ISSN: 0852-0682, EISSN: 2460-3945
Forum Geografi, Vol 33 (1) July 2019: 82-100
DOI: 10.23917/forgeo.v33i1.8021

FORUM GEOGRAFI
Indonesian Journal of Spatial and Regional Analysis

Tourist Attraction Popularity Mapping Based on Geotagged Tweets

Totok W. Wibowo [*], Ahmad F. Bustomi , Anggito V. Sukamdi
Faculty of Geography, Universitas Gadjah Mada, Yogyakarta, Indonesia 55281
[*)] Corresponding Author (e-mail: totok.wahyu@ugm.ac.id)

Received: 20 April 2019 / Accepted: 17 July 2019 / Published: 01 August 2019

Abstract. The development of tourist attractions is now highly influenced by social media. The speed at which information can be disseminated via the Internet has become an essential factor in enabling distinct tourist attractions to potentially gain high popularity in a relatively short time. This condition was not as prevalent several years ago, when tourism promotion remained limited to a certain kind of media. As a consequence, rapid change in the relative popularity of tourist attractions is inevitable. Against this, knowledge of tourist attraction hotspots is essential in tourism management. This means there is a need to study the means by which to both quickly determine the popularity level of tourist attractions and encompass a relatively large area. This article utilised tweet data from microblogging website Twitter as the basis from which to determine the popularity level of a tourist attraction. Data mining was conducted using Python and the Tweepy module. The tweet data were collected at the end of April and early May 2017, at times when there are several long holiday weekends. A Tweet Proximity Index (TPI) was used to calculate both the density and frequency of tweets based on a defined search radius. A Density Index (DI) was also used as a technique for determining the popularity. The results from both approaches were then compared to a random survey about people's perceptions of tourist attractions in the study area. The result shows that geotagged tweet data can be used to determine the popularity of a tourist attraction, although it still only achieved a medium level of accuracy. The TPI approach used in this study produced an accuracy of 76.47%, while the DI achieved only 58.82%. This medium accuracy does indicate that the two approaches are not yet strong enough to be used for decision-making but should be more than adequate as an initial description. Further, it is necessary to improve the method of indexing and the exploration of other aspects of Twitter data.

Keywords: Twitter, geotagged, hotspot, popularity, tourism.

Abstrak.
Perkembangan objek wisata pada saat ini tidak dapat terpisahkan dari media sosial. Kemampuan internet dalam menyebarkan informasi telah membuat suatu objek wisata dapat secara singkat meraih popularitas yang tinggi. Hal ini tentu berbeda dengan kondisi beberapa tahun yang lalu, yang mana promosi objek wisata masih sangat terbatas. Perubahan popularitas pun menjadi hal yang tak terelakkan karena tingkat penyebaran data yang begitu cepat. Di sisi lain pengetahuan tentang tingkat popularitas objek wisata sangat diperlukan dalam penentuan prioritas pengembangan yang menyeluruh. Dengan demikian diperlukan kajian untuk dapat memetakan tingkat popularitas objek wisata secara cepat dan dapat menjangkau daerah yang luas. Artikel ini akan memanfaatkan sumber data dari situs *Microblogging Twitter*, sebagai dasar untuk penentuan tingkat popularitas suatu objek wisata. Penambangan data (*data mining*) dilakukan dengan menggunakan bahasa *Python* dan modul *Tweepy*. Data dikumpulkan pada saat libur panjang di akhir bulan April dan awal bulan Mei tahun 2017, yang mana diasumsikan akan terdapat banyak wisatawan yang berlibur. *Tweet Proximity Index* (TPI) digunakan untuk menghitung kepadatan tweet dan frekuensi tweet, berdasarkan radius pencarian yang ditentukan. *Density Index* (DI) juga digunakan untuk

memberikan pendekatan lain untuk menentukan popularitas objek wisata. Kedua hasil analisis akan dibandingkan dengan survei secara acak tentang persepsi masyarakat terhadap objek wisata di wilayah kajian. Survei secara langsung juga dilakukan untuk mengetahui akurasi hasil analisis yang telah dilakukan. Hasil penelitian menunjukkan bahwa data geolocated Tweets dapat digunakan untuk penentuan popularitas objek wisata. TPI menghasilkan akurasi yang lebih tinggi (76,47%) daripada DI (58,82%). Akurasi menengah ini memang menunjukkan bahwa kedua pendekatan tersebut belum cukup kuat untuk digunakan untuk pengambilan keputusan, tetapi lebih dari cukup untuk digunakan sebagai deskripsi awal popularitas objek wisata. Perbaikan metode penyusunan indeks maupun eksplorasi aspek lain dari data Twitter perlu dikembangkan untuk mendapatkan nilai akurasi yang lebih tinggi.

Kata kunci: *Twitter, geotagged, hotspot*, popularitas, pariwisata.

1. Introduction

Indonesia has experienced rapid development of social media over recent years. Many factors have contributed to this development, including hardware, software and infrastructure development. Among such factors, however, information technology infrastructure plays a huge role in promoting and supporting the development of social media; for instance, the recent implementation of a 4G network in Indonesia. The latest generation of broadband Internet provides far higher speeds than the previous generation (Fauzi *et al.*, 2012). Around the same time, the smartphone has become a ubiquitous item. The competitive price of smartphones, combined with their inbuilt sensors and functionality, has led to their widespread use by people as an enhanced telecommunication device. Furthermore, the addition of a Global Positioning System (GPS) sensor in smartphones opens up the possibility of recording geospatial data.

Users have a choice of many different social media platforms, although it is relatively common for a user to be active across numerous different platforms. Twitter, a microblogging social media website, is a platform with a relatively large number of users in Indonesia. Statista (2016) noted that in 2016 there were 24.34 million active Twitter users in Indonesia, which means that Indonesia has the third-highest number of active Twitter users in the world after the United States and India. There are also various different groups of Twitter users, ranging from government officials, politicians, academics and advertisers, to students who are still at school (Huberman *et al.*, 2008). Even the president of the United States has a specific Twitter account called POTUS (President of the United States). In contrast to other social media platforms such as Instagram and Facebook, the Twitter Application Programming Interface (API) is more accessible, thus increasing the possibility of obtaining more data.

The growing number of users will directly result in massive transfers of data between users and the server. The server will also be affected by the very high volumes of data being stored, which can even exceed the limits of big data (exabyte/10^{18}). The concept of big data has existed since the beginning of computing because it was used incipiently to identify data that could not be processed efficiently using traditional database methods (Kaisler *et al.*, 2013). Thus, due to its different characteristics, big data required special handling for its processing. There are two main things to consider when handling big data, namely the design of a system that is capable of handling such large volumes of data and the ability to filter it according to specific objectives (Katal *et al.*, 2013).

The impressive thing about tweets is the option to add position data, which in this case is supported by the GPS found on smartphones. A tweet that incorporates location information (a geotagged tweet) can be used for the purposes of spatial visualisation and spatial analysis. Although, according to the data, only 5% of all tweets have position information (Carto, 2017), it is undeniable that their existence has added new data sources in mapping as outcomes of location-based social media (Thatcher, 2014), in addition to the data sources mentioned in several kinds of literature (Kraak & Ormeling, 2013). The current and recent use of geolocated tweet data has been very diverse, ranging from studies on happiness level (Frank *et al.*, 2013), sense of place (Jenkins *et al.*, 2016), global mobility patterns (Howelka *et al.*, 2014; Yin and Du, 2016), to Twitter network analysis (Takhteyev et al., 2012) and rainfall data correlation (Lwin *et al.*, 2015).

The results are able to reveal things that were previously difficult to do. Indeed, even the act of obtaining data for a study was more challenging. This opportunity is inseparable from the role of technology in transforming humans into active sensors for the purpose of data collection (Miller & Goodchild, 2015) in such a way as to engender a shift in the data collection paradigm. Whereas in the past data collection was based on data-scarce activity, there has now been a shift in the paradigm due to the fact that currently, respondents actively collect data (data-rich).

