International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org, editorijettcs@gmail.com
Volume 7, Issue 2, March - April 2018 ISSN 2278-6856

Spatial Analysis of GIS Data for Social Media Data Disaster Management

Neelima Gaur[1], Farheen Fauziya[2], Ajit Kumar Jain[3]

[1,3]Banasthali Vidyapith Jaipur, Rajasthan, 304022, India

[2] Indian Institute of Technology, New Delhi, 110016, India

Abstract : ***Today ease access of computers and modernization leads to increased Internet users day by day. Social media are a medium to communicate information to the public while Geographical Information System (GIS) is a technique, which focuses on communication of geographical information. The fact is this GIS communicates geographical information stored in digital form. This paper focuses on one of application of geographic data of popular social media platform Twitter users by generating heat map with Quantum Geographic Information System (QGIS) tool, supported by GIS technology. Which can also use to calculate the user density within a particular region to get useful information. The same method can be apply for other social media platforms as Facebook, LinkdIn, Pinterest, Instagram etc. to inferred meaningful results.***

Keywords: Analysis, GIS, Heat Map, Point Density, QGIS, Social media, Spatial data, Twitter, Visualization.

1. Introduction:

There is a growing number of social media users day by day. It is so because using social websites people communicate or connect with each other and give the real-time information throughout the world [10]. Tweets or tagged data having GPS coordinates with user's consent using any GPS enabled device such as smart phones, which offers users geo-location when they tweets. According to current statistics there is 500 million tweets per day. Which is a source of large data production of user-generated contents. As per the current statistics there are 4,156,932,140 Internet users throughout the world and 2.62 billion social media users in 2018 and it is also estimated that it may rise to 3.02 billion up to 2021[12][13]. The speed at which data is increasing, it seems like the problem of data disaster is arising. This huge data exposed us towards new perspective i.e. its use for various applications. As text data can be used to do sentiments analysis while spatial data have other applications as to find out twitter user density in specific area by mapping. This application, mapping of spatial data can compare with the traditional cartography process. The only difference between cartography maps are often simplifies as there are limits to the amount of data that can physically and meaningfully stored on a small map. The spatial data for this research work collected using GNIP and Restful State Transfer Application Programming Interface (REST APIs), which provides data is in JSON (Java Script Object Notation) format.

So this paper mainly focused on the following sections; in section–1, relevant data, we need to perform data cleaning for available data for various analysis using GIS tool i.e. QGIS and store it into relevant database. In section –2, methodology to explain map projection of this available Twitter data and the methods involved the entire procedure to get desired results. In section-3, (results) creation of heat map for this data model and calculation of Twitter user density within a specific region. Last section 4-includes conclusion and future work. The abbreviations used in this paper are given in Table-1.

In 1988, the National Science Foundation established the National Center for Geographic Information and Analysis (NCGIA) as a catalyst for the acknowledgement to number of challenges facing the rapidly growing geographic information science community. The various institutes, commercial firms, and teaching programs engaged in GIS activities. Identifying this requirement, on several occasions NCGIA proposed and accomplished initiatives focusing on the problem of dealing with spatial analysis in a GIS environment [7].

The role of location for spatial data is crucial, both in an absolute sense (coordinates) and in a relative sense (spatial arrangement, distance). Indeed, location has spatial effects in two ways: spatial dependence and spatial heterogeneity. Where spatial dependence, often also referred to as spatial autocorrelation or spatial association. According to First Law of Geography, "everything is related to everything else, but near things are more related than distant things." As a consequence, similar values for a variable will tend to occur in nearby locations, which leads to spatial clusters. While the second type of spatial effect i. e. spatial heterogeneity, includes the spatial or regional differentiation, which follows from the intrinsic uniqueness of each location [1].

