civil Engineering, Atma Jaya Yogyakarta University, Yogyakarta, Indonesia (e-mail: okkie.putriani@uajy.ac.id)

S. Priyanto is with the Department of Civil Engineering, Gadjah Mada University, Yogyakarta, Indonesia.

1.1.3 ATTN – people Survey Methodology

The ATTN data collection process obtained from the data collection of the sample OD-matrix using cellular data is carried out by the aim to obtain the data of the Origin Destination Matrix (movement) of the mobile phone users movement for a certain period to get the sample OD-matrix. Collecting data on the traffic flow of people on the road transportation network, by conducting a survey, is synergized by the aim to obtain data on the flow of traffic of people on the intermodal transportation network. The collection of data on the network at the transportation node is to obtain data on the daily traffic flow of people on the transportation node.

1.2 Sample Data Collection Using Cellular Data

Basically to form data from the origin of the people movement, the information needed is a series of movements of people from one place to another in a period of time. In simple terms, the information question is the location of people from time to time. In this case the position identification is very dependent on the zone that will be created or highly dependent on zone granulation.

If the zone system created is provincial based, then the location will inform from the province. Whereas the zone system is a district/city, then the location will identify the position of the district or city, and if the administrative location, the location needed is district/city and province. Especially for the formation of this National OD-matrix, the lowest administrative location required is district/city, but for further development and needs analysis, administrative locations at the village and sub-district levels should be included.

Administrative location: numeric (code) or string (description), the representation of the type above is not a necessity, because all data fields can be represented in a string form, and the needed for processing is begun. In the formation of the National OD-matrix, the OD-matrix that will be formed consists of 3 (three) granularity zones, namely OD-matrix between provinces, OD-matrix between certain islands/regions, OD-matrix between districts/cities in certain regions. In each matrix formed, an external zone will be defined, namely the zone outside the observed area. In general, the external zone will be divided into four parts, north, south, east and west. The determination of this zone is determined relative to the area being observed, and this can be obtained from the location of longitude and latitude of origin and destination. The concept of cellular data processing to produce mobile phone user movement patterns with LBA (Location Based Advertising) data developed into algorithms with two stages, namely the stages of cellular data sorting and the stages of forming people's movement patterns.

1.2.1 Collecting People Flow on Transportation Network

The survey is collecting the people flow simultaneously, at the same time span as cellular data-based OD-matrix data collection. The data flow unit required is person/hour; person/day. Traffic flow surveys are carried out by segments on a predetermined transportation network. The transportation network segment is a link survey in the road network system, data on traffic volume can be obtained by field survey using the traffic count method. The scope of the activity includes the recruitment of surveyor personnel, training of surveyor personnel, bring a survey of traffic flow at the specified point, especially at the point of the road, occupancy rate survey is needed. In accordance to the objectives of the roadside traffic volume survey, which is to determine the volume of movement of people on the link in the main transportation network of the area being studied, the traffic survey location is placed on the national road segment if used to determine the national OD-matrix scaling factor, and on provincial roads if the determine regional OD-matrix scaling factors is being used.

1.2.2 Collecting Data on The Network in The Transportation Node

Daily data on the OD-matrix data collection time range is based on cellular data and data collection of traffic flow of people on the transportation network. The link in the railway network system, the traffic volume data on the link in question can be obtained by collecting data from the train manifest during the specified period, the other source is the station manager. The link in the air freight network system, traffic volume data is obtained by collecting data from the flight manifest of the planes that use the track during the specified period, other sources are airport managers. The link in the naval network system, the traffic volume data on the link can be obtained by collecting data from the ship liners or from the port manager. The link in the crossed river and lake (*ASDP*) trajectories of traffic volume data can be obtained by crossing traffic managers.

1.2.3 The Sample Needs and The Implementation of ATTN-people Survey

Sample requirements include the location of traffic counting, nodes, and equipment needed. Based on the location of the traffic counting survey planned as many as 412 observation segments in 34 provinces in Indonesia. There are 97 observation segments in Sumatra Island, 152 observations in Java, 66 observation segments on Kalimantan Island, 31 observation segments in Nusa Tenggara and Bali, 43 observation segments on Sulawesi Island, 12 observation segments on Maluku Island, and 11 observation sections in Papua Island. The description of the number of current observations of the point of traffic counting Origin of 2018 National Transportation Purpose per Province can be seen in Table 1. Secondary data collection was carried out at the station node as many as 57 observation sections consisting of 51 regional trajectories and 6 KRL trajectories, the crossing port node as many as 16 observation sections consisting of 47 commercial lines and 116 pioneer trails, 118 harbor nodes of 377 observation sections, and 377 airport nodes.he ATTN survey includes the continuous stages of socialization, preparation of resources (organizational and institutional), conducting surveys, data processing, monitoring. Nonetheless, to ensure data accuracy, the Research and Development Agency continue to reduce the Monitoring Team to the field by the Traffic Counting Survey (TC) and Road Occupancy Interview (ROI).