Tourism was declared a national priority in the 2015-2019 Medium Term Development Plan (RPJM), with the hope that by the end of 2019 there would be 20 million visiting foreign tourists and 275 million local tourists (Setkab, 2017). The tourism sector is highly strategic in terms of its role in increasing economic activity and supporting regional development. Ideally, these efforts will be accompanied by improvements in the facilities and infrastructure at each tourist attraction. The management of tourist attractions that have been integrated into one administrative area will support the implementation of such regional development. Therefore, information is needed on the popularity of tourist attractions. Ideally, more popular attractions will require more resources than less popular attractions.

In recent years, social media has contributed significantly to the dissemination of tourism information. Some social media accounts are even created specifically for the purpose of tourism promotion. Interactions between social media users have the power to encourage users to visit certain tourist attractions. Moreover, the information presented on social media is not just textual in nature but also features multimedia content. The abundance of multimedia data on social media provides the opportunity to study a variety of things. Nevertheless, it is still necessary to process the data carefully, particularly in the stages of data collection and management. Data analysis can then be applied as needed.

New tourist attractions, such as Breccia Cliff Park, Amaryllis Park and Kalibiru Tourism Village, are notable for having rapidly gained popularity among social media users. It is important to be prepared for such popularity in order to be in a position to maximise the visitor experience. The influence of social media on the popularity of legendary tourist attractions is another interesting case to study. Adaptation is the key for any tourist attraction to retain its popularity and attract visitors. As an example, there is the transition from agriculture and fisheries to total tourism in Karimunjawa (Setiawan *et al.*, 2017). Borobudur, which had setbacks and was abandoned, was able to achieve a high level of popularity through a process of adaptive transformation (Baiquni, 2009).

Twitter allows users to access data on a server using an API which is limited by regulations. Users' Twitter data, especially geotagged tweets, can be used to map the distribution of the popularity of attractions quickly and efficiently. However, the accuracy of the method's use in determining popularity still needs to be assessed. This paper will examine the usefulness of Twitter data as an indicator to assess the popularity of tourist attractions.

2. Literature Review

2.1. Big Data

The development of mobile computing hardware has followed Moore's law for decades. The increase in hardware production has also had an impact on the volume of data collected owing to the fact that almost every electronic device has a mechanism for obtaining data. However, in the current information age, the ability to handle large volumes of data continues to evolve (Tsai *et al.*, 2015). That is why Fisher *et al.* (2012) showed that big data involves data that cannot be handled and processed by most current methods or information systems.

The characteristics of big data that are often discussed are 3V, namely volume, velocity and variety (Laney, 2001). These three characteristics explain the "big" term in big data. Volume refers to a massive data size, velocity refers to transfer rates and variety refers to the large variety of data structures. However, the concept of 3V is now no longer suitable for describing big data (Rijmenam, 2013; Borne, 2014). To describe the characteristics of the current trend in big data, we need to add several additional features, namely veracity, validity, value, variability, venue, vocabulary and vagueness.

Data mining is the study of the collection, cleaning, processing, analysis and acquisition of meaningful information from a data set (Aggarwal, 2015). In its utilisation, there are numerous variations in the problem domain, application, formulation and data representation. Thus, the term data mining is wide-ranging in its use to explain several aspects of data processing. The abundance of data is a direct impact of technological development and computerisation in various aspects of life.

The systematics of data collection must accommodate the purpose of data usage. However, there is also the possibility of reusing the same data for different purposes. In this case, data mining can be used as a medium for extracting data from various sources for its later management and presentation (Aggarwal, 2015). Raw data will be collected, cleaned and transformed into a standard format for processing. Data can be stored in commercial database systems and then processed using various analytical methods to gain insight/ information. Within the entire process, the majority of data mining work is focused on data preparation.

2.2. Twitter API

Founded by four people in 2006, Twitter is a microblogging site that allows users to post messages comprising a maximum of 140 characters of text. Despite its simple concept, in its development, Twitter has become a choice of social media platform that is widely used by various different groups. Within five years of its release, there were 100 million active Twitter users (O'Reilly & Milstein, 2012).

A follower is the most basic level of user interaction on Twitter. The first account will always get the latest tweets from the second account. Furthermore, the first account has the option to distribute specific tweets from other accounts (known as retweeting). Users can also mention other accounts on Twitter, while the feature of many more interactions among other users is what differentiates Twitter as unique compared to other social media.

Every tweet by a user will be stored on the Twitter server that is certainly equipped with cybersecurity. However, like most web services, Twitter has an API that allows users to download data using predetermined rules. Streaming API provides low latency access to stream tweet data globally. A streaming client will receive a push notification about tweets that match their search criteria. Streaming API enables data to be obtained in real time. As at the time of the research, Twitter has three types of streaming API, namely:

a. Public streams: enable the tracking of public data on the Twitter timeline. Used to find out specific topics and for data mining.

b. User streams: allow searching on a Twitter user account. The result of the research is data that corresponds to the desired account.

c. Site streams: a multi-user version of user streams. Connections to Twitter are required to use a server and represent multiple users.

3. Research Method

3.1. Data Mining

Data mining was carried out using the Public Streaming Twitter API. In this case, four keys needed to be generated from the Twitter developer page, namely access tokens, access token secret, consumer key and consumer secret. The function of the keys is to get legitimacy to stream to Twitter's data via OAuth.

The scripting was carried out using the Python programming language. Not all tweets were collected in this study as only geotagged tweets were relevant. Thus, in streaming, it is necessary to limit the search area, in our case to the Central Java Province and the Special Region of Yogyakarta (Figure 1). The search area limit parameters were included in the script as one of the query criteria.

Python requires an additional Tweepy library to communicate with the Twitter API. Installation of the Tweepy module is done directly in the Python storage directory that is associated with QGIS. This is done to maintain the independence of the Python installation from the various software on the computer. The first part of the script contains several functions from the Tweepy module, which is then followed by providing the four previously obtained accesses and keys. The command to stream tweets is written in the next section, which is then followed by authentication and entry of the keyword as the basis for the query.

The script can be executed through the Console / Terminal / Command Prompt, which is available on any desktop operating system. In addition, some GIS software provides direct access to Python through the GUI Console, with one way being to use Quantum GIS. In this study, Python script was executed from QGIS because it can be set to directly display geospatial data. Data collection was carried out over two long weekends at the end of April and early May 2017 because generally, the number of tourists will increase over both of these holidays. As the method chosen was Streaming, the script continued during the time the query was being run.

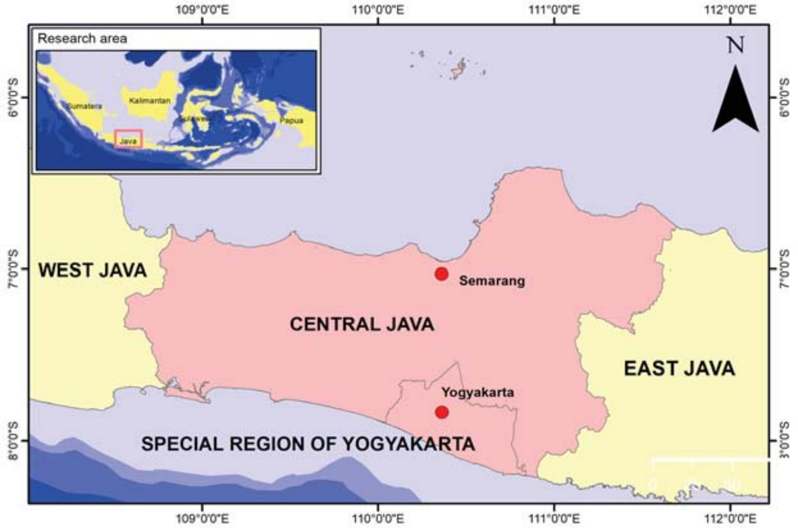


Figure 1. Research area (indicated by light red colour).