Table-1: Abbreviations

CSV	Comma Separated Values
GIS	Geographic Information System
JSON	Java Script Object Notation
NCGIA	National Center for Geographic Information and Analysis
QGIS	Quantum Geographic Information System
REST API	Restful State Transfer Application Programming Interface
SNS	Social Networking Sites

In context of social media (e.g. Twitter), geo-tagged data has two major perspectives: First, users can geo-tagged (which are basically geographic locations of users) their messages with precise location in the form of latitude, longitude means the exact position where the user was when the message was recorded using latitude, longitude measurements. Second, profile based geo-tagged data can tell about other things of users like where they were born, their educational places, lived, employed etc. [4].

In 2010, Facebook launched a service called "Places" which allow its users to share their point of interest about their physical locations. Only a small percentage of Facebook users with increasing use of social media and mobile devices, that shared content (i. e. text messages, photos etc.) has geographical identification added called "geo-tagged". Past research has mainly focus to analyze the location data collected from either indirect sources of location or via small-scale experiments on users while the work on explicitly provided location data has been limited as well [3].

Social networking sites (SNS) are used for social and professional interaction among public. The growing data for SNS, has encouraged to analyze the relationship of activities performed on SNS with user locations. Therefore, this work is aimed to identify and analyze the characteristics associated with social media user locations to answer the research questions designed to conduct this research. A mapping study (also called scoping study), which is a type of systematic literature review, is appoint to identify potential studies from digital databases through a developed protocol [5].

1.1 Relevant Data Set:

The spatial data of Twitter has been collected using GNIP, which provide access to online data through REST API. It is a commercial service and world's largest social data provider. The process of collection of the online data of social site is known as 'scraping'. User provided metadata (e.g., location and time zone) is readily accessible in the tweet JSON objects. This metadata can be appended as extra text features, in addition to features derived from tweet text. That data is in JSON format i.e. the key-value pairs are there. It is needed to perform the "Data Cleaning" process to fetch some certain fields from this huge amount of data (i.e. geographical information or geo-coordinates of Twitter users) as latitude, longitude values of the locations of a region, in which they are highly dense and to maintain the relevant database to store the data.

In order to a deep examination of Twitter user density of all the locations of Twitter users, we consider all the geo-locations (i.e. latitude, longitude) database. The available data is in given format in Fig.1.

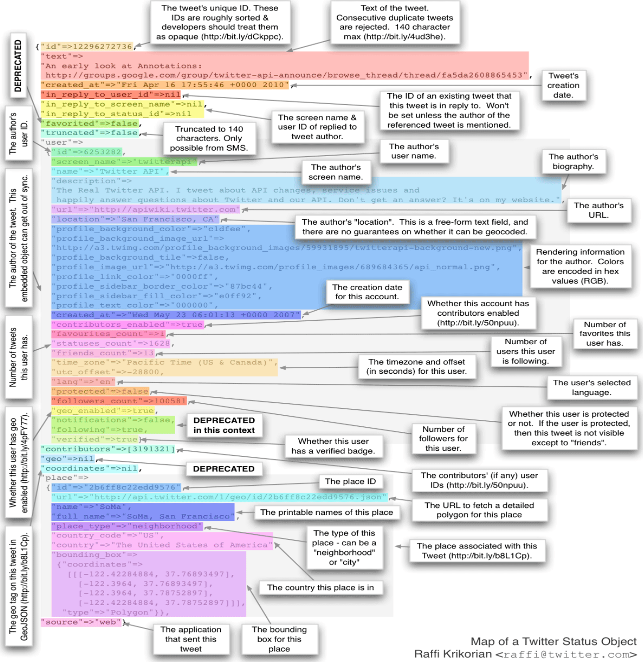
To perform data cleaning process, it is required to remove all the noisy data (irrelevant data which is not required for this analysis) from the available data set. For this, we have to write a code using any programming language (as java, python) for fetching some specific fields as latitude, longitude and time from the given data set and store it into text file. MYSQL database is used to read that text file and finally run a query to get locations of Delhi region only.

2. Methodology:

2.1 Data Model (Mapping):

QGIS 2.14 Essen is open source software of GIS, which provide various interesting functionality to make it possible to visualize and analyze the spatial data of Twitter users and get useful results to support scientific research. To import data into QGIS, it must be in Comma Separated Values (CSV) format because QGIS only supports CSV format.