TABLE I
SURVEY TRAFFIC COUNTING ATTN 2018 PER PROVINCE

No	Province	Total	No	Province	Total
1	NAD	9	18	KALTENG	18
2	SUMUT	10	19	KALTIM	14
3	SUMBAR	10	20	KALTARA	4
4	RIAU	10	21	KALSEL	15
5	KEPRI	9	22	BALI	8
6	JAMBI	10	23	NTB	13
7	BENGKULU	10	24	NTT	10
8	SUMSEL	11	25	SULUT	8
9	BABEL	7	26	GORONTALO	8
10	LAMPUNG	11	27	SULTENG	7
11	DKI	26	28	SULBAR	5
12	BANTEN	15	29	SULSEL	10
13	JABAR	29	30	SULTRA	5
14	JATENG	24	31	MALUKU	6
15	DIY	16	32	MALUKU UTARA	6
16	KALTIM	42	33	PAPUA	7
17	KALBAR	15	34	PAPUA BARAT	4

Source: *Kemenhub Balitbanghub 2018.*

II. BIG DATA TRANSPORTATION

Digital signals are captured by accurate conditions in progress, both from driver coordinates, telephone bills, message status on social media, until spatial location recordings can be gathered into data bases for transportation research [1]. Big Data has four (4) aspects: volume (quantity of data), velocity (speed), variety (data variation), veracity (truth, so that the correlation in processing to decision making [2]. Social media and social networking platforms like Facebook, Twitter, Instagram, Google Map, Waze, Gojek provide opportunities for people to share ideas, emotions and information publicly or per community, with a huge volume of social signals with real time accuracy as the Big Data concept. Real-time traffic information is very important to support ITS (Intelligent Transportation Systems) applications: accident detection, vehicle navigation, traffic signal control, traffic monitoring and others [3]. It is shown in Fig. 1.

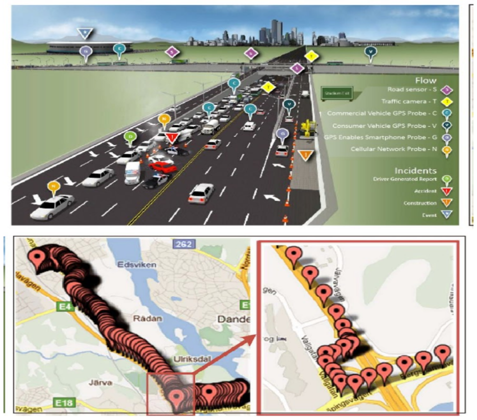

Fig. 1. Volume Traffic Transportation Using Real Time Mobile Data [4] and Example of Vehicle Trajectories Collected by A-GPS Phones [3]

One of the social transportation studies about traffic analysis planning. Social transportation focuses on five areas: (1) transport analysis with big data using data mining, machine learning, and process language programming methods, (2) crowdsourcing mechanisms for social media-based transportation, Internet of Things (IoT), Internet of Everything (IoE), Vehicle of Everything (V2X). Vehicle to Everything (V2X) in an automotive company refers to technology around the Internet of Things from Vehicle to Vehicle (V2V), Vehicle to Infrastructure (V2I), Vehicle to Device (V2D), Vehicle to Pedestrian (V2P), Vehicle to Home (V2H), Vehicle to Grid (V2G). This V2X concept is prioritized on driver safety, mobility and environment oriented to the challenges of Intelligent Transport System (ITS), (3) the next stage of service from Location-Based Services (LBS), (4) Web-based technology that regulates transportation, (5) Application and further research development [5].

