Inferring Commuting Statistics in Greater Jakarta from Social Media Locational Information from Mobile Devices

Imaduddin Amin, Ni Luh Putu Satyaning P.P, Yulistina Riyadi, and Jong Gun Lee

Pulse Lab Jakarta - United Nations Global Pulse

Email: {imaduddin.amin, ni.paramita, yulistina.riyadi, jonggun.lee}@un.or.id

Setia Pramana and Robert Kurniawan

Institute of Statistics, Jakarta, Indonesia

Email: {setia.pramana, robert}@stis.ac.id

ABSTRACT

Jakarta is the biggest city in Indonesia with a population of more than 10 million and a megacity with 1.38 million people commuting around Greater Jakarta [1]. In order for the Government of Indonesia to understand the commuting behaviors of citizens and better plan the transportation system in Greater Jakarta, the Indonesian Central Bureau of Statistics first conducted its commuting survey in 2014. In this paper, we produce commuting statistics from locational information on social media and show that the information from mobile phone apps is a promising source of data. It implies that the commuting behaviors from big data can allow the public sector more frequent statistics.

1 INTRODUCTION

It is known that about 1.38 million people commute daily in Greater Jakarta or Jabodetabek, which includes Jakarta, Bogor, Depok, Tangerang and Bekasi. The scale of the commuting population along with the population of Jakarta (about 10 million) has caused Greater Jakarta to face many urban challenges, including transportation. This led the Government of Indonesia to conduct the first commuting survey by the Indonesian Bureau of Statistics (Badan Pusat Statistik or BPS) in 2014.

Among many development issues, transportation is one of the key issues that directly affect citizens' daily lives. In many cases, a survey-based data collection method is widely used but one of its inherent drawbacks is that such an approach takes some time from the design of a survey to the release of its final result. For instance, the first commuting survey in Greater Jakarta took longer than one year. Consequently, there are many attempts to use other types of data to produce similar statistics using different kinds of geolocated information such as GPS devices, sensors, social media and mobile phone data, in a shorter cycle [2, 3, 4, 5, 7, 8, 9].

In Indonesia, social media is recognised as a promising data source to understand the behavioral patterns of people, specifically as about 88.1 million of people use the Internet and among them about 79 million of people use the Internet for accessing social media[1]. Jakarta is often named as the Twitter capital of the world with 10 million tweets every day. As another example, in Greater Jakarta, between January and May 2014 for five months, about 40 million tweets were posted with GPS information.

In this paper, we test how the locational information from social media on mobile devices can reveal commuting patterns in Greater Jakarta. First we produce Origin - Destination statistics given 10 cities in Greater Jakarta from the entire GPS-stamped tweets posted by mobile devices, by identifying a set of people who commute in these areas. Second, we calibrate the initial result based on the population distribution and Twitter user distribution. Finally we verify the result with the official commuting statistics. Our result confirms that geo-located tweets has the potential to to complement official commuting statistics in order to fill information gaps.

2 DATA SET

2.1 Official Commuting Statistics

The Indonesian Central Bureau Statistics (BPS) conducted the first commuting survey in household level covering 13,120 households from 1,312 census blocks. This survey aims to understand the behavior of commuting users in 13 cities or regencies in Greater Jakarta which includes five Jakarta cities (Central, South, North, East, and West Jakarta), five satellite cities (Bogor, Bekasi, Depok, Tangerang, and South Tangerang), and three regencies (Bogor, Bekasi, and Tangerang). Greater Jakarta is a loosely defined term with few definitions. For instance one defines it with 13 cities and regencies but another understands Greater Jakarta only with 10 cities. We use the latter, which is widely perceived by citizens, excluding three

[1] "E-Government Indonesia" (Source: Ministry of Communication and ICT) http://www.detiknas.go.id/wp-content/uploads/2016/08/PAPARAN-SNGIDC-2016-Bambang-Dwi-Anggono.pdf

regencies, i.e., Kabupaten (Regency) Bekasi, Kabupaten Tangerang, Kabupaten Bogor, which rather connect other close provinces such as Banten and West Java.

Table 1 shows a part of the commuting survey results particularly showing commuters from the 10 cities in Greater Jakarta to five cities in Jakarta. For instance, among the total number of commuters to the five cities in Jakarta, 2.31% of people commute from South Jakarta to East Jakarta.

