EXAMINING ASSOCIATIONS OF THE SOCIOECONOMIC CHARACTERISTICS WITH THE NUMBER OF GEO-TAGGED TWEETS IN CENSUS BLOCK LEVEL (CASE STUDY: BOSTON)

M. Molavi Gonabadi[1], P. Mojtabaee[1*], M. Taleai[1]

[1] Faculty of Geomatics, K.N.Toosi University of Technology, Tehran, Iran (moein.molavi@email.kntu.ac.it; pooya.mojtabaee@email.kntu.ac.ir; taleai@kntu.ac.ir)

Commission VI, WG VI/4

KEY WORDS: Spatial Analysis, Socioeconomic Parameters, Environment Characteristics, Social Media, Twitter, Geo-tagged

ABSTRACT:

In this study, the aim is to help uncover some facts about who the Twitter users really are. To this purpose, geotagged twitter data from the city of Boston together with socio-economic data were used. In the first step, tweets in each census block were counted and using the Getis-Ord Gi* index the hotspots and coldspots of tweet locations were extracted. Then, a multiple linear regression was employed, having the number of tweets as the response variable and the population data, age, education, occupation and income as the explanatory variables. Hence, more insight into the relationship between the number of tweets and some socio-economic factors is obtained. Results show that the central parts of Boston are the hotspot and the southern areas are the coldspot locations with regard to tweet numbers. The regression results imply that the number of tweets shared by users in an area is related to the income, the number of people having a university degree and the number of people having each type of job in that spatial unit. The results achieved in this paper could lead to a better vision and understanding in analyzing the tweeter users' behavior in any area of research and application.

1. INTRODUCTION

Today, social networks have gained enormous popularity as a medium to communicate and share contents. Twitter is one of the social networks which is very popular and in which approximately 500 million tweets are shared daily (Sloan and Morgan, 2015). Therefore, in different areas of research, social media has become a favorable data source (Dodds et al., 2011). Previous research have proven the potential of Twitter data in a wide range of applications, such as crisis identification and management, outbreak, human behavior (Allen,et al, 2016; Davis Jr, et al, 2011; Funayama, et al, 2014; Singh, Dwivedi, et al, 2017; Yang and Mu, 2015). The prevalence of GPS-equipped smartphones enables users to share their locations along with the content. Research has shown that roughly 0.85% of tweets are geotagged, which means that the accurate position of the person who tweeted was captured as longitude and latitude when he/she posted the tweet (Sloan et al., 2013) This geographic information is valuable because it enables linking user-shared content to its location. Having the location of the shared tweet gives the ability to relate the content with demographic and contextual data (Gayo-Avello, 2012). For example, by examining crime data recorded in neighborhoods and tweets about fear and insecurity located within the surrounding area, one can analyze the relationship between violent behavior and a sense of security (Sloan and Morgan, 2015). Curini et al. measured happiness by analyzing the content of the tweets in each province of Italy. Then, by examining the spatial relationship between happiness and meteorological and socio-economic parameters, the effect of each criterion on the degree of happiness was determined (Curini et al., 2015). Using the user-generated contents like twitter text content can yield information about different events or issues. Previous research has aimed to predict the behavior of populations only by looking at the tweet content. For example, there were attempts to predict disease outbreaks, film box office gross, election results, and stock market movements (Gayo-Avello et al., 2013). It is important to investigate more on who uses Twitter and there were research attempts to use the content of tweets to understand more about the population behavior. Researchers have previously tried to estimate demographic characteristics of Twitter users based on their tweets and other public data such as their Twitter profiles (Mislove et al., 2011; Pennacchiotti and Popescu, 2011; Rao et al., 2010; Sloan et al., 2013)

Looking at the tweets' locations puts forward the idea that their distribution is heterogeneous. Such a distribution is the result of a spatial process affected by environmental parameters which cause the considered phenomena not to follow a uniform distribution. Identifying these parameters can lead to a better understanding of Twitter users.

In this area, using Twitter data together with user profile's data, Grant Blank investigated the relationship between the twitter use and age, education level, economic situation and employment status of the users in UK and United States (Blank, 2017). The results showed that using Twitter in the UK was significantly correlated with age, income, education and lifestyle. Each year of age reduces the use of Twitter by about 6%. Also, people with a level of income above $ 40,000 per year are 3-4 times more likely to be Twitter users, whereas Twitter use in the United States has been significantly correlated with age and income. With each year of age, Twitter usage drops by about 3%. The study concluded that the non-representative characteristics of Twitter users indicate that using Twitter data for research where representativeness is important, such as predicting elections is inappropriate (Blank, 2017).