Data signals from cellular networks can be a means of analysing transportation systems to help formulate transportation models to predict future users. An approach based on this type of data is very interesting for transportation systems that require large-scale expansion, because it has additional benefits that do not require special equipment or installations, so they can be very cost efficient. In this study examines how the data obtained can be processed and used to act as an enabler for the transport analysis model. Modular layered which is integrated into the entire process and presents the results of the initial analysis of cell phone call data in the context of mobility, transportation, and transportation infrastructure [6].

The split model aims to determine the number of trips in different modes to request requests for trips between different zones and nodes. Systematically, the mode selection phase of the sequential demand analysis procedure will be described. Split models are an important step in estimating travel demand. Development of split mode models based on widely available mobile data and geographical data on transportation networks. Parts of all three transportations (cars, public transportation and pedestrians) that will describe the conditions and validation with existing data [7].

Traffic management is increasingly in need of fast access to data diversity in direct and immediate decision making. However, there are problems that hinder the rapid need to obtain and integrate spatial data through the web, namely the heterogeneity of the Geographic Information System (GIS) system and data sharing system, in this case focusing on Web Feature Service (WFS) and Web Map Service (WMS) [8]

III. FLOATING CAR DATA/ FLOATING CELLULAR DATA

The European Union 2010/40/EU defines ITS as a technology and communication system that is applied in the field of land transportation, including infrastructure, vehicles, users, traffic management, mobility management of other modes of transportation. ITS includes wireless communications, computer technology, floating car data

(FCD) or floating cellular data, sensing technologies. FCD is a method for determining the speed of traffic on a road network.

FCD is needed to collect transportation routes and in general there are 4 methods to get raw data: triangulation methods, vehicle re-identification, GPS-based methods, and cellular phone-based monitors [9]. This is based on the collection of localization data, speed, direction of travel information and time from the cellular telephone on the vehicle. Based on this data, each vehicle actively acts as a sensor for the road network. Traffic congestion can be identified, travel times can be calculated, and traffic reports can be made more quickly [10].

IV. Algorithm

According to Torp [11] to be able to detect traffic queues, there are two measuring stretch concepts and report stretch. The main idea of the traffic queue detection algorithm is to detect the traffic queue at stretching the report if there is a queue in one of the measurement ranges on stretching the report. The initial challenge in designing this algorithm is: how to scale the algorithm to a big way, how to use both manual queue detection, and how to report it in time when the queue occurs. The algorithms used for transportation mode detection can be categorized as discriminative or generative, it was shown in Table II [12]. Generative algorithms model class-conditional probability density functions and prior probabilities. Discriminative algorithms do note attempt to model underlying probability distribution.

TABLE II
Methods Used For Transportation Mode Detection

No	Methods Used For Transportation Mode Detection	
1	Generative	Discriminative
2	Naïve Bayes	Support Vector Machines
3	Bayesian Networks	Neural Networks
4	Mixture Models	Nearest Neighbor
5	Hidden Markov Models	Decision Tree
6		Random Forests
7		Clustering

Source: M. Nikolic & M. Bierlaire, 2017

The transportation mode detection approaches based on smartphone data considered in the study differ in the type and the number of used input data, the considered transportation mode categories, and the algorithm used for the classification task, which affects their prediction capabilities.

V. Machine Learning

Commonly machine learning application can be divided into two main classes: tasks whose main goal is to find complex patterns in large amounts of high-dimensional data and identifying object in images [13]. The relevant information can be broken down as metadata (camera location/ID, timestamps, tracking frame-by-frame timestamps, frame statistics) and extracted features (classes of objects, positions of objects, direction of objects, and counts of objects). The analytical approach can be broken down in three categories: data extraction and descriptive analysis, computer vision tasks, pipeline development. The study held by Pulse Lab Jakarta used the CCTV (Fig.2) and computerized by calculation vision task using python and YOLO3 deployment method.

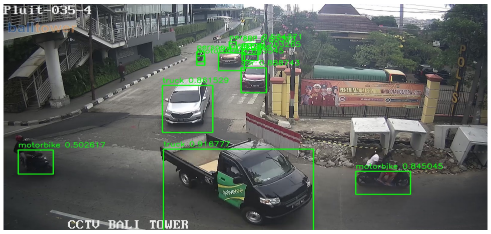

Fig.2. Object are detected and classified from raw CCTV footage (Source: PulseLab Jakarta, 2018).

