Spatial Point Data Analysis of Geolocated Tweets in the First Day of Eid Al-Fitr 2017 in Java Island

T W Wibowo
Faculty of Geography, Universitas Gadjah Mada
totok.wahyu@ugm.ac.id

Abstract. Eid Al-Fitr is a worldwide Muslim feast day, which in Indonesia generally accompanied by tradition of going home (mudik). The demographic patterns at the time of the holiday are generally shifted, in which some urban residents will travel to their hometowns. The impact of this shifting is that there is a quite massive mobility of the population, which is generally accompanied by traffic congestion. The presence of location sensors on smartphone devices, open the opportunity to map the movement of the population in realtime or near-realtime. Especially now that social media applications have been integrated with the capability to include location information. One of the popular social media applications in Indonesia is Twitter, which provides microblogging facilities to its users. This study aims to analyze the pattern of Geolocated Tweets data uploaded by Twitter users on the first day of Eid Al-Fitr (1 Syawal 1438H). Geolocated Tweets data mining is done by using Streaming API (Application Programming Interface) and Python programming language. There are 13,224 Geolocated Tweets points obtained at the location of the study. Various point data analysis techniques applied to the data have been collected, such as density analysis, pattern analysis, and proximity analysis. In general, active Twitter users are dominated by residents in major cities, such as Jakarta, Bandung, Surabaya, Yogyakarta, Surakarta and Semarang. The results of the analysis can be used to determine whether the Geolocated Tweets data mined by the Streaming API method can be used to represent the movement of the population when mudik.

Keywords: Text data mining, Twitter, Geolocated Tweet, Spatial point data analysis

1. Introduction
Population mapping is one challenging subject, since the object is moving continuously and dynamically. Traditional mapping technique such as choropleth map, is now still popularly used since it’s conception in 1826 by Baron Pierre Charles Dupin [1]. The usage of administrative boundary in the choropleth map sometimes can be misleading, therefore a new technique was initiated by George Julius Poulett Scrope and Henry Drury Harness [2]. The new technique was then popularized by using a Russian term “dasymetric” which means density measurement [2][3]. Generally, the dasymetric mapping is used to estimate the population by using one or several ancillary data, such as street, address, parcel distribution, and limiting weight [4]. Modifications were made by several researchers to improve the original binary idea of settlement and non-settlement. Instead of using binary logic in the original
The idea of dasymetric map, several researchers have tried to modify the mapping unit, such as dividing the low and high population density of the settlement [5][6] and population density fraction [7]. Nevertheless, the mapping techniques discussed above still relying on population data availability, which is also relatively immobile.

The world we know today is a very different place than before with the emergence of internet. Data is now rich, hence there is a new concept called big data. It is called big data because of the high velocity, volume and variety of the data [8]. Since big data is having its unique characteristics, the means of processing will also be special. Among other factors that influencing rich data environment, social media is gaining the most popularity in society.

By definition social media is online platforms that allow users to create and exchange content [9]. The data and/or information sharing in social media is made easier by the increasing number of smartphone user. Many social media applications have the option to share user’s location (geolocation), and smartphone has enable user to activate the feature conveniently. Geolocation data, which is designed primarily for mobile devices, is mainly used for geo-positioning, geocoding, and geotagging [10]. This geolocated data will enriched the geospatial data sources, which can be used further in spatial analysis especially in spatial distribution of population [11].

Twitter can provide its user to activated geolocated features to generate Geolocated Tweets. The existence of social media such as Twitter can provide opportunity to explore public opinion, social interaction, identity and cultural values [12]. Even though there are several assumptions that considered Twitter Data as “data-light”, many researchers have proved its potentially rich data [11][12]. As of 2016, Indonesia was ranked 3rd in active Twitter user, just behind United States and India [13]. As Twitter is a microblogging site, there is limitation in the number of characters that a user can post in a single tweet. This fact actually can inhibit the attribute data analysis, particularly in thematic related study [11]. For the last reason, this research only aims to map the distribution of population in Java Island in the first day of Eid Al-Fitr.