QGIS 2.14 supports, plugins which explore its functionalities. There is two type of plugins - core plugin and external plugin. The plugins, which are already available in standard QGIS installation are "Core Plugin" while the plugin which is available in QGIS plugin repository but needed to install(as per application) before using them are called "External Plugin". OpenLayer plugin (i.e. an external plugin) supports the open streets map which is taken as base map here and import CSV file of geo-coordinates having latitude and longitude values, which are already stored within relevant database, to project the location of the twitter users over the base map.



```
{"id"=>12296272736,
 "text"=>
 "An early look at Annotations:
 http://groups.google.com/group/twitter-api-announce/browse_thread/thread/fa5da2608865453",
 "created_at"=>"Fri Apr 16 17:55:46 +0000 2010",
 "in_reply_to_user_id"=>nil,
 "in_reply_to_screen_name"=>nil,
 "in_reply_to_status_id"=>nil
 "favorited"=>false,
 "truncated"=>false,
 "user"=>
  {"id"=>6253282,
   "screen_name"=>"twitterapi",
   "name"=>"Twitter API",
   "description"=>
   "The Real Twitter API. I tweet about API changes, service issues and
    happily answer questions about Twitter and our API. Don't get an answer? It's on my website.",
   "url"=>"http://apiwiki.twitter.com",
   "location"=>"San Francisco, CA",
   "profile_background_color"=>"c1dfee",
   "profile_background_image_url"=>
   "http://a3.twimg.com/profile_background_images/59931895/twitterapi-background-new.png",
   "profile_background_tile"=>false,
   "profile_image_url"=>"http://a3.twimg.com/profile_images/689684365/api_normal.png",
   "profile_link_color"=>"0000ff",
   "profile_sidebar_border_color"=>"87bc44",
   "profile_sidebar_fill_color"=>"e0ff92",
   "profile_text_color"=>"000000",
   "created_at"=>"Wed May 23 06:01:13 +0000 2007",
   "contributors_enabled"=>true
   "favourites_count"=>1
   "statuses_count"=>1628,
   "friends_count"=>13
   "time_zone"=>"Pacific Time (US & Canada)",
   "utc_offset"=>-28800,
   "lang"=>"en"
   "protected"=>false
   "followers_count"=>100581
   "geo_enabled"=>true,
   "notifications"=>false,
   "following"=>true,
   "verified"=>true},
 "contributors"=>[3191321],
 "geo"=>nil,
 "coordinates"=>nil,
 "place"=>
  {"id"=>"2b6ff8c22edd9576",
   "url"=>"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json"
   "name"=>"SoMa",
   "full_name"=>"SoMa, San Francisco",
   "place_type"=>"neighborhood"
   "country_code"=>"US",
   "country"=>"The United States of America"
   "bounding_box"=>
    {"coordinates"=>
      [[[-122.42284884, 37.76893497],
        [-122.3964, 37.76893497],
        [-122.3964, 37.78752897],
        [-122.42284884, 37.78752897]]],
     "type"=>"Polygon"}},
 "source"=>"web"}
```

DEPRECATED

The author's user ID.

The author of the tweet. This embedded object can get out of sync.

Number of tweets this user has.

Whether this user has geo enabled (http://bit.ly/4pFY77).

The geo tag on this tweet in GeoJSON (http://bit.ly/b8L1Cp).

The tweet's unique ID. These IDs are roughly sorted & developers should treat them as opaque (http://bit.ly/dCkppc).

Text of the tweet. Consecutive duplicate tweets are rejected. 140 character max (http://bit.ly/4ud3he).

Tweet's creation date.

The ID of an existing tweet that this tweet is in reply to. Won't be set unless the author of the referenced tweet is mentioned.

The screen name & user ID of replied to tweet author.

Truncated to 140 characters. Only possible from SMS.

The author's user name.

The author's screen name.

The author's biography.

The author's URL.