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	-	2.31%	4.69%	1.88%	0.97%
East Jakarta (EJ)	5.33%	-	4.66%	1.62%	3.34%
Central Jakarta (CJ)	1.83%	0.86%	-	1.70%	1.20%
West Jakarta (WJ)	2.65%	0.46%	4.65%	-	4.19%
North Jakarta (NJ)	0.87%	1.14%	3.17%	1.82%	-
Bogor (Bo)	0.34%	0.27%	0.39%	0.25%	0.05%
Bekasi (Be)	3.36%	7.27%	3.48%	1.10%	1.73%
Depok (De)	7.40%	2.35%	2.37%	0.72%	0.54%
Tangerang (Tg)	2.62%	0.17%	1.62%	4.23%	0.45%
South Tangerang (ST)	6.13%	0.33%	1.89%	1.30%	0.27%

Table 1 Commuting Survey 2014 - Statistics (%) from 10 cities in Greater Jakarta to 5 cities in Jakarta

2.2 Twitter Data

We collect all GPS-stamped tweets posted in Greater Jakarta from a data firehose and specifically for this study, we use a set of tweets posted between January 1st 2014 and May 30th 2014, for five months, considering that the official commuting survey was conducted during the first quarter of 2014. Given all tweets in the Greater Jakarta area, we first filter tweets posted using mobile devices only and then map and use the locations of tweets to the administrative resolution. Finally we use 38,491,430 tweets from 1,456,927 unique users.

3 METHODOLOGY

3.1 Inferring Origin-Destination Locations

First, per user, we infer two locations, Origin (Home) and Destination. It is worth noting that even though we finally produce city-level statistics, we process the locations of one's tweets at sub-district level, when determining the Origin and Destination cities.

- **Home Location** is inferred as the most tweeted sub-district location between 9pm and 7am in order to avoid using the locational information during commutes.
- **Destination Locatio** is determined as the most tweeted sub-district location during weekdays, excluding Home Location.

Using this approach, among 1,456,927 unique users who posted tweets in Greater Jakarta for five months from January 2014, we find the origin information of 877,054 users, and the origin and destination information of 305,761 users in sub-district level, which is about 2.8% (14%) of the whole population (the commuting population, respectively) in Greater Jakarta.

3.2 Calibrating the Initial Result

Once we map the Origin-Destination information at sub-district level to city level, we do a calibration based on the population data from 10 cities, due to the unequal penetration rates of Twitter in the cities. For instance, people in South Jakarta post tweets much more than people who are based in other cities in Greater Jakarta. After the calibration with Twitter penetration information, the cross correlation score between two forms of statistics, official statistics and the statistics from our approach improved from 0.92 to 0.97

4 RESULTS

The final result is presented in Table 3(a) and its chord diagram is shown in Figure 1. In Table 3(b) we show the rank difference between Table 2 and Table 3(a). For instance, the value for SJ ⇒ CJ is calculated as '0' because two ranks from two statistics are same as people in South Jakarta most commute to Central Jakarta than other three cities. Table 3(b) shows that our simple approach produces good results without significant differences. Even though three origin cities have three different rank pairs, it is worth mentioning that the values in the official statistics are originally close each other, and our results also present similar values. For instance, the values of De ⇒ EJ and De ⇒ CJ are 2.35 % and 2.37% in the official statistics (2.46% and 2.39%, respectively and in our results).

5 SUMMARY

We have shown that social media is a promising source of data to infer commuting statistics in Greater Jakarta, even with a simple approach and we will extend the method with further investigation and calibration methods using demographic information.

