Implementation of Big Data to Develop Origin-Destination Matrix Estimation Model

Ofyar Z. Tamin¹, Heriansyah², Siti Raudhatul Fadilah³

¹ Department of Civil Engineering, Institut Teknologi Sumatera, Lampung, Indonesia
² Department of Electrical Engineering, Institut Teknologi Sumatera, Lampung, Indonesia
³ Department of Civil Engineering, Institut Teknologi Bandung, Bandung, Indonesia

sitiraudhatulfadilah@gmail.com

Abstract. Big data is a huge collection of electronic data which is currently popular to be developed in various scientific fields, including the transportation sector. Several data categories that can be used for transportation infrastructure planning are real-time location/position and timestamp of road users. This technological sophistication addresses a very crucial transportation planning problem that has been faced for many years, that is the inaccurate and unrepresentative prediction of the Origin-Destination Matrix. Up to this point, the Origin-Destination Matrix has been obtained using conventional and unconventional methods. Data was collected through manual traffic counts survey and direct interview in the study area. This results in unreliable data and inadequate sampling rates, which are only able to cover 20% of the population. In addition, the application of these techniques is expensive, time-consuming, and require a lot of human resources. Therefore, this study tries to develop a model capable of predicting the origin and destination of mobility, which in this case is represented by the sub-district boundary as a zone. As a result, an Origin-Destination Matrix can be obtained with a high degree of accuracy. The case study used is Bandung city. In the future, this research is expected to become a milestone in elaborating a validated Origin-Destination Matrix estimation model that can be used in various areas.

Keywords : big data, origin-destination matrix, accuracy

1. Introduction

Transportation planning aims to estimate the amount and location of transportation demand, for instance determining the total trips of vehicles, passengers, and/or goods, present and future. The results of the planning will be used for various transportation infrastructure investments. Therefore, information describing the mobility from the origin to the destination zone, which in this case is represented by an origin-destination matrix, is needed. An origin-destination matrix can provide a detailed indication of travel demand. It can be concluded that the origin-destination matrix has a very important role in various transportation planning and management studies.

The origin-destination matrix can be defined as a two-dimensional matrix that contains information on the number of trips between zones within a study area, including zones outside the study area as external zones. Two main methods have been developed for constructing an origin-destination matrix, namely conventional and unconventional [1-6].
The conventional methods rely on primary survey results to obtain the total trips from origin to destination zone. As a consequence, the application of this method will be highly expensive since it requires very large data collection activities. Besides, conventional methods take a very long data collection time, as well as data analysis, but produce a matrix with a low level of accuracy. For example, the Indonesian Ministry of Transportation is only able to carry out a national origin-destination survey every five years because of the enormous costs and long processing time. One of the reasons that the original destination matrix has a low level of accuracy is the limited sampling rate, which mostly only reaches 20% of the population. Similarly, incongruity number and location of samples taken, along with the high human error in data collection and processing, also lead to a reduction in the accuracy of the matrix.

Meanwhile, the unconventional method is an alternative that can complement the origin-destination matrix generated by conventional methods. This method is processed by utilizing traffic flow data. However, the limited amount of traffic flow data obtained and the lack of reliability of the data resulted in low accuracy of the origin-destination matrix.

Industrial revolution 4.0, which combines automation technology with cyber technology, is currently a big leap in the industrial sector where information and communication technology can be fully utilized to achieve the highest efficiency. In response, transportation planners are trying to take advantage of rapid technological improvements to identify travel patterns more accurately and comprehensively. One thing that could be developed is the estimation of the number of trips from one point to another within a particular study area, represented by an origin-destination matrix, using the coordinates of the location of each vehicle, human, and/or goods. This information can be obtained by utilizing available Android-based applications, such as Google Maps, considering that currently almost all residents use smartphones. Therefore, the location and timestamp of each road user can be acquired as a whole and in real-time. This is expected to be a solution to the sampling rate problem in data collection.