The majority of population in Java Island is moslem, and each year there is mudik tradition. Mudik can be interpreted as someone that is going home for a relatively short time after they settled in big city, mostly because they are working in big city. However, people that are conducting mudik usually will not stay in their hometown for a long time. This tradition can be seen as social, cultural, and/or psychological phenomena. This paper will examine the usage of Geolocated Tweets data, to provide the population movement during the first day of Eid Al-Fitr, in which many people committing mudik tradition. Generally, mudik is characterized by rapid and massive population movement, in a relatively short time. This will be an ideal time to test Geolocated Tweets data to get the knowledge of population movement.

2. Research Methods
2.1. Research Area
Java is the most populated island in the world. Based on 2010 census data, there are 106.45 million inhabitants in an island that has an area of 128,297 km². Generally, the population movement in Mudik tradition, is always focusing on the migrants in the capital/main cities back to their hometown. The most highlighted movement when mudik is migration from Jakarta to other cities in Java. Indonesian Government highly anticipated mudik activity in Java, since the risk of traffic accident usually rising. There are two main road that can be used by people, the north path (Jalur Pantai Utara) and the south path (Jalur Lintas Selatan). Mostly the north path is more congested than the south path (Figure 1). Not only by bus or car, mudik are also done by other transportation mode as well such as airplane, train, and even motorcycle. However, the people are usually not stay for long, as they have to go to work in their current residence.

2.2. Data
Twitter allows user to access their data through Application Programming Interface (API), one of which can be done by using streaming method. Typically there are three kinds of Streaming API, namely the
Public Stream, the User Stream, and Site Stream. Public Stream will allow the user’s application to get public data on Twitter, while User Stream enable user to stream Twitter data in real-time. Site Stream allow the user to monitor real-time Twitter feeds for a large number of users. The data used in this study were collected using Streaming API method, which employs Python Script to run the data mining [14]. This method dictate the user’s web server to always stay connected to Twitter in order to get real-time Tweets [15]. Once the connection is interupted, then the requested data will be forfeited. However, there is limitation on how much the user can stream the data, which is only 1% of the data pool each time request is made.

![Figure 1. Study area](image)

2.3. Data Analysis

As mentioned before that big data needs to be processed with special technique. Though the data used in this study is relatively small to be called big data, it still needs a way to harness insight through spatial analysis. In this study there are four analysis conducted to the data, which are distance analysis, density analysis, pattern analysis, and proximity analysis. All point data analysis used in this research aim to generalize the abundant points and get a new perception about the data.

2.3.1. Global Density. Global density was the first measure used in this study. It is simple yet can be used to get the first impression of the data. The density can be calculated by dividing the number of Tweets with its respective administrative area. In this research, provincial area based on Minister of Internal Affairs Regulation No. 66/2011 was used to calculate the global density of Tweets.

2.3.2. Quadrat Density and Point Density. Regular tessellation (square and hexagonal) was used as a based for quadrat count method, in which the number of point data is summed up according to its respective area [16]. Both tessellation polygons were created with an area of 1 km2 in each polygon. Points that are located within a polygon were then calculated, and presented in choropleth map. However, density calculation has difficulty in considering the definition of the study area [16]. The area for density calculation can be changed easily, but determining which is more representative area can be a challenging task. Therefore, point density analysis were done to provides an alternative of density calculation, as it offers different area determination of the previous two methods. Point density would produce a raster based density surface, in which the density calculation based on the points located within each cell and it’s neighbour cell.
2.3.3. **Variance-Mean Ratio (VMR).** The quantitative pattern measurement of point data could give the illustration of how the data is distributed, evenly spaced, random, or clustered. VMR calculates the ratio between variance and mean of quadrat count density, in a Poisson distributed data. VMR greater than 1.0 indicates a tendency toward clustering in the pattern, and a VMR less than 1.0 is indicates an evenly spaced arrangement. However, this approach is considered as an unreliable unless the volume of the data is very large and high intensity of events per quadrat [16].