3.2. Data Visualisation

The point data visualisation technique has a unique characteristic. However, there is a need to generalise its appearance should the volume of data become too high. Simple data visualisation can be valuable for analysis. In this case, density value can be used to simplify spatial point data. Global density constitutes the simplest density calculation, which divides the population over the administrative boundary area. This visualisation method provides a very effective aggregation of point data. However, the use of arbitrary administrative boundaries tends to lead to subjectivity and to details being missed from the display.

The issue of which areas to select to represent the data point is referred to as the Modifiable Areal Unit Problem (MAUP). If not treated carefully, MAUP can lead to bias. The tessellation polygon technique can be used to address this problem and divides the study area into a grid with predetermined shapes and sizes. An area of 1 square km is assumed to be sufficient to represent the effective service area of an average tourist attraction.

3.3. Data Analysis

While visualisation is intended to produce a general picture regardless of the tourist attraction, point analysis is carried out as an approach for determining the popularity of attractions. Radius of Gyration is a measure that is often used to determine and quantify the effect of distance reduction on mobility patterns (Gonzalez *et al.*, 2008). However, since the moment of inertia effect does not impact on the creation of a tweet, an alternative approach is needed.

The first approach emphasises the measurement of the number of tweets and the distance to the measured point. The Tweet Proximity Index (TPI) is used for this purpose. TPI is calculated based on two parameters, namely the index of the number of tweets and the average distance index of tweets (Wibowo, 2017). Both calculation parameters are carried out at a defined radius from the tourist point. The TPI value ranges from 0 to 2, where 0 indicates no tweets at all and 2 denotes many tweets and that the location is in the tourist attraction. In this study, we used a search radius of 1 km. Figure 2 illustrates the spatial depiction of TPI in each tourist attraction.

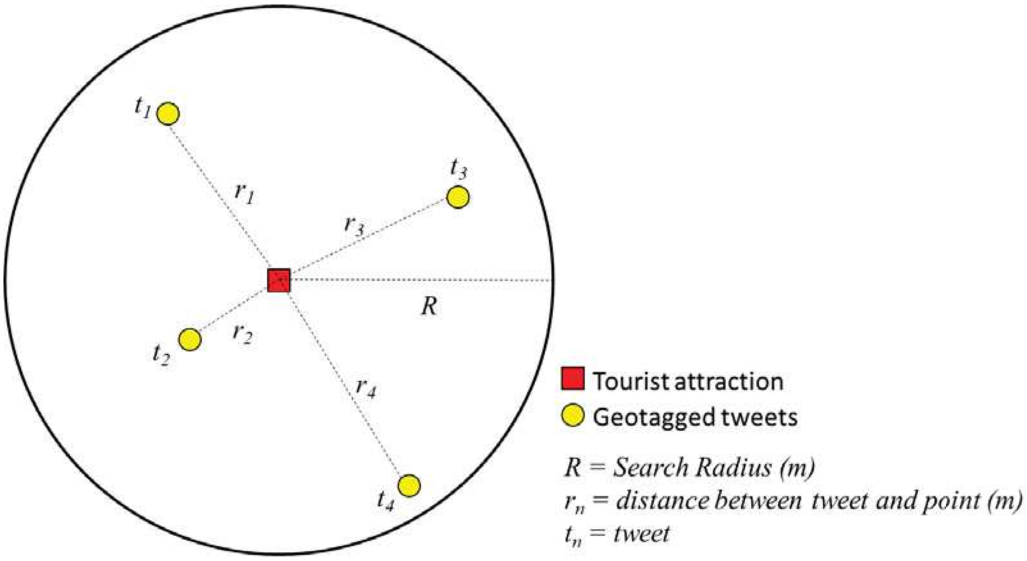


Figure 2. Spatial depiction of TPI calculation.

Point density was used as a second approach to determine the popularity. The density index (DI) was calculated based on the results of point density analysis using the kernel density estimation (KDE) principle. The grid size used was equated with a search radius for TPI calculations. Theoretically, the denser the Twitter data on a tourist attraction, the more popular the tourist attraction. Measurement of density level was carried out using the point density algorithm in GIS software, which in principle will also pay attention to neighbouring cell density.

3.4. Popularity Assessment

In recent years social media has become a very effective means of disseminating tourism information. Many new attractions have become very popular as a result of information uploads, which act as a chain message for social media users. According to official data, there are more than 100 tourist attractions in the study area, although this figure does not include attractions that are popular because of social media. The popularity of tourist attractions was measured through the random dissemination of questionnaires using an online survey form. The items in the questionnaire were divided into four stages (sections), namely identity, Twitter data, tourism data, and social media and tourism. The target respondents were tourists in several tourist locations. The age limit of the respondents was determined by selecting respondents who were most likely to have social media and actively use it (ages 15–50 years).

Seventeen tourist attractions that were rated popular by 144 respondents, as indicated by a high number of votes, were used as the reference data. Meanwhile, the same number of tourist attractions with the greatest TPI and the highest DI was also selected. The accuracy of both approaches was assessed by comparing them with the reference data. Accuracy was indicated as a percentage, denoting the extent to which TPI and DI can predict the correct tourist attractions.

4. Results and Discussion

4.1. Tweet Data

A total of 85,096 tweets were obtained from the data mining before going through a data cleaning process. The aim of the cleaning was to remove data from outside the research area. The amount of Twitter data from within the study area stood at 76,859 (90.32%), with the remainder found from within Indonesia but outside the search area. This query imperfection was likely caused by various data that did not have location information but were nevertheless captured by the query script.

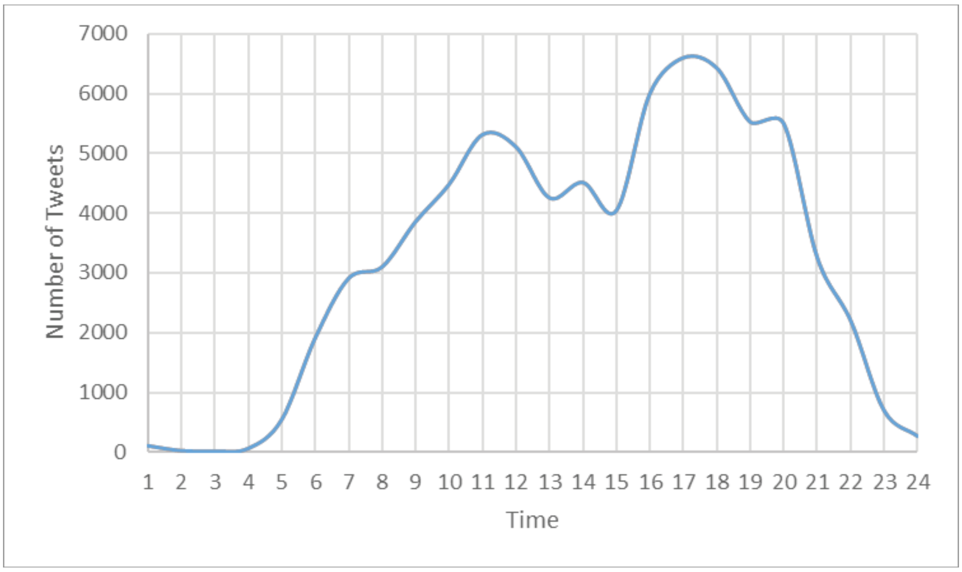


Figure 3. Frequency of tweets per hour.

Data acquisition generally began in the morning and ended at night. Figure 3 shows there were two peaks in terms of the number of tweets, which occurred during the day and evening. Data acquisition fell dramatically between midnight and 4 am, which we can assume is because many of the tourist attractions in the study area were closed during this time.

Twitter data mining using the Streaming API method requires users to always be connected to the Twitter server. If the query is met with a connection problem, then the data mining will be forfeited, which is one disadvantage of using the Streaming API method. One option for overcoming this problem is to reduce the amount of data for queries that can be implemented. In this case, the user must diligently perform a re-query if the previous task has finished running. The duration of a query depends on the desired area; the wider the area, the shorter the query time will be. Conversely, a narrower search area will require a longer query time.