Furthermore, in order for data collection in transportation infrastructure planning to be carried out more effectively and efficiently, both in terms of time, cost, and resources, as well as to produce a reliable origin-destination matrix, this research will implement big data for the development of an estimation of the origin-destination matrix. The case study used is Bandung city, on a sub-district basis. In the end, it is expected that this research will become a milestone of future studies that utilize big data to identify travel patterns and generate an origin-destination matrix, as a substitute for conventional methods.

2. Literature Review

2.1 Origin-destination matrix

Mobility is a daily activity carried out for a variety of reasons and purposes. Although the need is very high, several problems, such as congestion and insufficient availability of infrastructure, still occur. Therefore, it is necessary to correctly identify current and future travel patterns, for example from where and where to go, at what time, and how much traffic flows, so that transportation investment policies can be utilized properly and as needed.

Travel patterns in the transportation system are frequently described in the form of traffic flow (vehicles, passengers, and goods) moving from the origin zone to the destination zone within a certain area and over a certain time period. To describe these patterns, transportation planners often use the term origin-destination matrix. Travel patterns can be generated if an origin-destination matrix is assigned to a transportation network system. By studying this, transportation problems that arise can be immediately identified.

By definition, the origin-destination matrix is a two-dimensional matrix containing mobility information between zones that have been defined within a study area. The information is arranged in a table which is shown in the Table 1 below. Rows in the table represent the zone of origin, and the column represents the zone of the destination. So, the matrix cell represents the amount of flow from origin to destination. In this case, the $T_{id}$ notation states the amount of traffic flow (vehicles, passengers, or goods) moving from origin zone $i$ to destination zone $d$ during a certain time interval.
Two main methods have been developed to generate an origin-destination matrix: conventional and unconventional [1-6]. The generation of the origin-destination matrix using the direct method is highly dependent on the results of data collection and primary surveys. There are several obstacles in conducting the survey, for example, the interview process that could disrupt road users and cause traffic delays, also insufficient sample size due to time and cost constraints. This resulted in errors in the data collection process. In addition, the choice of survey method depends on the availability of surveyors. Therefore, technical and human factors often arise under conventional direct methods, such as errors in recording and interpreting data.

Based on the description above, the conventional "direct" method produces an origin-destination matrix with quite high errors. There are several types of errors which are described below.

1. Errors in daily / seasonal variability and expansion of survey data will occur when a factor is required to convert the original existing data to produce an origin-destination matrix at certain time intervals.
2. Errors in data collection occurred during the survey process, mostly due to human errors, such as: errors in identifying vehicles, calculating flows, incomplete questionnaires, or errors in writing information. These errors can be reduced by good quality control, but it is impossible to eliminate them completely.
3. Errors in data processing, come up in the process of moving raw data and also because of human errors. The main sources of these errors are coding, typing, double counting, missing or illegible data, editing and creating files, creating tables, or in programming.
4. Sampling error occurs because the survey cannot cover all vehicle trips during a certain time interval, except in very simple cases.

### 2.2 Big data

Big data is a term for very large and complex electronic data sets. The understanding of big data includes data sets with sizes beyond the ability of common software tools to capture, create, manage, and process data within tolerable elapsed time [8]. Basically, technological sophistication drives the availability of big data, which is changing very rapidly and consists of many types of data categories. Big data has become very popular after billions of people use the internet to fulfill various daily activities and needs. Lots of data is stored on computers and the internet today in the form of text, images, sound, video, animation, blogs, books, weather, position on earth, temperature, flights, and even markets. Apart from “super big” big data users, such as Facebook, Google, Twitter, Amazon, and Alibaba, there are many local Indonesian companies that involve big data in their business, such as Traveloka, Gojek, and Tokopedia. Besides, in the telecommunications sector, various operators also apply big data in their own operations, such as Telkomsel, Indosat, and XL Axiata.