The author's "location". This is a free-form text field, and there are no guarantees on whether it can be geocoded.

Rendering information for the author. Colors are encoded in hex values (RGB).

The creation date for this account.

Whether this account has contributors enabled (http://bit.ly/50npuu).

Number of favorites this user has.

Number of users this user is following.

The timezone and offset (in seconds) for this user.

The user's selected language.

Whether this user is protected or not. If the user is protected, then this tweet is not visible except to "friends".

Number of followers for this user.

DEPRECATED in this context

Whether this user has a verified badge.

DEPRECATED

The contributors' (if any) user IDs (http://bit.ly/50npuu).

The place ID

The URL to fetch a detailed polygon for this place

The printable names of this place

The type of this place - can be a "neighborhood" or "city"

The place associated with this Tweet (http://bit.ly/b8L1Cp).

The country this place is in

The application that sent this tweet

The bounding box for this place

Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Fig. 1 Available data format of Twitter users over which data cleaning needs to be performed. [16]

In mapping process, Coordinate Reference System (CRS) is a way to define how the two dimensional projected map is related to real places on this earth. The base map already has WGS 84 / Pseudo Mercator CRS. Sometimes, it is needed to manually set the CRS for particular layer into QGIS. The available data of Twitter users for 15 days, to show the active Twitter users within a particular region. By visualizing the map it is clear that Twitter users were active at the locations which are shown by dark circles. The final map projection will be like Fig 2(a).

The same analysis can be done using other tool box available in ArcGIS, a desktop application. The map projection can be seen in the above Fig. 2(b).

2.2 Heat Map:

If there is a large volume of data and a number of overlapping points, heat Maps are used to easily identify the clusters of higher concentration of activity or it is the occurrence of a set of points to represent the density of points over a surface. It is a graphical representation of any data using color coding method. Where each color shows the clusters of higher concentration of activity or it is the occurrence of a set of points to represent the density of points over a surface. It is a graphical representation of any data using color coding method. Where each color shows a particular value from low-density value (cool) to high-density value (hot) or vice versa as shown in fig. 3. So heat map can be a good application here because there are various overlapping points due to twitter feeds from the same geo-location (when the user is in stationary condition).