Sloan and Morgan (2015) have concluded in their study that there are statistically significant differences in demographic characteristics of the people using geo-services and the people

* Corresponding author

who geotag their tweets. Men are more likely to share geotagged tweets than women. Also, people sharing geotagged tweets are 0.82 years older than people who do not use location-based services of Twitter. (Sloan and Morgan, 2015). In another paper, a comparison was made between twitter population and census data of U.S. population in three axes of geography, gender and race/ethnicity, and it was found out that twitter population is a sample of the population which is not uniform (Alan Mislove et al., 2011). Generally, in the mentioned papers, the demographic, social and economic data were either obtained from the user's profiles or by the analysis of their tweet content. This study attempts to take a spatial view on investigating the relationship between demographic, social and economic parameters with the number of tweets shared by the Twitter users using their geotagged tweets. The following of the paper is structured as follows. Section 2 discusses the methodology. In Section 3, the results are presented and the conclusions are drawn in Section 4.

2. METHODOLOGY

2.1 Dataset

In this study, data of five-month geotagged tweets from 2018-11-17 to 2019-4-7 were used which were downloaded from Twitter Public API. These data included longitude and latitude of the tweeting location. Together with the location data, socioeconomic data include age, income, education level and job type are aggregated in block group scale were utilized (Figure 1).

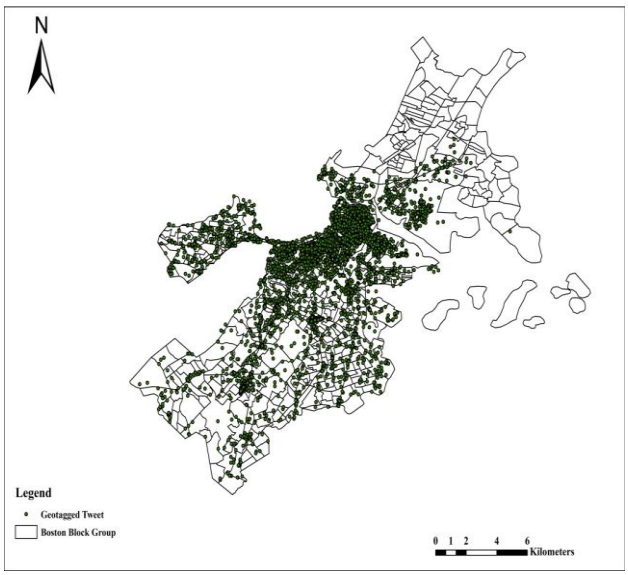

Figure 1. Tweet's Spatial Distribution in Boston

2.2 Spatial Autocorrelation

To investigate the spatial autocorrelation in the tweet data, Moran's I index was measured. The Moran index is obtained as follows. (Eq. 1)

$$I = \frac{n \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}\, z_i\, z_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} \sum_{i=1}^{n} z_i^2} \quad (1)$$

also the z-score for the statistic is computed as: (Eq. 2)

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}}$$
$$E[I] = \frac{-1}{n-1}, \quad V[I] = E[I^2] - E[I]^2 \quad (2)$$

where z_i is the deviation of an attribute for feature i from its mean $(x_i - X)$, $w_{i,j}$ is the spatial weight between feature i and j, n its equal to the total number of features.
The Moran's I values lie between -1 and 1. The negative values show negative autocorrelation and the positive values indicate positive autocorrelation. The z-score and p-value show whether or not the clusters are the outcomes of a random process. A z-score>2.5 or a z-score<-2.5 and a p-value<0.01 suggests that with a likelihood of 1% the observed pattern could be random and the null hypothesis is rejected.

2.3 Hotspot Analysis

Hotspot analysis considers the areas having statistically significant high values and low values as hotspot locations and coldspot location, respectively. A statistically significant hot spot is a feature having a high value which is also surrounded by other features with high values. The Getis-Ord Gi* is given as (Eq. 3).

$$G_i^* = \frac{\sum_{i=1}^{n} w_{ij} x_j - \overline{X} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\frac{\left[n \sum_{j=1}^{n} w_{ij}^2 - \left(\sum_{j=1}^{n} w_{ij} \right)^2 \right]}{n-1}}}$$
$$\overline{X} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (3)$$
$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\overline{X}\right)^2}$$

In these equations x_j is the attribute value for feature j, and w_{ij} is the spatial weight between feature i and j, and here n is equal to the total number of features.
In the current study to identify hotspots and coldspots of the tweets in the census blocks Getis-Ord Gi* is used.