Over the last few years, big data has succeeded in changing various systems and experiences in industry lines, including the transportation industry. Big data makes everyday trips easier, more accessible, and enjoyable with the application's features that show which route is the shortest and fastest way to reach the desired destination. The use of big data allows people to obtain a real-time location for every single vehicle, human, and/or item with a high degree of accuracy. This data can easily be accessed by utilizing cellular phone user data connected to the Base Transceiver Station (BTS). Position information is collected in a database with a very large capacity.
3. Methodology

Figure 1 below shows several different methods that can be used to estimate the origin-destination matrix. This study adds the use of big data as one of the conventional methods to determine the number of vehicle trips between zones, which previously only consisted of the direct and indirect methods.

In this study, the availability of big data, particularly location and timestamp data, will be studied and tried to be implemented in compiling the origin-destination matrix. Furthermore, a validated origin-destination matrix that represents vehicle trips in a given area will be very useful in transportation infrastructure planning. To achieve this goal, an estimation model is developed by combining real-time location data obtained from big data with the zoning system within a study area and existing transportation network.

Large scale data also requires a large program to be converted into useful information in order to achieve certain objectives. Big data cannot be processed only by conventional database programs called SQL (Structured Query Language) or RDBMS (Relational Database Management System). However, big data requires a database program that supports NoSQL (Not only SQL) and is capable of processing unstructured data.

![Figure 1. Origin-destination matrix estimation method](image)

The software used for distributed big data processing in this research is Apache Hadoop MapReduce 3.3.0. MapReduce is a programming model used for big data processing by utilizing a distribution system that is capable of executing parallel processes. Not only open source, but MapReduce is also quite tolerant of errors as well as it has a high level of reliability and scalability. The MapReduce software works in two phases, “Map” and “Reduce”. The process that occurs in the “Map” phase is splitting and mapping data. Meanwhile, the process carried out in the “Reduce” phase is shuffle and reduce data. Figure 2 shows the processes of each phase.
Figure 2. Map and reduce process [9]

In implementing MapReduce software, the call of each map function does not depend on other map functions. This allows MapReduce to perform divide and conquer working principles in parallel computing. It is therefore possible to duplicate the execution of reprocessing tasks in the event of failure. Nevertheless, this process will not affect the computational results. MapReduce divides the input data sets into independent pieces which are processed in parallel by “map” tasks. This application sorts the output from the map, which then becomes the input to reduce tasks. The input and output of these tasks are usually stored in system files. The application also takes care of scheduling, monitoring, and rerun failed tasks. Compute nodes and storage nodes in the Hadoop MapReduce application are able to be located on the same computer. Such a configuration allows the framework to schedule tasks effectively and efficiently on individual nodes that already contain data. Moreover, this configuration results in very high aggregate bandwidth across clusters.

Also, the MapReduce framework consists of one JobTracker Master and one TaskTracker Slave on each cluster node. JobTracker Master is responsible for scheduling component job tasks on the TaskTracker Slave, monitoring them, and rerun failed tasks. Whereas TaskTracker Slave performs tasks according to commands given by JobTracker Master. Overall, using MapReduce will shorten the data processing time by 30% of the normal time compared to other techniques [10].

4. Results and Discussion

Based on the methodology previously described, this study utilizes cellular data detected by Base Transceiver Station (BTS) data from 30 sub-districts in Bandung city, which is further referred to as zones, to develop an origin-destination matrix estimation model. The coordinates of each BTS are then inputted into the map of the Bandung City area, as shown in Figure 3 below. Hence, it helps to find out the location of the sub-district where the BTS stands.
Location information data for each road user is processed using Hadoop MapReduce. This aims to shorten processing time, but accurate, considering that the entire data is collected in a large capacity database. Then, the Hadoop MapReduce software is operated using a parallel system by utilizing several servers together to get optimal results.

Basically, MapReduce works in two phases, Map and Reduce. The Map phase consists of a process of splitting and data mapping, while the Reduce phase involves shuffling and reducing data. The process that occurs in the application of origin-destination matrix estimation using MapReduce can be seen in Figure 4.