In general, the distribution of the spatial data displays a clustering pattern in locations such as Yogyakarta, Surakarta, Semarang and Magelang (Figure 4). The amassing of data in the four cities seemed to dominate the distribution of tweets at the study site. Further examination of the map indicates a longitudinal pattern which has a strong association with road network data. Some coastal areas have a relatively large volume of Twitter data; for example, Cilacap, Bantul, Tegal, Pekalongan and Jepara Regencies.

Several areas around the Kendeng Hills, such as Grobogan Regency, Rembang Regency and Blora Regency, have a very small number of tweets compared to other regions. A quite similar pattern can be seen in the western central zone, which has a hilly and mountainous topography that would certainly hinder Internet infrastructure. Data from various cellular operators in Indonesia confirms this condition, especially for the Kendeng Hills region. On the contrary, large cities are widely covered by cellular operator services from various networks, and this can act as a growth stimulant for social media users.

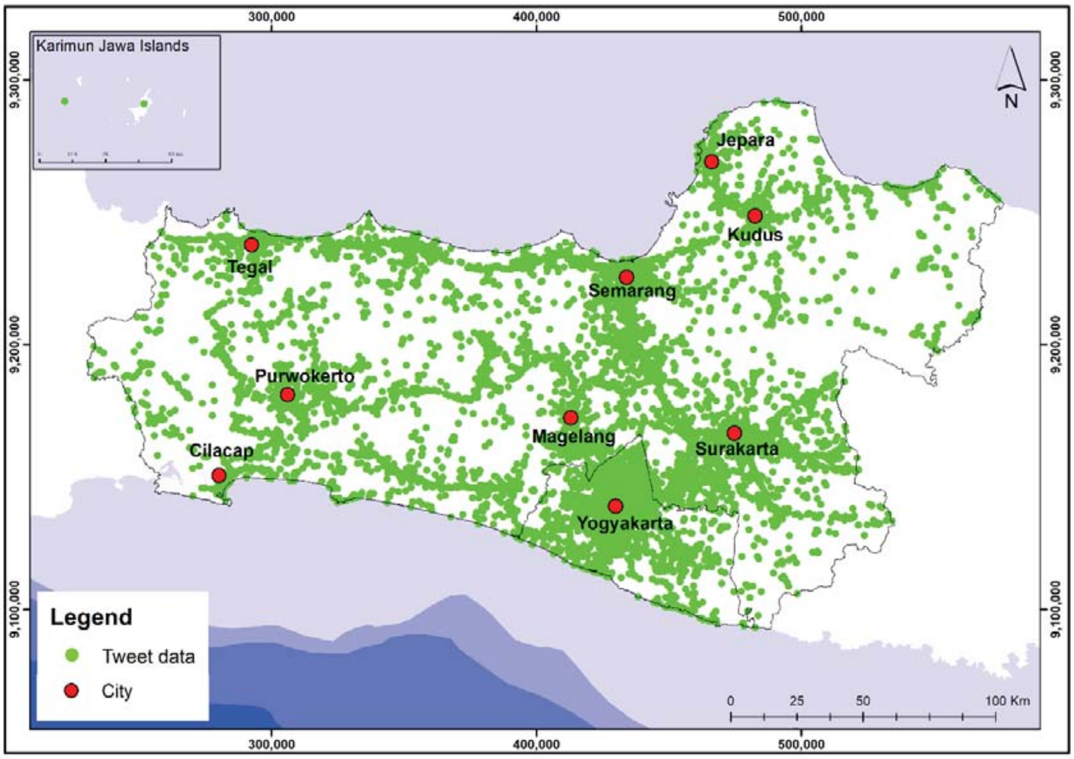


Figure 4. The result of Twitter data mining.

4.2 Global Density Analysis

Figure 5 exhibits the global density of tweets calculated based on regency boundary. Semarang, Magelang, Yogyakarta and Surakarta Regency dominate when it comes to high tweets density. The tweets density in those regencies exceeded 4 tweets/km^2. Moreover, the tweets density in Yogyakarta City stood at 421 tweets/km^2. The latter is far in excess of the average tweets density, which was only 18 tweets/km^2. In addition to the regencies/cities in the area, only Banyumas Regency, Pekalongan City, Tegal City and Kudus Regency have relatively high density values. Other districts/cities have a density of 1 tweet/km^2 or lower. Global density analysis tends to be very subjective and can sometimes be misleading because there is a rather forced data aggregation. This visualisation method can be used to give a global perspective or perform a regional analysis.

4.3. Tessellation Polygon Density Analysis

The substituting of administrative boundaries with uniform boundaries can provide a more objective assessment of density. In this case, we used a square tessellation polygon with an area of 1 km^2. A more uniform division of the unit analysis allows for a more thorough calculation of tweets. An area size of 1 km^2 is assumed to be sufficient to represent the average area of tourism since activity would only be practical within close proximity. One of the advantages of using tessellation polygon visualisation is that it conveys the dramatic difference between neighbouring polygons.

The results of the tweet density calculation based on the tessellation polygon can be seen in Figure 6. Clusters of tweet density can be observed in Yogyakarta City, Semarang City and Surakarta City. Linear patterns along the road found in the initial data can be represented well in this visualisation method, unlike the previous visualisation. Of course, this is an advantage because it enables a more detailed pattern to be presented, but with a level of information that is simpler than the original data.

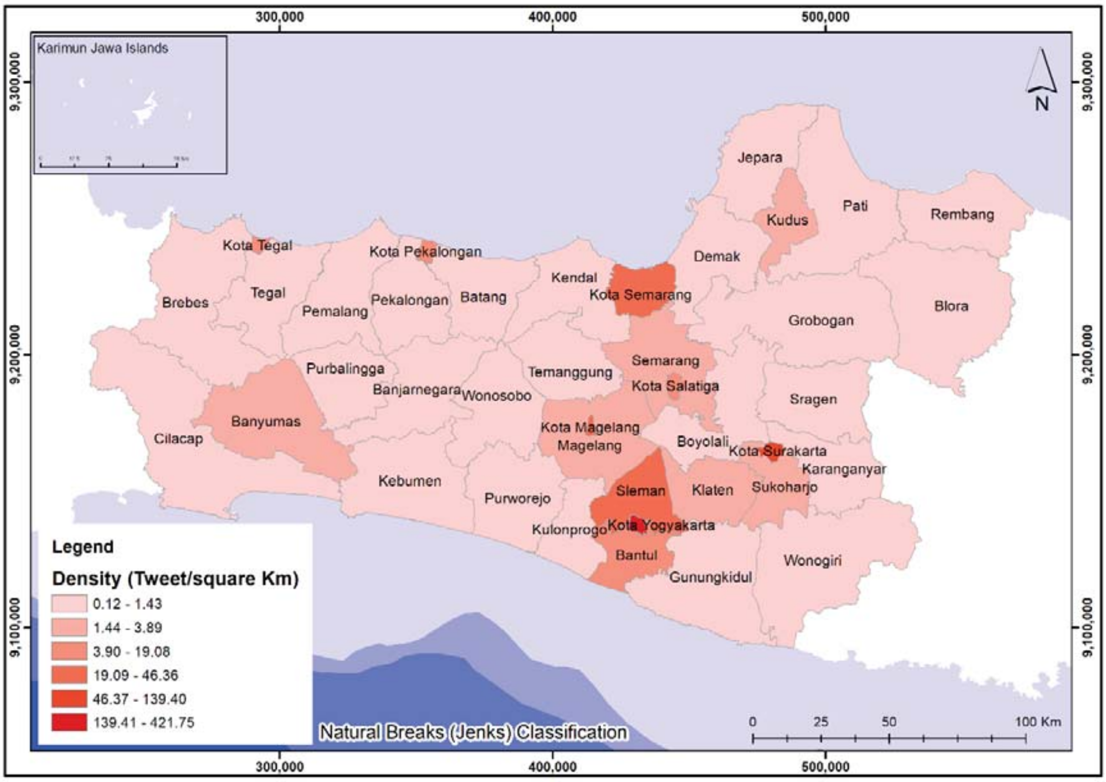


Figure 5. The global density of each Regency.