2.3.4. **Average Nearest Neighbour (ANN).** Empirical point patterns, though studied for many reasons, is mainly expected to reveal the process that is responsible for creating the pattern [17]. Moreover, study of point pattern will inform us about fundamental relationships in space or about possible causes for the observed patterns [18]. The second order effects was detected using pattern analysis through Average Nearest Neighbour (ANN) analysis. It calculates a nearest neighbor index based on the average distance from each points to its nearest neighboring points. The index is the result of ratio between observed distance and the hypothetical data [19]. The following formula were used to calculate ANN:

\[
\text{ANN} = \frac{\overline{D}_O}{\overline{D}_E}
\]

where \(\overline{D}_O\) is the observed mean distance between each points and it’s nearest neighbour:

\[
\overline{D}_O = \frac{\sum_{i=1}^{n} d_i}{n}
\]

and \(\overline{D}_E\) is the expected observed mean distance between each point and its nearest neighbour:

\[
\overline{D}_E = \frac{0.5}{\sqrt{n/A}}
\]

2.3.5. **Tweets Proximity Index (TPI).** To determine the influence of a city on the data, an index that combines the normalization of data number with the average distance was built. Point distance analysis is used as basic analysis, in which the result can be used to calculate the number of points and also providing the average distance from source point. Among all cities in Java, 35 were selected as qualitatively assessed using the raw Tweets data. The city center were used in the calculation of Tweets Proximity Index (TPI), acted as the representation of the center of activity in the city. Search radius in point distance analysis can be set subjectively in regards characterisics found in study area. TPI can be calculated using the following formula:

\[
TPI = N + D
\]

where normalization of data number (N) can be calculated based on data in each city (Xc) as follows:

\[
N = \frac{X_c - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\]

and the average distance index (D) is the inversed value of average distance divided (\(\overline{x}\)) by search radius (r):

\[
D = 1 - \left(\frac{\overline{x}}{r}\right)
\]

The summary of methods used in this study is illustrated in Figure 2.
3. Results and Discussion

3.1. Spatial Point Analysis

By using Public Stream API, 13,224 Tweets were collected during the first day of Eid Al-Fitr in the research area. The data were then imported in GIS Software as point data, showing exactly where the Tweets has been made (Figure 3). General distribution of the points, showed that major cities residents still dominating in the study area.

![Figure 3. Geolocated Tweets acquired in 25th June 2017](image)

Global density of the points showed that DKI Jakarta has the highest point density which reached 4.33 Tweets/km², meanwhile East Java has the lowest (0.04 Tweets/km²). However, the global density might have a bias, since the area of each provinces is not the same. Among all provinces in study area, DKI Jakarta has the smallest area, whilst West Java is the largest. Table 1 provide the calculation of global density analysis.
Table 1. Global density

| No | Province       | Tweets Count | Area (km²) | Global Density (Tweets/km²) |
|----|----------------|--------------|------------|----------------------------|
| 1. | Banten         | 1,173        | 9662.92    | 0.12                       |
| 2. | DI Yogyakarta  | 765          | 3133.15    | 0.24                       |
| 3. | East Java      | 1,875        | 47799.75   | 0.04                       |
| 4. | Central Java   | 1,871        | 32800.69   | 0.06                       |
| 5. | West Java      | 4,667        | 35377.76   | 0.13                       |
| 6. | DKI Jakarta    | 2,873        | 664.01     | 4.33                       |

Quadrat density provides a more balanced density calculation by using tessellation area, since the density is calculated using the same area. The result of square and hexagon tessellation quadrat density shows similar pattern, with the latter having a higher maximum density (Figure 4a & Figure 4b). The shape of hexagon made it possible to have a higher point density, since it is the most effective tessellation shape. However, the pattern shown by both tessellations is quite the same.

The result from point density analysis shows a wider area of density compared to quadrat density. Addition of search area parameter in point density analysis makes the probability of a point entering a calculation area becomes higher. Jakarta and its vicinity still filled with high density Tweets, and so does other provincial capital and main cities (Figure 4c). From three calculated density analysis, it can be concluded that the Tweets still dominated by Twitter users in provincial capital and several main cities.