The splitting process is done by dividing the input data containing the location of each cellular network user. This process is completed by three workers in the map phase so that the load is distributed and calculations can be done in parallel. Meanwhile, in the mapping process, the number of cellular network users who change locations within a given interval is calculated. As in the Map phase, the Reduce phase also uses three workers to process the 30 zones that are being analyzed in this study; each worker handles 12 to 13 districts. In this phase, the shuffling process is carried out to organize the data generated in the Map phase. This process is completed by adjusting the origin zone data into input data in the Reducing process, which is a process to accumulate all the data generated by the previous process to get the desired origin-destination matrix.
Furthermore, Map and Reduce process will be translated into a user interface menu that can calculate the origin-destination matrix by adjusting the analysis time and interval, as a result in Table 2.

Figure 4 MapReduce flow in the origin-destination matrix estimation model
Table 2. Origin-destination matrix of Bandung city

| Origin | Destination | 0 | 1 | 2 | 3 | 4 | 5 |
|--------|-------------|---|---|---|---|---|---|
| 0      | 0           | 1 | 2 | 3 | 4 | 5 |
| 1      | 1           | 1 | 2 | 3 | 4 | 5 |
| 2      | 2           | 1 | 2 | 3 | 4 | 5 |
| 3      | 3           | 1 | 2 | 3 | 4 | 5 |
| 4      | 4           | 1 | 2 | 3 | 4 | 5 |
| 5      | 5           | 1 | 2 | 3 | 4 | 5 |
5. Conclusion
The initial stage of developing an origin-destination matrix for Bandung city has been carried out in this research by utilizing the data of the cellular telephone users’ location connected to the Base Transceiver Station. Based on these data, the trips taken by those users from one zone to another in real time can be determined by using the Map-Reduce method. In the end, it is expected that this estimation model can produce an origin-destination matrix with a high level of accuracy so that it can be useful in transportation infrastructure policy planning.

6. Acknowledgements
This research was supported by Research, Community Service, and Innovation Program at Institut Teknologi Bandung.

References
[1] Tamin, O. Z. 1985. *Estimation of Matrices for Freight Movement from Traffic Counts Using a Non-Linear Regression Approach*. MSc Thesis of the University of London, Imperial College and University College London.
[2] Tamin, O. Z. 1986. *The Estimation of a Transport Demand Model from Traffic Counts and Simple Zonal Planning Data*. Report of Transport Studies Group, University College London.
[3] Tamin, O. Z. 1988a. *The Estimation of Transport Demand Models from Traffic Counts*. PhD Dissertation of the University of London, University College London.
[4] Tamin, O. Z. 1988b. *Transport Demand Model from Traffic Counts*. Proceedings of the 20th Annual Conference of Universities Transport Studies Group, London, UK.
[5] Tamin, O. Z. 1988c. *Simplified Transport Models Based on Traffic Counts, Workshop in Transport Models*. University of Napoli, Napoli, Italy.
[6] Tamin, O. Z. 1988d. *Metoda Estimasi Matriks Asal-Tujuan (MAT) dengan Menggunakan Data Arus Lalu lintas*, Seminar Fakultas Teknik Sipil dan Perencanaan, Institut Teknologi Bandung, Bandung.
[7] Tamin, O. Z. 2008. *Perencanaan, Pemodelan, dan Rekayasa Transportasi: Teori, Soal, dan Contoh Aplikasi*. Penerbit ITB, Institut Teknologi Bandung, Bandung.
[8] Snijders, C., Matzat, U., and Reips, U. 2012. “Big Data”: Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science, 7 (1), 1-5.
[9] Luo M., Liu, G. 2010. *Distributed Log Information Processing with Map-Reduce: A Case Study from Raw Data to Final Models*. IEEE International Conference on Information Theory and Information Security. Available at: http://dx.doi.org/10.1109/icitis.2010.5689760
[10] Dayalan, M. 2018. *MapReduce: Simplified Data Processing on Large Cluster*. International Journal of Research and Engineering, 5 (5), pp.399–403. Available at: http://dx.doi.org/10.21276/ije.2018.5.5.4.