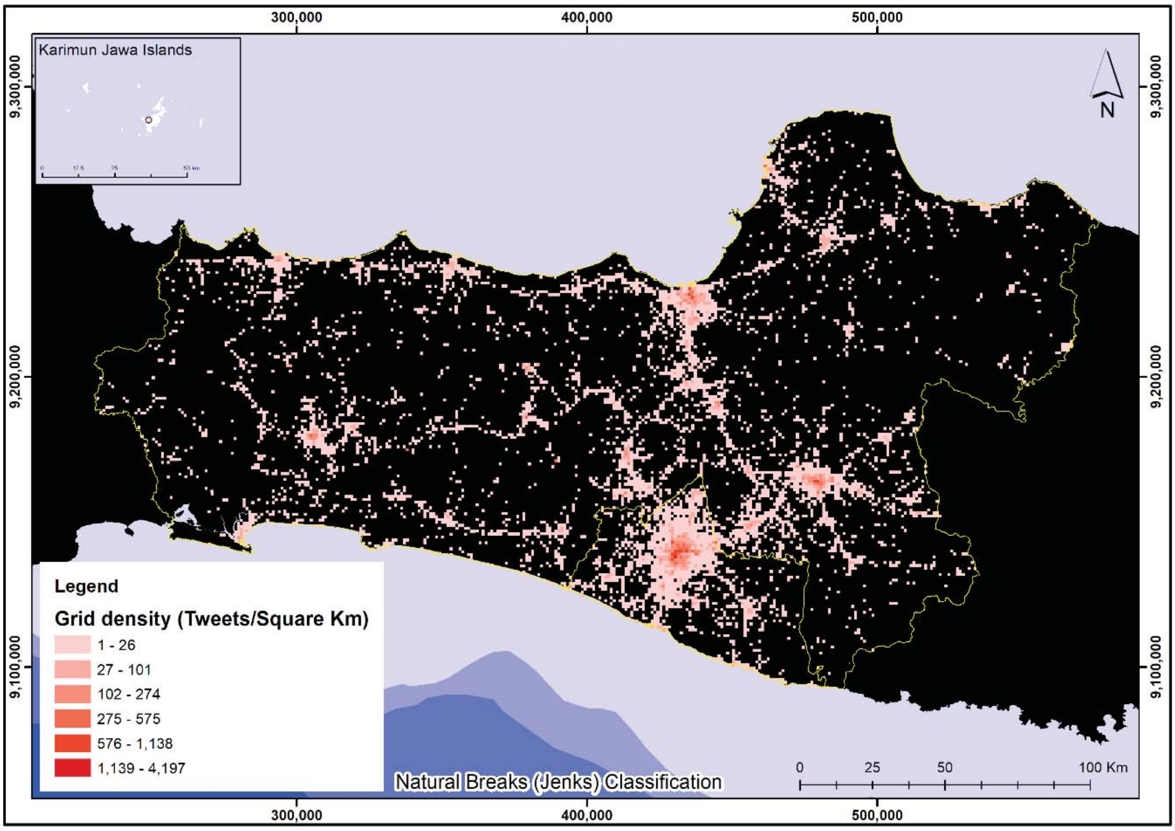


Figure 6. The global density of each Regency.

Each polygon contains an average of 13.5 tweets with a standard deviation value of 76.16. However, the 4,196 data range is very wide. This result indicates the emergence of inequality within the study area. It is interesting to investigate the factors further. For example, in addition to infrastructure factors, as previously thought, the tweet-making behaviour of Twitter users also has an influence on the creation of data patterns.

4.4. Tweet Proximity Index (TPI)

Based on the results of the point distance analysis, a TPI was developed which stated the average distance and the amount of data within a predetermined radius. In general, the TPI value ranges from 0 and 1.35 with a mean of 0.58 and a standard deviation of 0.26. This relatively poor result is due to the significant difference in the number of tweets (Figure 7). Based on the results of the distance index calculation, the calculation of the average distance is not consistent because each tourist attraction has a different number of tweets. This condition is advantageous to tourist attractions that have relatively close tweet distances and a small number of tweets.

The distribution of TPI values in the study area is less affected by the density pattern discussed in the previous section. The TPI classes are distributed equally in the west and centre of the study area, despite the relatively low tweets density value. An exception is the Karimun Jawa National Park (TNKJ), although this cannot be included in the calculation of the TPI value as the distance to the nearest tweets is 4 km, which exceeds the search radius limit of only 1 km. This result is not unexpected as the access to cellular networks in the Karimun Islands is not as good as in Java Island, thus limiting the movement of social media users.

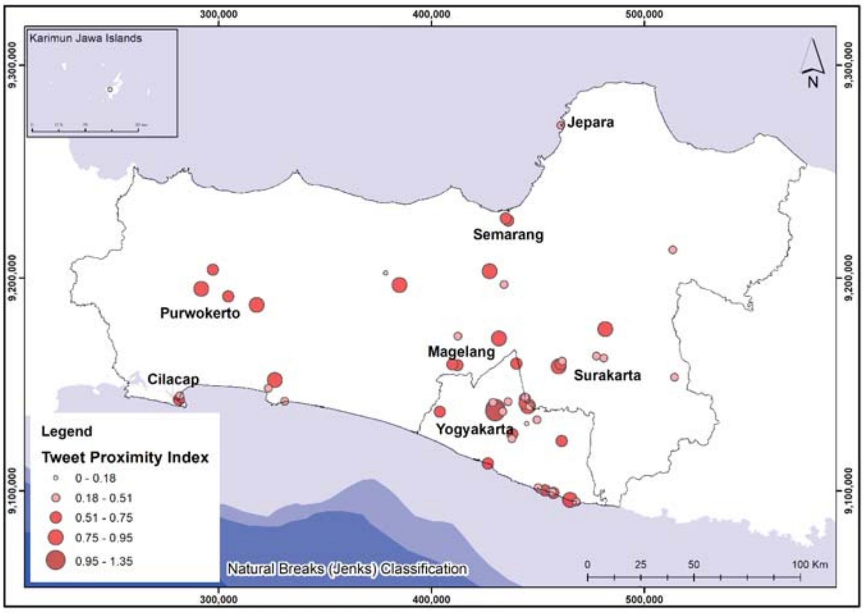


Figure 7. Tweet Proximity Index in the study area.