VI. Multi-modal Transportation Routing Planning Models

Optimization models are formulated to describe the multi-modal transportation routing planning problem mathematically. By inputting the practical transportation data into the optimization models and then solving them by exact solution methods and branch-and-bound method or approximate solution methods, optimal solution can be attained [14]. The distinguish of the formulation characteristics shown in Fig. 3.

VII. Web-Mapping

The direction for the Web Mapping function is generated from big analytic data, modeling results and combined with other knowledge. The available functionality needs to be expanded to get more data. The challenge is to determine the high level of functionality needed in the application and end-user context [15]. The rapid development of cellular telephone technology, the method of extracting traffic data through cell phones began to play an important role in urban transportation planning and network analysis. This study investigated cellular-based methods to extract travel distribution data for urban transportation planning [16].

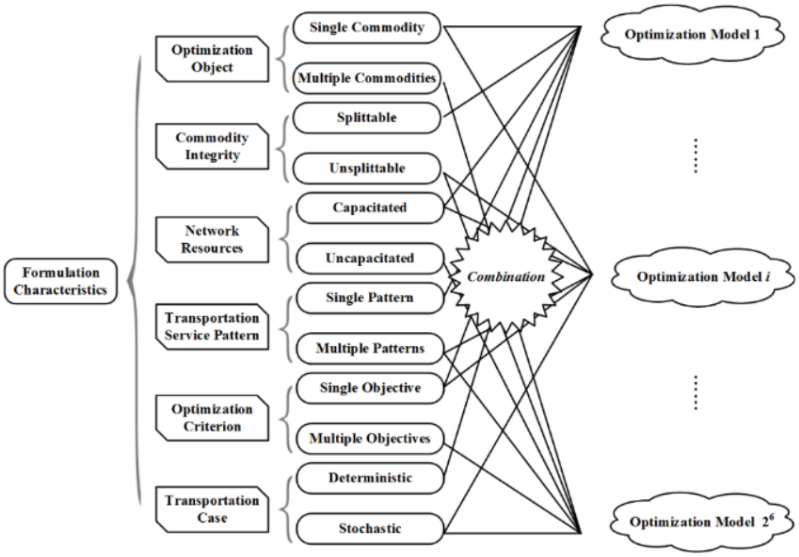

Fig. 3. Formulation Characteristics of The Optimization Models (source: Y. Sun, M. Lang, D. Wang, 2015)

Large and diverse group participation in process planning has long been encouraged to increase the effectiveness and acceptance of plans. The challenge of stakeholder participation is very important in every transportation planning process. Crowdsourcing is used to collect data from various stakeholders in transportation projects both data collection and feedback on the quality of public transport services and real-time information quality [18].

VIII. Methodology

The approach of this research is aimed to obtain a combination real-time analysis model of movement distribution and simultaneous mode selection. The intention of this research is to introduce the benefit of big data using in transportation, to find appropriate methodology that big data can be used effectively and easily to users, and to trigger input and feedback to tackle the usefulness of big data.

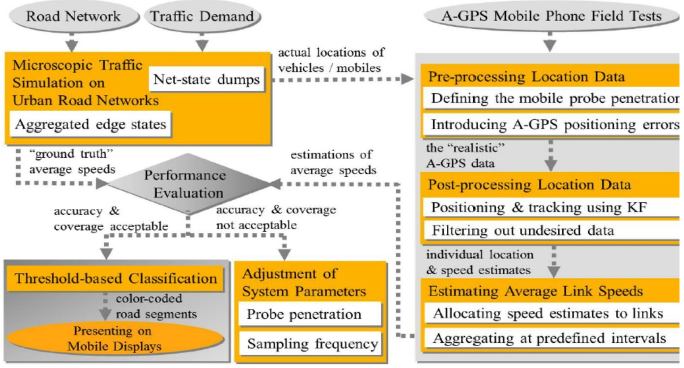
Road Network
Traffic Demand
A-GPS Mobile Phone Field Tests
Microscopic Traffic Simulation on Urban Road Networks
Net-state dumps
Aggregated edge states
actual locations of vehicles / mobiles
Pre-processing Location Data
Defining the mobile probe penetration
Introducing A-GPS positioning errors
the "realistic" A-GPS data
"ground truth" average speeds
Performance Evaluation
estimations of average speeds
Post-processing Location Data
Positioning & tracking using KF
Filtering out undesired data
accuracy & coverage acceptable
accuracy & coverage not acceptable
individual location & speed estimates
Threshold-based Classification
color-coded road segments
Presenting on Mobile Displays
Adjustment of System Parameters
Probe penetration
Sampling frequency
Estimating Average Link Speeds
Allocating speed estimates to links
Aggregating at predefined intervals

Fig. 4. Simulation-based Framework [3]

The data that will be obtained from cellular phone data, in this case is the main choice that will later be verified by the field.