Further analysis using average point distance method was calculated to determine which provincial capital holds the most Tweets. Two search radius were used to determine the average distance of points from city center. The lower the average distance, it means that there are a lot of Tweets that posted around the city center. In the first search radius (25 km), Yogyakarta has the lowest average distance points (Figure 5). But the second search radius (50 km), shows that Tweets in Bandung has the lowest distance. Although has a high density, evidently the Tweets that were made in Jakarta has the lowest average distance.
Figure 4. a) Quadrat density – square; b) Quadrat density – hexagon; and c) Point density

Figure 5. Average point distance to provincial capital

In order to obtain the quantitative measurement of pattern, Variance to Mean Ratio and Average Nearest Neighbour were performed. VMR was performed by using the same tessellations structure as quadrat density calculations. Both VMR analysis were yielded a relatively high above the threshold of an evenly distributed event, in which the threshold is 1 [16]. The findings of VMR calculations was
confirmed by Average Nearest Neighbour analysis. Given the z-score of -169.05, there is a less than 1% likelihood that this clustered pattern could be the result of random chance. Table 2 summarize the VMR analysis on both tessellations, meanwhile Figure 6 shows the result of ANN analysis.

| No | VMR Analysis | Tessellations |
|----|--------------|---------------|
| 1  | Number of Tweets | 13,224         | 13,224         |
| 2  | Number of tessellations | 326,508       | 330,948       |
| 3  | Mean quadrat count | 0.04           | 0.04           |
| 4  | Observed variance | 0.86           | 0.88           |

VMR | 21.62 | 22.08

Figure 6. The result of ANN analysis

3.2. First and Second Order Impact

From various analyzes that have been done, it is known that the characteristics of tweet data collected is clustered and located around several main cities. The average distance from provincial capital is also relatively close. Based on the evidence, it can be drawn that the main city people still dominate active users Twitter who is also happened to be active (Tweeted) at the time of Eid-Al Fitr. Presumption elaborated previously in the background, in this case is not proven, since evidently there are much geolocated Tweets in main cities (especially Jakarta) instead of destination city of mudik. This leads to another assumption that the first order impact is stronger than the second order, which means that infrastructure and supportive environment to access the internet still plays a large role in social media activity particularly Twitter. With far better internet facilities and infrastructure in main cities, the citizen can easily become active in social media. Furthermore, Twitter users who usually make Geolocated Tweets, do not migrate even though the mudik tradition is going on.

The findings also indicated that there are differences between citizen and netizen. The purpose of this study aims to connect the movement of citizen and the Geolocated Tweets. However, apparently not all citizen are also active netizen, particularly who use Twitter. As a result, the data is showing a clustered pattern in which many netizen actively using Twitter during the first day of Eid Al-Fitr. A more elaborated study is needed to explain this behaviour.

During the first day of Eid-Al Fitr in Indonesia, Muslims have a tradition to apologize to each other. Not only verbally, apologies are also often delivered through social media. The majority of Tweets mentioned Eid Al-Fitr or something that are still associated. It means that Geolocated Tweets data
acquired from Public Streaming API provides the content of user’s Tweets, in which can provide many applications regarding thematic mapping.

3.3. Tweet Proximity Index

The Tweet Proximity Index (TPI) was developed and used to determine which city has the highest influence in Geolocated Tweets. This index was based upon two approaches, which are average distance and the counts of Tweets. The smaller average distance indicates the stronger impact of a city, on the other hand the lessTweet counts indicates the less impact of a city. For the purpose of TPI calculation, 35 cities were selected by considering the point density.

Table 3 shows TPI calculation using search radius of 25 km. Megapolitan cities, that consist of Jakarta, Depok, South Tangerang, Bekasi, and Tangerang, were still dominating the top list of TPI, only Bogor that placed outside the top 10. Despite having a slightly significant number of Tweets, the average distance of Tweets in Bogor is relatively distant. The usage of circular shape search radius might not be an advantage for an elongated shape city like Bogor, so that the TPI was low. In this case an ellipsoid search radius is more suitable, but there must be a consideration of the angle of the axes.