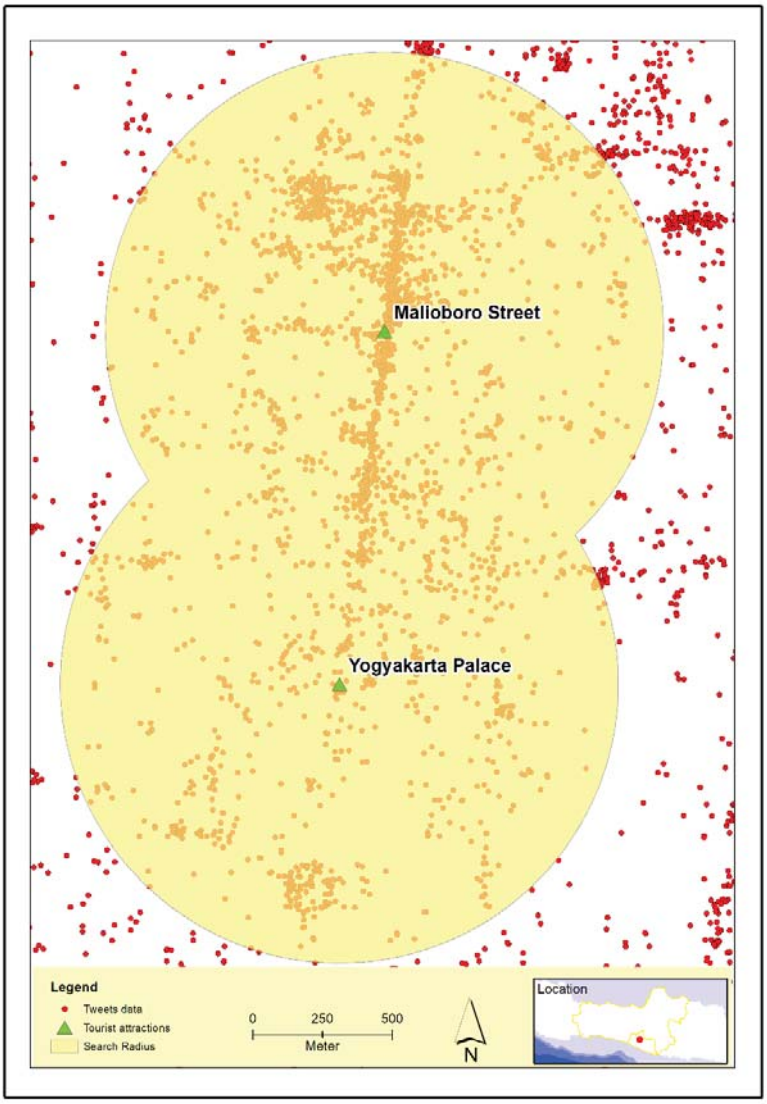


Figure 8. Distribution of tweets within a 1 km radius of Malioboro Street and Yogyakarta Palace.

Kaligua Beach and Siung Beach have the smallest average distance values of 56.24 m and 78.11 m, respectively. However, these values were derived from only a very small number of tweets, with only one tweet for Kaligua Tourism. As an impact, the TPI value for these two attractions is in the very high category. In contrast, Pahlawan Street, with 1,287 tweets, is ranked 15th since the average distance reached 516.5 m.

The results of the TPI calculations show that Malioboro Street and Yogyakarta Palace are the two tourist locations with the highest index values, registering 1.35 and 1.31, respectively. The number of tweets within a 1 km radius from these two points is indeed very large and displays a longitudinal pattern along Malioboro Street (Figure 8). The second location in the centre of Yogyakarta offers easy transportation and accommodation for tourists. There is also a wide variety of tourist attractions, thus making it easier for tourists to access an all-in-one destination. Repairs to the quality of the pedestrian route on the east side of Jalan Malioboro has attracts more tourists.

4.5. Density Index (DI)

The results of the point density calculation present different things from the TPI calculation as the quantification of tweets is also calculated based on the density in neighbouring cells. The density values of points at the study location itself range from 0 to 948.11 tweets/km^2. The distribution of density values is similar to the density patterns shown in Figures 4 and 5, with differences seen in the tourist attractions in the City of Magelang, which have a relatively low density (see Figure 9). Semarang and Surakarta City each have a tourist attraction with a comparatively high density value, while Yogyakarta City has the highest value.

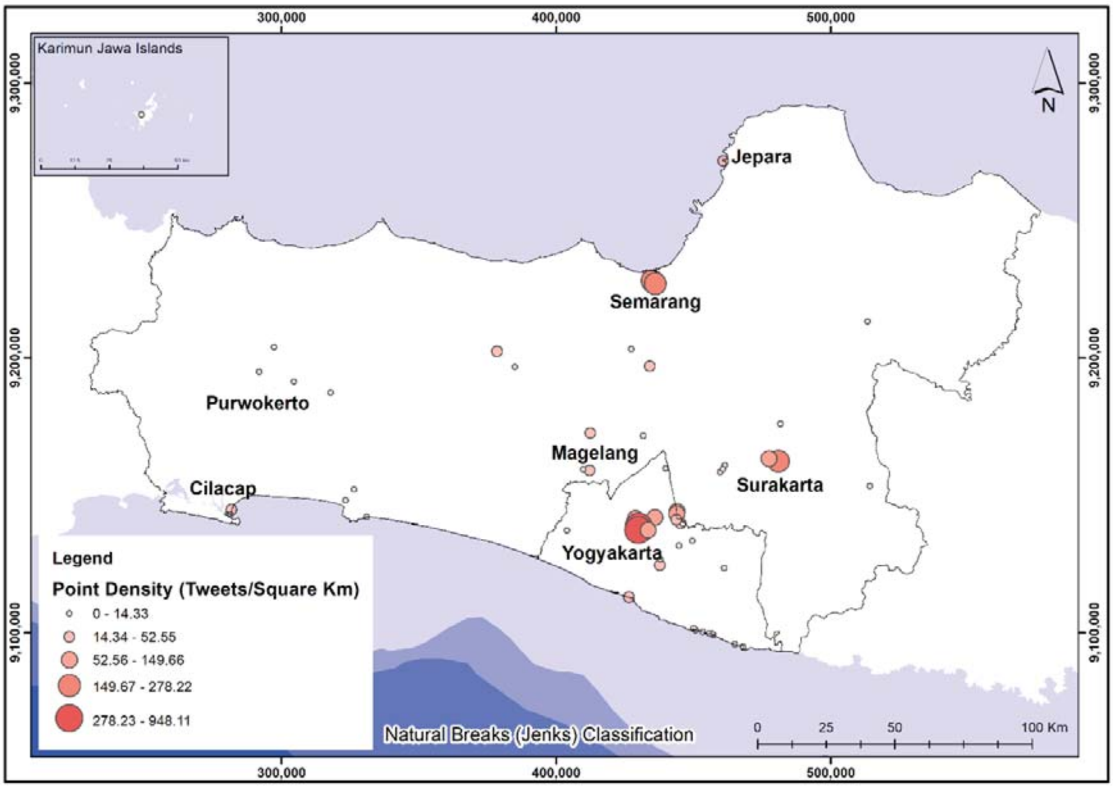


Figure 9. The result of point density analysis.

Aside from the traditional tourist attractions, several other tourist attractions have in recent times rapidly gained in popularity through social media. However, the new social media tourism sensation has not resulted in a high tweet density. Among others, Mangunan Pine Forest is the most popular and has the highest density value (23.33 tweets/km^2). Meanwhile, Amaryllis Flower Park scored the lowest density value of only 3.56 tweets/km^2. This result is quite reasonable since at the time of the study the amaryllis was not in its flowering phase and thus an attraction based on it was not regarded as a high priority for tourists to visit. Without any certain waiting or peak period, Mangunan Pine Forest is attractive to visitors. Elements such as this must be considered by the management of tourist attractions when looking at increasing the number of visitors to tourist attractions, although each tourist attraction already has its own characteristic.

The Kalibiru tourism objects that rose in 2016 through social media recorded only 10.78 tweets/km^2. This result is far below Prangtritis Beach (20.44 tweets/km^2), which remains one of the most popular tourist destinations in the Special Region of Yogyakarta. The ease of finding tourist sites may need to be improved. In the digital era, the location of tourist attractions can be included on a digital map for searching by potential tourists, including the route to take to access the attraction.

Borobudur Temple, which is an iconic tourist attraction, has a high tweet density (42.33 tweets/km^2). At the time of the field survey, Borobudur Temple was seeing large numbers of tourists as they generally visit in groups. Not far from Borobudur Temple is *Gereja Ayam*, which has become a tourist attraction and social media sensation. The results of the density calculations show that *Gereja Ayam* has a value of 11.33 tweets/km^2.

4.6 Survey Result

The questionnaire survey was started after the Twitter data mining had been completed, with a total of 159 respondents from various backgrounds. The identity section of the survey contained general questions regarding the respondents' personal data and social media accounts. The interesting finding here was that Instagram was the social media with the highest appeal, as shown by the fact that many respondents use it, with an activity level of 90.73%. Meanwhile, Twitter was ranked third, below Facebook (Figure 10). Despite having a relatively large number of users, the respondents' level of activity in using their Twitter account stood at only 36.59%. Based on this, a high number of social media users does not always equate to a high level of user activity. Social media platforms must certainly have strategies in place aimed at increasing their user activity since a lack of data from users diminishes the power of social media as an alternative data provider.