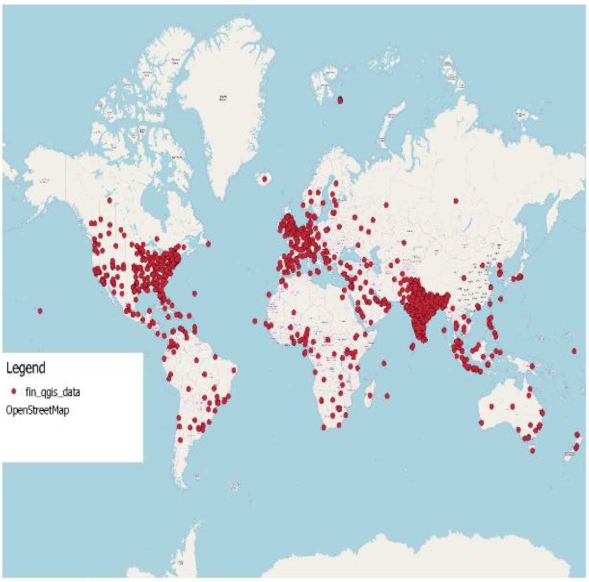


Fig. 2(a) Map projection of twitter users using QGIS

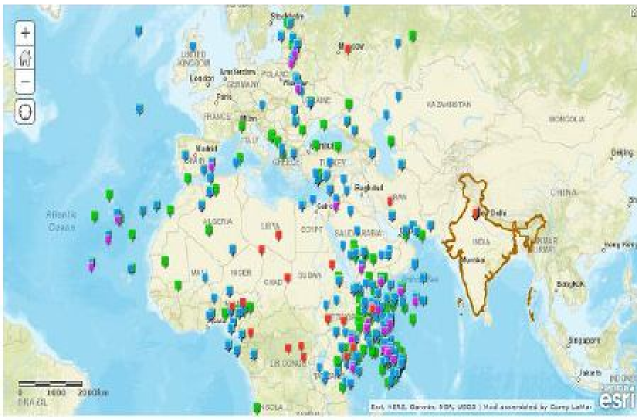


Fig. 2(b) Map projection of twitter users using ArcGIS

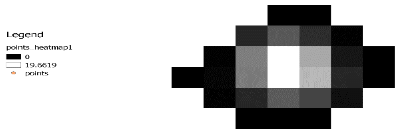


Fig 3: Heat map of point data

2.3 Point Density Calculations:

Density analysis receives known quantities of any phenomenon and spreads them across the background based on the measured quantity at each location and the spatial relationship of the locations of the measured quantities.

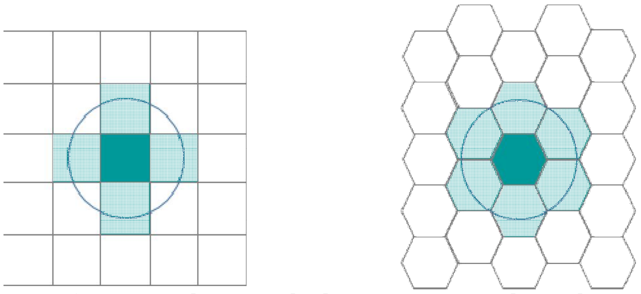


Fig.4 Point density calculation in rectangular and hexagonal grids [14]

To calculate twitter user density within the specific region (i.e. Delhi here), it is required to calculate the number of Twitter users within the unit area. Here twitter users are represented by point layer. First, the whole region is divided by grid layer as shown in Fig. 6(a). But it is better to choose the hexagonal grid shape because it has equal length of its edges and hexagonal centroids are equidistant. So it is more straightforward to find neighbors in the hexagonal grid, which can be easily understand in Fig. 4.

For this calculation, MMQGIS plugin is used i.e. an external plugin, already present in QGIS plugin repository but need to install before it use. Import the point layer of geo-locations of Twitter users and create a grid on it (It is rectangular and hexagonal grids in this work in Fig. 6(a) and 6(b)), then apply point count polygon algorithm to get the point density within particular cell of a grid.

The user density is calculated with the given formula in equation (1).

$$no.of\ users\ within\ a\ neighbourhood\ (User\ density) = \frac{total\ no.of\ points\ that\ fall\ within\ neighbourhood}{area\ of\ the\ neighbourhood} \quad (1)$$

Using this formula the user density is calculated for North, South, East, West and New Delhi regions and the results are given in table-2.

3. Results:

3.1 Heat Map Analysis:

The heat map plugin available in QGIS 2.14 (Essen) creates the relevant heat map. In which the geo-locations of tweet feeds has been taken as the unit of clustering and the darker portion of the map represent the higher concentration of Twitter users. It does not create any new data while it takes vector point layer as input in the form of shape file and produces the raster layer as result of the corresponding input layer. Which looks blur, as viewing an image using a translucent screen. The following map will produce.

By visualize this heat map for Twitter users it is clear that the whole region is divided into matrix or grids having different shades of colors whose range (max-410.093000, min-0.000000) is given in legends of the relevant map. By this heat map it is clear that the darker region is representing the higher concentration of users and lighter region is representing the lower concentration of users.