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	-	2.05%	4.69%	1.42%	0.67%
East Jakarta (EJ)	5.49%	-	4.77%	0.85%	2.18%
Central Jakarta (CJ)	1.89%	0.87%	-	0.93%	0.87%
West Jakarta (WJ)	3.53%	0.66%	5.07%	-	2.49%
North Jakarta (NJ)	1.52%	2.05%	3.88%	2.21%	-
Bogor (Bo)	1.47%	0.74%	2.16%	0.48%	0.39%
Bekasi (Be)	3.64%	6.48%	4.26%	0.96%	1.53%
Depok (De)	6.06%	2.46%	2.39%	0.58%	0.41%
Tangerang (Tg)	2.99%	0.64%	3.01%	3.56%	0.69%
South Tangerang (ST)	3.99%	0.40%	1.64%	0.77%	0.25%

ORIGIN	DESTINATION				
	SJ	EJ	CJ	WJ	NJ
South Jakarta (SJ)	0	0	0	0	0
East Jakarta (EJ)	0	0	0	0	0
Central Jakarta (CJ)	0	0	0	0	0
West Jakarta (WJ)	-1	0	0	0	+1
North Jakarta (NJ)	0	0	0	0	0
Bogor (Bo)	0	0	0	0	0
Bekasi (Be)	0	0	0	0	0
Depok (De)	0	+1	-1	0	0
Tangerang (Tg)	-1	0	+1	0	0
South Tangerang (ST)	0	0	0	0	0

Table 2 (a) Commuting Statistics from Social Media (left) (b) Rank Comparison with Official Statitics

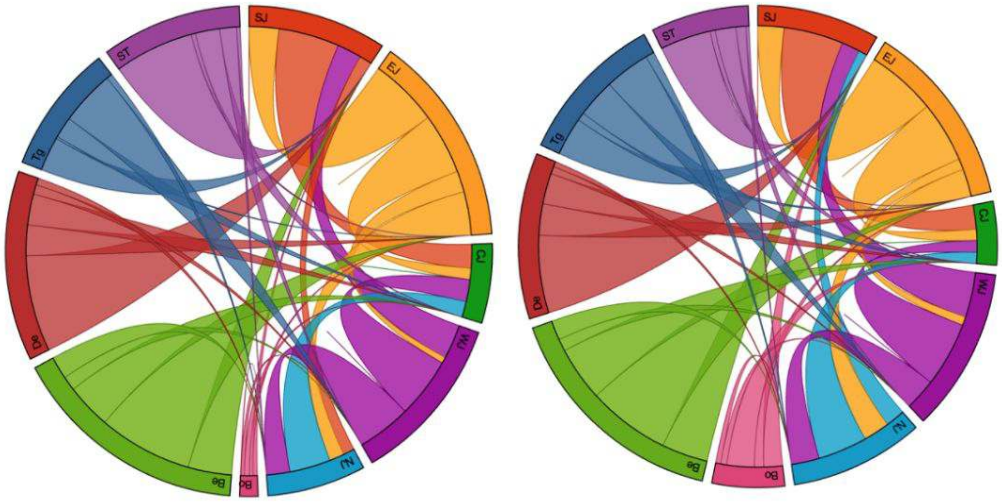

Figure 1 [Best Shown in Color] Official Commuting Flows (Left) and Statistics from Twitter (Right)

REFERENCES

[1] Jakarta Commuting Survey http://www.bps.go.id/index.php/publikasi/4416

[2] Palchykov V, Mitrović M, Jo HH, Saramaki J, Pan RK. Inferring human mobility using communication patterns. Scientific reports. 2014;4. doi: 10.1038/srep06174. pmid:25146347

[3] A Kurkcu, K Ozbay, EF Morgul, "Evaluating the Usability of Geo-Located Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for New York City"

[4] B Hawelka, I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos, C Ratti, Geo-located Twitter as proxy for global mobility patterns

[5] Liu, W., Zamal, F.A., & Ruths, D., (2012) Using social media to infer gender composition of commuter populations.

[6] Swier, N., Komarniczky, B., & Clapperton, B. (2015) Using geolocated Twitter traces to infer residence and mobility. GSS Methodology Series No 41

[7] Gonzalez MC, Hidalgo CA, Barabási A-L. Understanding individual human mobility patterns. Nature. 2008;453(7196):779–782. doi: 10.1038/nature06958. pmid:18528393

[8] Palchykov V, Mitrović M, Jo HH, Saramaki J, Pan RK. Inferring human mobility using communication patterns. Scientific reports. 2014;4. doi: 10.1038/srep06174. pmid:25146347

[9] Jiang S, Fiore GA, Yang Y, Ferreira F Jr, Frazzoli E, González MC. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (p. 2). ACM. (2013, August).