Nowadays, a range of devices can be used to access social media activities. Those devices, whether mobile or not, are now widely enjoyed by people due to their tremendous market penetration in recent years. Among other devices, the smartphone is the foremost choice among respondents when interacting with the community through social media. This type of social media activity opens up the possibility of geolocation/geotagged data being available because smartphone devices are generally equipped with GPS receivers that can be activated/deactivated.

The second part of the questionnaire contained questions enquiring about the respondents' Twitter accounts, if indeed they had any. The aim was to capture the behaviour patterns of Twitter users in Indonesia, at least in terms of what the respondents indicated. This section was not compulsory for all respondents to complete as not all of the respondents had a Twitter account. The low performance of Twitter users raised in the previous discussion is supported by the data in this section. The pink colour grouping in the upper-left corner of Table 1 generally indicates that during the time when there is a high frequency of tweet creation, we also see a decrease in the average number of tweets made by the user. The highest values from this data indicate that three respondents made six tweets per day, with the last tweet made less than 24 hours ago. However, this equates to only 2.5% of all respondents who completed the questionnaire.

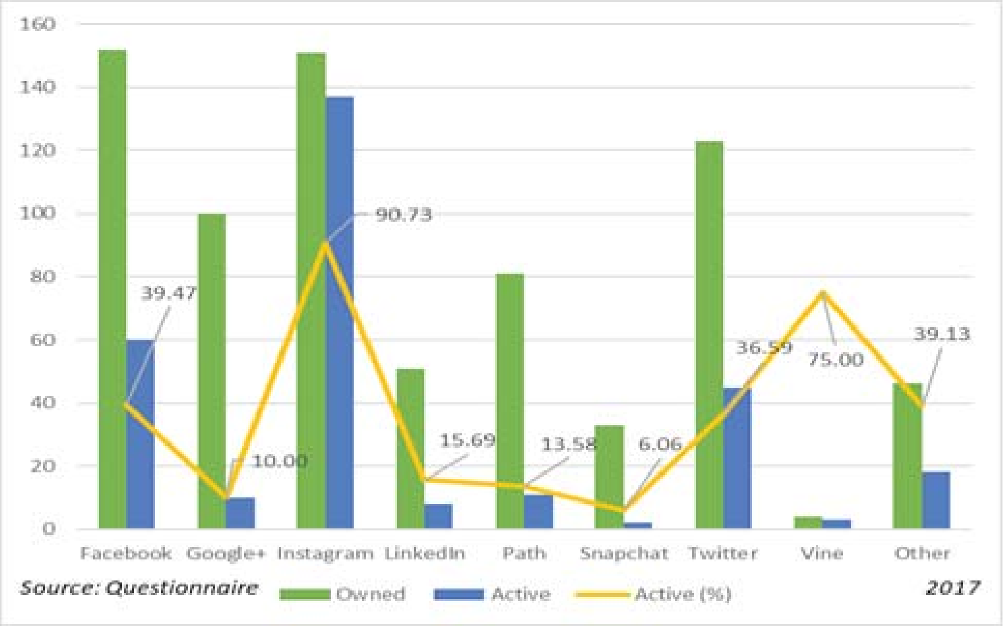


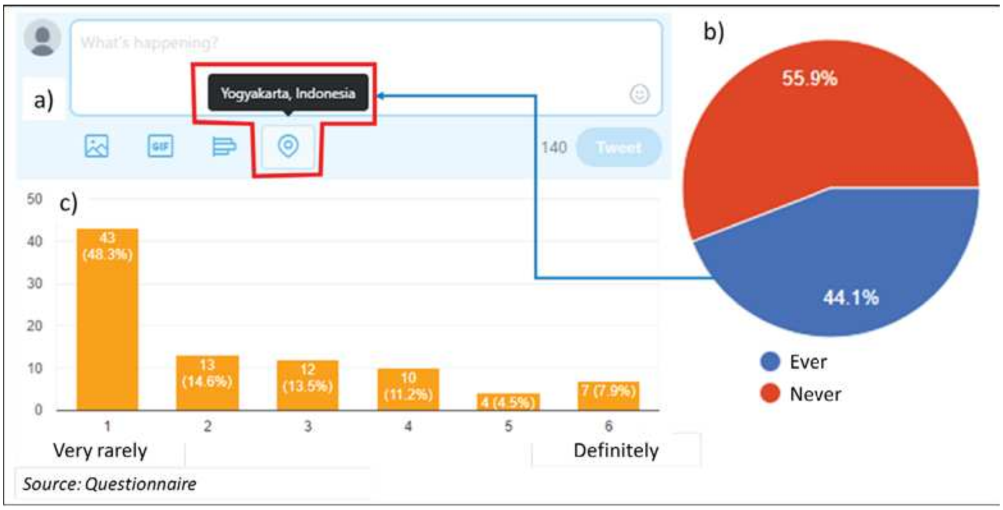
Figure 10. The number of social media accounts owned by respondents and their levels of usage.

Table 1. The relationship between the average number of tweets made in one day and the time when the last tweet was made.

		Latest tweets				
		< 24 hr	1-2 days	2-7 days	1-2 weeks	> 1 month
Average tweets in one day	< 1	1	8	4	7	50
	1	4	4	1	0	4
	2	5	1	1	0	5
	3	3	1	1	1	1
	4	1	0	0	0	2
	5	2	1	2	2	1
	6	3	0	0	0	0
	7	0	0	0	0	1
	8	0	0	0	0	0
	9	0	0	0	0	0
	10	0	0	0	0	0
	> 10	0	0	0	2	2

Source: Questionnaire

The upper-right corner of Table 1 displays proof of the decreasing activity level of the Twitter user. A majority of the respondents who indicated that they had not posted a tweet for more than one month had also posted less than an average of one tweet per day. Overall, only 15.7% of tweets had been made within the previous 24 hours, which was far below the figure for tweets made more than one month ago, which stood at 54.54%.



Source: Questionnaire

Figure 11. a) Add location information to a tweet; b) survey results related to the adding of location information to Twitter; c) survey results related to the frequency of adding location information to Twitter.

The feature of adding a location to a tweet will further increase the chances of data being created that include coordinates. The option to add location information is presented every time a user makes a tweet (Figure 11a). However, since it is optional, not all users will opt to show their location. According to the respondents, only 44.1% had ever used this feature within their Twitter account (Figure 11b). A more detailed look at the data shows that of the respondents who had used the location feature in Twitter, only 7.9% always activated it, while 48.3% of the respondents very rarely used it (Figure 11c).

The results of the survey indicate that there is the possibility to acquire Twitter data that contains location information in only limited quantities. This finding confirms other studies which reveal that only 0.71% of all tweets in Indonesia were geotagged (Carley *et al.*, 2015).

The third part of the questionnaire presented questions related to tourism in Central Java and the Special Region of Yogyakarta. Most of the respondents favoured nature tourism activity. The data also show that 30% of the respondents undertake tourism activities more than ten times per year. The fourth part of the questionnaire looked at the relationship of social media with tourism. Based on the data, many respondents obtained tourism information through social media, followed by information from friends and web pages respectively. As a form of media that benefits from a relatively quick speed of data transfer, social media is indeed an effective and efficient form of promotional media. The past few years have seen the sudden emergence of new famous tourist spots after they have gone viral on social media.

As mentioned above, there is little probability of geotagged tweets being created by the user. However, looking at the survey data, 17.7% of respondents answered that they had added location information to tourist attractions. Thus, among the various data contained on Twitter, it still offers the potential for use in tourism research.

4.7. Popularity

The popularity of tourist attractions was assessed by comparing the results from the questionnaire with the TPI calculation and DI. The assessment involved data from a total of 17 tourist attractions. The tourist attractions were selected based on the results of the respondents' choice of favourite, with a minimum of 2 voters required. Appendix 1 presents a comparison of the popularity of attractions based on the three above-mentioned elements.