Jember unexpectedly, despite having the least number of Tweets (within a radius of 25 km), was able to occupy the eighth rank because it has the closest average tweet distances. This shows that Jember residents mostly make Tweets around the city center, which can be attributed to the completeness of facilities and infrastructure that centered around city center as has been alluded to earlier. Cilacap, Pekalongan, and Madiun have the similar characteristic to Jember, in which they are only having small number of Tweets (below 100) but the average distance is relatively close.

4. Conclusions

At least for the time being, the abundance of Twitter Data couldn’t represent the movement of people during mudik. Apparently there is still a big differences between citizen and netizen, in this case netizen who posted Geolocated Tweets. Twitter user, especially those who Tweet with Geolocated features, still
References

[1] Friendly M and Denis D J 2001 Milestones in the history of thematic cartography, statistical graphics, and data visualization [Online]. Available: http://www.datavis.ca/milestones/index.php?page=home.

[2] Mennis J 2009 Dasymetric Mapping for Estimating Population in Small Areas Geography Compass 3(2) pp 727-745.

[3] Wright J. K 1936 A method of mapping densities of population with Cape Cod as an example Geographical Review 26 pp 103-110.

[4] Tapp A F 2010 Areal Interpolation and Dasymetric Mapping Methods Using Local Ancillary Data Sources Cartography and Geographic Information Science 37(3) pp 215-228.

[5] Mennis J 2003 Generating surface models of population using dasymetric mapping The Professional Geographer 55(1) pp 31-42.

[6] Weischelbaum J, Petrini-Monteferri F, Papathoma M, Wagner M and Hackner N 2005 Sharpening census information in GIS to meet real-world conditions: the case for Earth Observation WIT Transactions on Ecology and The Environment 84 pp 143-154.

[7] Steiner P and Paulus G 2007 Dasymetric mapping for public health planning Aalborg.

[8] Kaisler S, Armour F, Espinosa J A and Money W 2013 Big data: Issues and challenges moving forward Hawaii.

[9] Gandomi A and Haider M 2015 Beyond the hype: Big data concepts, methods, and analytics International Journal of Information Management 35 pp 137-144.

[10] Ranitovic J, Nikolic V and Velmirovic L 2015 Geolocation-based social networks and mapping in emergencies.

[11] Mislove A, Lehmann S, Ahn Y, Onnela J and Rosenquist N 2012 Understanding the Demographics of Twitter Users in Fifth International AAAI Conference on Weblogs and Social Media.

[12] Sloan L, Morgan J, Burnap P and Williams M 2015 Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data PLoS ONE 10(3) pp 1-20.

[13] Statista 2016 Statista [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ [Accessed 12 08 2017].

[14] Klinger R 2015 The Geography of Tweets: Reading Tweets with QGIS [Online]. Available: http://www.digital-geography.com/the-geography-of-tweets-reading-tweets-with-qgis/#.WM75ODt97IU. [Accessed 13 08 2017].

[15] Reifman J 2014 Building With the Twitter API: Using Real-Time Streams [Online]. Available: https://code.tutsplus.com/tutorials/building-with-the-twitter-api-using-real-time-streams--cms-22194. [Accessed 12 08 2017].

[16] O'Sullivan D and Unwin D J 2010 Geographic Information Analysis, 2nd ed (New Jersey: John Wiley & Sons, Inc).

[17] Okabe A, Boots B, Sugihara K and Chiu S N 2000 Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd ed (Chichester: John Wiley & Sons Ltd).

[18] Rogerson P 2006 Statistical Methods for Geography, 2nd ed (London: Sage).

[19] Mitchell A 2005 The ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics, Esri Press.

[20] Langford M and Unwin D 1994 Generating and mapping population density surface within a geographical information system The Cartographic Journal 31 pp 21-26.

dominated by Metropolitan/Urban/Downtown resident. However, further analysis is needed to elaborate this fact. Not only assess the data spatially, but also collecting Twitter’s user opinion regarding their behaviour of using Twitter. On the other side, there is a huge possibility that Twitter Data can be used in attribute data analysis through it’s user Tweet, although a robust data cleaning will be needed.