The flow of this research includes three main stages: imputation and path-matching trip from cellular phone data, matrix projection based on survey data commuting trip, and matrix correction from a combination of modelling road flows and modelled traffic.

Data needed with multiple real-time main entities: time of each cellular ID, location of each cellular ID, and data traffic on the study area road network as a calibration process and validation. The development of algorithms forms the flow or volume of traffic on the road network. Validation on certain days with the same time surveyed the flow of vehicles on the road network.

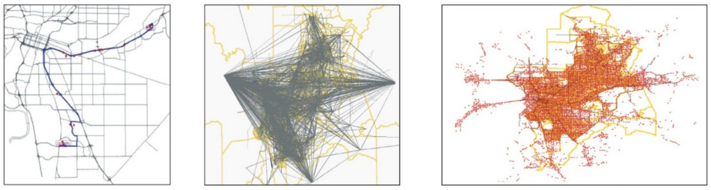

Fig. 5. The choice of observation path from the groove and output model, the blue line indicates the shortest path from the free flow travel time combination of the highway and arteries. OD flow collected from vehicle trips identified on cellular phones for 1 hour. Display mobile phones from 400-600 customers randomly from the entire dataset.imulation-based Framework [18].

IX. Discussion

In optimization big data real-time analytics using mobile phone data in Origin Destination National Transportation (ATTN) survey, we need to propose and compare combinations of several methods for classifying transportation activity data from smartphone [19]. The two main objectives are to classify the data as accurately as possible and to reducing the dimensionality of the data as much as possible in order to reduce the computational burden of the classification. Smartphones generates a constant stream of data describing the phone's acceleration and location. Unfortunately, based on *Databoks* statistics and data portal January 2017, cell phone users in Indonesia reached 371,4 million users or 142 percent of the total population 262 million, in Fig. 6. It showed every population used 1,4 cellular phone, the same meaning that one person is using more than one cell-phone, even more 2-3 cell phone cards. While Indonesia's urban population reaches 55% of the total population. These approach has become the opportunity for lack of accuracy of the survey using cell-phone data.

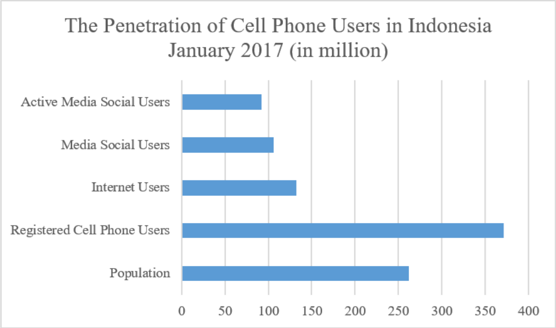

Fig. 6. The Penetration of Cell Phone Users in Indonesia January 2017 (source: databoks.co.id)

The interesting point of view based on the previous research was the gap studies between the transportation theory and the big data analysis. Combination of the algorithm to classify five different modes of transportation (i.e. walking, biking, car, bus, and rail) while being computationally simple enough to run a typical smartphone [19]. In particular, Bayesian network as the algorithms method had a better performance than other algorithms in terms of the percentage indenfitified of detection transportation mode [20]. The effort to control the high cost, efficiency and safety issues for calculating household trip daya survey in Indonesia are emerging, based on the ATTN 2018 that nowadays is held. The transportation in Indonesia, a vast polygot nation, as the developing countries is different from developed countries like Singapore has a highly developed and successful transportation management. The crouwded traffic condition make the data mining is more challenging.

X. Conclusion and Future Work

The deeper researches are needed for classify the best algorithm using data mining and machine learning can be used for optimizing big data real-time analytics using mobile phone data in Origin Destination National Transportation (ATTN) survey. The survey in Indonesia needs a massive attention on the cleansing data based on the massive and mixed users in the transportation, including the double cell phone in one users.

References

[1] W. Zheng, W. Chen, D. Shen, S. Shen, X. Wang, and L. Zang, "Big Data for Social Transportation," *IEEE Intelligent Transportation System.* 17, no. 4, 620-630, 2016.