The TPI, despite appearing to be overestimated, turns out to have a greater accuracy than the DI, although with a very weak difference. The accuracies of the calculation indexes were 76.47% and 58.82%, respectively. These accuracy values are quite high considering that the data used in the calculation of TPI and DI were unfiltered by Twitter content. If the raw data processed correspond with the purpose of the mapping, then the result of the index analysis is expected to be able to provide a higher level of accuracy.

Analysis of non-geotagged data is needed for exploration because the volume of data on the server is much higher than the geolocated data. The lack of access to official and easily accessible data on tourist numbers also acts as an impediment to testing accuracy in this study. If data on the number of tourists can be acquired at the same time as the data mining is carried out, then an accuracy assessment can be conducted more precisely. However, a lack of tourist categorisation will make the analysis much more difficult as the analysis will include a large volume of tourist data.

5. Conclusion

Geolocated tweet data can be accessed using the Public Streaming API via Python scripts and the Tweepy module. Queries can be performed by determining a search location or by keyword. The wider the search area, the more quickly data can be queried as it increases the opportunity for capturing the tweets. The results derived from the query data can be utilised for mapping activities, especially thematic mapping. Many themes can be developed based on tweets from Twitter users.

Two approaches were used in this study to analyse the popularity of tourist attractions, namely the Tweet Proximity Index (TPI) and density index (DI). Neither approach delivered a satisfactory level of accuracy.

Further exploration is needed of both index drafting methods and the examination of non-geotagged tweet data. It is also interesting to study data from Instagram, which currently has the highest percentage of activity among other social media platforms in Indonesia.

Acknowledgements

This research was conducted with funding from the Faculty of Geography, Universitas Gadjah Mada through the Lecturer Research Grant. The author is grateful for the opportunity to carry out this research through the grant scheme.

References

Aggarwal, C. (2015). *Data mining: The textbook*. Cham: Springer International Publishing.

Baiquni, M. (2009). Belajar dari Pasang Surut Peradaban Borobudur dan Konsep Pengembangan Pariwisata Borobudur. *Forum Geografi, 23*(1), 25-40. https://doi.org/10.23917/forgeo.v23i1.4997

Borne, K. (2014). Top 10 big data challenges a serious look at 10 big data V's. [Online]. Accessed: 20/4/2017. Retrieved from: https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/

Carley, K., Malik, M., Kowalchuk, M., Pfeffer, J., & Landwehr, P. (2015). *Twitter usage in Indonesia*. Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA.

Carto (2017). *Connecting Twitter data* [Online]. Accessed: 23/4/2017. Retrieved from: https://carto.com/learn/guides/data-and-sql/connecting-twitter-data

Fauzi, F., Harly, G. S., & Hanrais, H. S. (2012). Analisis penerapan teknologi jaringan LTE 4G di Indonesia. *Majalah Ilmiah Unikom, 10*(2), 281-288.

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *Magazine Interactions, 19*(3), 50-59.

Frank, M. R., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2013). Happiness and the patterns of life: A study of geolocated tweets. *arXiv preprint*. arXiv:1304.1296.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature*, *453*(7196), 779.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260-271.

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045*.

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a collective sense of place. *PloS one*, *11*(4), e0152932.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In System Sciences (HICSS), *2013 46th Hawaii International Conference on System Sciences*. pp. 995-1004. IEEE.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and good practices. In Contemporary Computing (IC3), *2013 Sixth International Conference on Contemporary Computing*. pp. 404-409. IEEE.

Kraak, M. J., & Ormeling, F. J. (2013). *Cartography: visualization of spatial data*. Routledge.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *META Group* [Online]. Accessed: 21/4/2017. Retrieved from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lwin, K. K., Zettsu, K., & Sugiura, K. (2015). Geovisualization and correlation analysis between geotagged Twitter and JMA rainfall data: Case of heavy rain disaster in Hiroshima. In *2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, 71-76. IEEE.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal, 80*, 449-461.

O'Reilly, T., & Milstein, S. (2012). *The Twitter book*. Sebastopol, CA: O'Reilly Media, Inc.

Rijmenam, M. V. (2013). Why the 3v's are not sufficient to describe big data. *BigData Startups* [Online]. Accessed: 20/4/2017. Retrieved from: https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Setiawan, B., Rijanta, R., & Baquni, M. (2017). Sustainable tourism development: The adaptation and resilience of the rural communities in (the tourist villages of) Karimunjawa, Central Java. *Forum Geografi, 31*(2), 232-245.https://doi.org/10.23917/forgeo.v31i2.5336

Setkab (2017). Tahun 2017 Kita Genjot Sektor Pariwisata [Online]. Accessed: 23/4/2017. Retrieved from:http://setkab.go.id/tahun-2017-kita-genjot-sektor-pariwisata/

Statista (2016). Number of active Twitter users in leading markets as of May 2016 (in millions) [Online]. Accessed: 23/4/2017. Retrieved from: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks, 34*(1), 73-81.

Thatcher, J. (2014). Big Data, big questions | Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication, 8*. p.19.

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: A survey. *Journal of Big Data, 2*(1), 21.

Wibowo, T. W. (2017). Spatial point data analysis of geolocated tweets in the first day of Eid Al-Fitr 2017 in Java Island. Proceedings of the 5th Geoinformation Science Symposium. Yogyakarta.

Yin, J., & Du, Z. (2016). Exploring multi-scale spatiotemporal Twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information, 5*(10), 187.

Appendix

Appendix 1. Comparison of the popularity of tourist attractions based on survey results, Tweet Proximity Index (TPI) and Density Index (DI).

No	Tourist Attraction	Vote	Tourist Attraction	TPI	Tourist Attraction	DI
1	Dieng Plateau	32	Malioboro Street	1.35	Malioboro Street	1.00
2	Malioboro Street	28	Yogyakarta Palace	1.31	Yogyakarta Palace	0.94
3	Borobudur Temple	13	Sangiran Museum	0.90	Lawang Sewu	0.29
4	Nglanggeran Ancient Volcano	12	Ketep Pass	0.82	Pahlawan Street	0.27
5	Lava Tour Merapi Volcano	8	Breccia Cliff Park	0.80	Surakarta Palace	0.26
6	Parangtritis Beach	8	Parangtritis Beach	0.75	Gembira Loka Zoo	0.16
7	Baturaden	7	Pahlawan Street	0.72	Kampung Batik Laweyan	0.15
8	Ketep Pass	7	Kalibiru Tourism Village	0.71	Sindu Kusuma Edupark	0.14
9	Karimunjawa National Park	6	Borobudur Temple	0.69	Upside Down World	0.12
10	Yogyakarta Palace	4	Pindul Cave	0.69	Borobudur Temple	0.04
11	Lawang Sewu	4	Baturaden	0.69	Breccia Cliff Park	0.03
12	Beaches in Gunungkidul	4	Lawang Sewu	0.66	Dieng Plateau	0.03
13	Rafting in Progo River	3	Indrayanti Beach	0.60	Rafting in Progo River	0.03
14	Kalibiru Tourism Village	2	Lava Tour Merapi Volcano	0.57	Train Museum	0.02
15	Pindul Cave	2	Drini Beach	0.57	Parangtritis Beach	0.02
16	Gembira Loka Zoo	2	Krakal Beach	0.51	Baron Beach	0.01
17	Siung Beach	2	Train Museum	0.47	Kukup Beach	0.01
	Accuracy		76.47%		58.82%	

The above table contains the three popularity assessments of the vote results from the survey, TPI and DI respectively. The number of tourist attractions was adjusted based on the results of the voting; in this case, there are 17 tourist attractions. The results from the TPI and DI calculations were also sorted and the same number were taken for comparison with the voting result. The colour represents the correspondence between voting data and TPI/DI data.

Yellow: corresponds to TPI and DI
Green: corresponds only to TPI
Orange: corresponds only to DI.