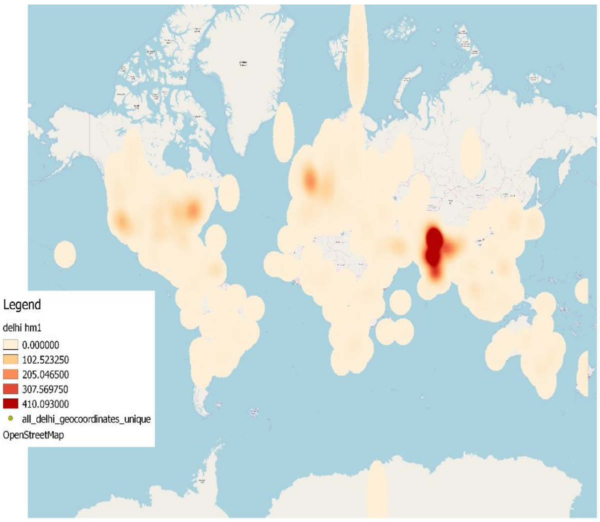


Fig. 5 Heat Map of given data model

3.2 Twitter User Density Calculation

In the given figure only specific locations (where Twitter users are active) are shown by rectangular shapes (in Fig 6(a)) and hexagons (in Fig 6(b)) and user density of each particular cell can be easily visualized by color codes of legends of produced map. Where each rectangle is approx. 1.40 square km and each hexagon is approx. 40.5 hectare or 0.405 square km in area.

With the below maps and using formula given in equation (1), the following table-2 is calculated. In this table the area is calculated by multiplying the number of cells to the area of a single cell and number of users are calculated by adding the number of users represented by corresponding color code as given in the legend of the map in Fig. 6(a) and 6(b).

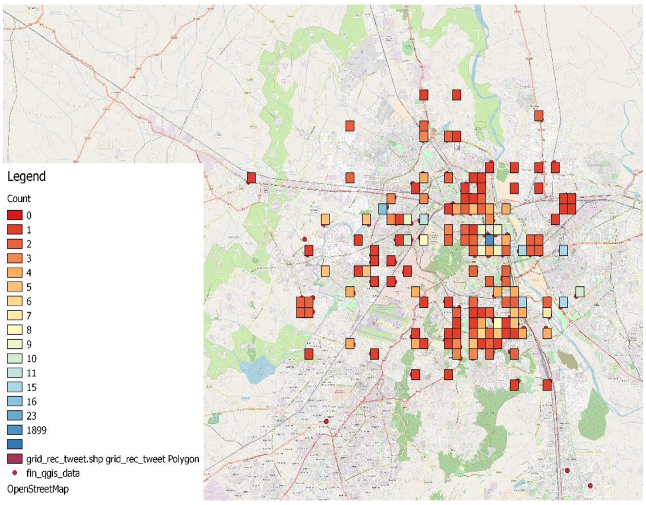


Fig 6(a): Twitter user density in rectangular grid of area 1.40 sq.km (approx.)

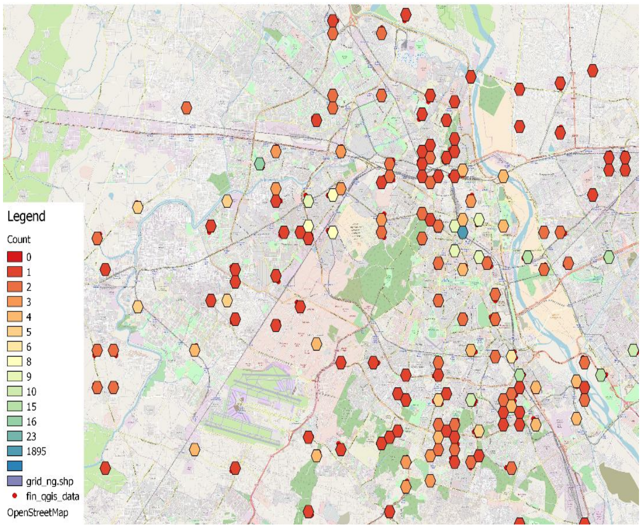


Fig 6(b): Twitter user density in hexagonal grid of area 0.405 sq.km (approx.)