[2] G. Elena, C. Florina, A. Anca, and V. Manole, "Perspectives on Big Digital and Big Data Analysis," *Data Base System Journa,*3, no. 4, 3-14, 2012.

[3] S. Tao, V. Manolopoulos, S. Eodriguez, and A. Rusu, "Real-Time Urban Traffic State Estimation with A-GPS Mobile Phones as Probes," *Journal of Transportation Tech*., 2, 22-31, 2012.

[4] D. Harris, "Why Better Traffic Data Means More Than Just A Faster Commute," 2012.

[5] I. Ivan, L. Sang-Woo, W. Tim, and M. Carsten, "Cyber Security Standards and Issues in V2X Communication," *International Journal Of Electrical, Electronics and Data Comm*., 5, issue-4, 2017.

[6] V. Angelakis, D. Gundlegard, B. Rajna, C. Rydergen, K. Vrotsou, R. Carlsson, J. Forgeat, T. Hu, E. Liu, S. Moritz, S. Zhao, and Y. Zheng, "Mobility Modelling for Transport Efficiency – Analysis of Travel Characteristics Based on Mobile Phone Data," In *Netmob 2013-Third International Conference on the Analysis of Mobile Phone Datasets, May 1-3, 2013, MIT, Cambridge, MA, USA*. 2013.

[7] F. Wang, and C. Chen, "On Data Processing Required To Derive Mobility Patterns From Passively-Generated Mobile Phone Data," *Transportation Research Part C: Emerging Tech*., 87, 58, 2018.

[8] C. Zhang, W. L, "The Roles of Web Feature and Web Map Services in Real-time Geospatial Data Sharing for Time-critical Applications," *Cartography and Geographic Information Science*, 32, no. 4, 269-283, 2005.

[9] M. Treiber, and A. Kesting, "Traffic Flow Dynamics," Springer-Verlag, Berlin, 2013.

[10] N. Chon, and H. Biscoff, "Explorative Study to Technique and Application and Sample Properties of GDP Data," Floating Car Data For Transport. Planning, TomTom, Technische Universiteit Einhoven, Univeristy of Technology, 2012.

[11] K. Torp, and H.S. Lahrmann, "Floating Car Data For Traffic Monitoring," In *5th European congress and exhibition of intelligent transport systems and services, Hannover, Germany*, 2005.

[12] M. Nikolic, and M. Bierlaire, "Review of Transportation Mode Detection Approaches Based on Smartphone Data," *17th Swiss Transport Research Conference*, 2017.

[13] K. Durpe, J. Walsh, A. Fout, J. Caldeira, A. Kesari, and R. Sefala, "Smart City Traffic Safety Technical Report," Pulse Lab Jakarta, Jakarta, 2018.

[14] Y. Sun, M. Lang, and D. Wang, "Optimization Models and Solution Algorithms for Freight Routing Planning Problem in the Multi-Modal Transportation Networks: A Review of the State-of-the-Art," *The Open Civil Engineering Journal,* 9, 714-723. 2015.

[15] B. Veenendaal, M.A. Brovelli, and S. Li, "Review of Web Mappings: Eras, Trends, and Directions," *Int. Journal of Geo-Information*, 6, 317, 2017.

[16] C. Pan, J. Lu, S. Di, and B. Ran, "Cellular-based data-extracting method for trip distribution," *Transportation Research Record: Journal of The Transportation Research Board*, 1945, 33-39, 2006.

[17] O.Z. Tamin, "Perencanaan, Permodelan, dan Rekayasa Transportasi", ITB, Bandung, 390-391, 2008.

[18] A. Misra, A. Gooze, K. Watkins, M. Asad, and C.A.L. Dantec, "Crowdsourcing and Its Application to Transportation Data Collection and Management," *Transport. Research Record: Journal of The Transport. Research Board*, 2414, 1-8, 2014.

[19] B. D. Martin, V. Addona, J. Wolfson, G. Adomavicius, and Y. Fan, "Methods for Real-Time Prediction of the Mode of Travel Using Smartphone-Based GPS and Accelerometer Data," *MDPI Sensors 2017,* 17, 2058, 2017.

[20] T. Feng, and H. J. P. Timmermans, "Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data," *Routledge Taylor & Francis Group, Transportation Planning and Technology*, 39, No. 2, 180-194, 2016.