Table-2

Region	With square grids		With hexagonal grids	
	Area (km^2.)	No. of users(user density)	Area (km^2.)	No. of users(user density)
North Delhi	9.8	23	4	10
South Delhi	26.6	42	1.6	12
East Delhi	7	5	2	4
West Delhi	2.8	11	2.4	27
New Delhi	32.2	65	2	1931

From Table 2, the following analysis can be done that in North Delhi region (Sadar market is considered) 23 users are active in 9.8 sq.km. area (while area is calculated using square grids). Similarly, for South Delhi region, 42 users are active in 26.6 sq.km. area, for East Delhi region, 5 users are active in 7 sq. km. area, for West Delhi region 11 users in 2.8 sq.km area and for New Delhi region 65 users are active in 32.2 sq.km area. The locations which are considered within particular region are Hauz Khas in South Delhi, Dilshad Garden and Karkarduma in East Delhi, Hari

Nagar in West Delhi and Paharganj in New Delhi respectively.

The same analysis can be done to calculated user density using hexagonal grids. The Number of users for particular cell is given by legends of map given in Fig. 6(b).

4. Conclusion and Future work:

It is a desktop application and would efficiently utilize GIS capability. GIS technology is becoming a part of our daily life by playing a vital role for route finding, weather forecasting etc. Although GIS has various applications as navigation, weather forecasting, telecommunication and network services, agriculture applications, disaster and natural resource management, network traffic density study etc. This paper basically presents how QGIS capabilities involve in heat map analysis and calculation of user density based on geo-locations of Twitter users. Therefore it would work as good decision- making a tool to estimate the network traffic by finding out the user density within a particular region. Which can have an application to any service provider to estimate an amount of specific service needed to supply in the corresponding region. So that service can be efficiently utilized and prevented from wasted. It would be a useful application for telecommunication department as well.

With all the calculation and analysis has been done in this paper leads towards figure out the highly dense region e.g. Paharganj in New Delhi region (also visualize by color code specified in legends of the map given in Fig. 6(a) and 6(b)).

This work has been done to calculate user density for Delhi region using QGIS tool but one can perform mathematical interpretation for the same and the same analysis can be done for the larger or smaller areas that has been used here.

References

[1]. Anselin, Luc. "Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences (92-10)." (1992).

[2]. Goodchild, Michael F., and Robert P. Haining. "GIS and spatial data analysis: Converging perspectives." Papers in Regional Science 83.1 (2004): 363-385.

[3]. Chang, Jonathan, and Eric Sun. "Location 3: How users share and respond to location-based data on social networking sites." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. AAAI Press, 2011.

[4]. Sloan, Luke, and Jeffrey Morgan. "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter." PloS one 10.11 (2015): e0142209.

[5]. Waheed, Hajra, et al. "Investigation of user behavior on social networking sites." PloS one 12.2 (2017): e0169693.

[6]. Daniel Sui, Michael Goodchild"The convergence of GIS and social media: challenges for Science"International Journal of Geographical Information Science Vol. 25, No. 11, November 2011, 1737–1748.

[7]. Anselin, Luc, and Arthur Getis. "Spatial statistical analysis and geographic information systems." *The Annals of Regional Science* 26.1 (1992): 19-33.

[8]. Han, Bo, Paul Cook, and Timothy Baldwin. "Text-based twitter user geolocation prediction." Journal of Artificial Intelligence Research 49 (2014): 451-500.

[9]. Ježek, Jan, et al. "Design and Evaluation of WebGL-Based Heat Map Visualization for Big Point Data." The Rise of Big Spatial Data. Springer, Cham, 2017. 13-26.

[10]. Sui, Daniel Z., and Michael F. Goodchild. "GIS as media?." International Journal of Geographical Information Science 15.5 (2001): 387-390.

[11]. Kumar, Shamanth, Fred Morstatter, and Huan Liu. *Twitter data analytics*. New York: Springer, 2014.

[12].https://www.fipp.com/news/insightnews/chart-of-the-week-almost-one-third-world-social (accessed on 16/04/2018)

[13].https://www.internetworldstats.com/top20.htm (accessed on 16/04/2018)

[14].http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-whyhexagons.htm (accessed on 16/04/2018)

[15].https://www.omnicoreagency.com/twitter-statistics/ (accessed on 16/04/2018)

[16].http://socialmedia-class.org/twittertutorial.html (accessed on 16/04/2018)