2.4 Regression

To explain the relationship between the independent variables and the dependent variable multiple linear regression model is utilized in this study. (Eq. 4)
The formula for the multiple linear regression is as below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon \quad (4)$$

where y_i is the dependent variable, x_i are the explanatory variables, β_0 is the constant term or intercept, β_p are the slope coefficients for the explanatory variables and ε is the model's error term.

The response variable is the number of tweets in each census block which is normalized by its population. The explanatory variables include the number of people in each census block who are enrolled at college or graduate school, per capita income, median age, the number of people with sales and office jobs, the number of people having job in service sectors, the number of people having jobs in natural resources, construction and maintenance sectors, the number of people with jobs related to production transportation and material moving, the number of people having bachelor or higher degrees, the number of private wage and salary workers, the number of self-employed in own not incorporate business workers, the number of government workers and the number of unpaid family workers. All these mentioned variables were normalized by the total population except the median age and the per capita income.

3. RESULTS

3.1 Spatial Distribution

In this step, to investigate the spatial autocorrelation, the Moran's I index was used. The obtained z-score and p-value for this index was 11.06 and 0.001 respectively, proving the clustered distribution of the tweet locations. Getis-Ord Gi* index shows the hotspot and coldspot for the tweet locations (Figure 2). According to Figure 2 the northern central regions of Boston are the hotspots and the southern regions appear to be the coldspot locations of the tweets. The outcome of the hotspot analysis can visually give valuable information on understanding the spatial patterns and trends of the user's tweet sharing activity which can improve the vision in many applications related to the understanding of the Twitter user's behaviors.

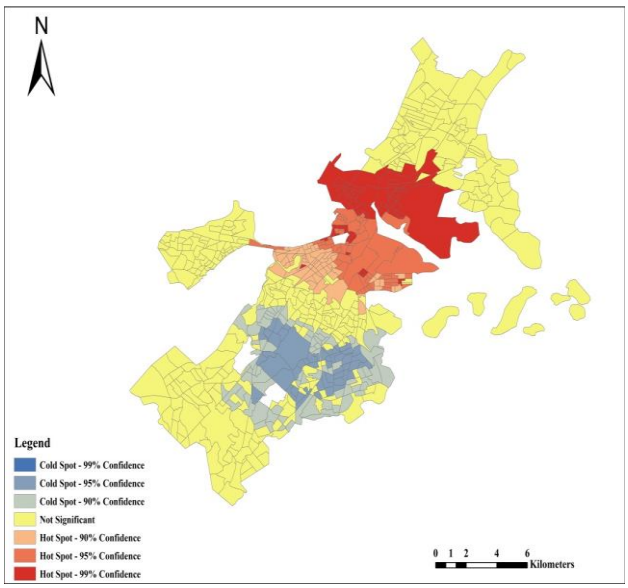

Figure 2. Spatial Clusters of Tweets in Boston

3.2 Regression Results

To discover the influential factors on the number of shared tweets, multiple linear regression was used. To this purpose, the total count of geotagged tweets in each census block was normalized by the population of their respective block so that the effect of the total population of each census block is removed. Considering that the dataset in this paper includes a limited number of tweets, some census blocks have zero tweet counts. Therefore, the census blocks containing no tweets were not entered in the regression model. Table 1 shows the results of the developed linear regression model. The obtained R-sq is 11.19%.

Dependent Variables	Coe.	p-Value
Constant	0.271	-
Median_Age	0.00164	0.831
Per_capita_income	0.00001	0.000
college_or_graduate_school_Enrollment_Per_population	0.00020	0.582
Bachelor_or_higher_degree_per_population	-1.777	0.001
Service_per_population	-1.917	0.054
Sales_and_office_per_population	-2.51	0.038
Natural_resources_construction_and_maintenance_per_population	-4.22	0.063
Production_transportation_and_material_moving_per_population	-4.11	0.044
Private_wage_and_salary_workers_per_population	1.842	0.013
Government_workers_per_population	1.16	0.394
Self_employed_workers_in_own_not_incorporated_business_per_population	0.74	0.738
Unpaid_family_workers_per_population	-5.6	0.733

Table 1. Regression Results for the Effect of Socioeconomic Characteristics on the Number of Tweets

According to the Table 1, among the considered parameters in this study, age variable which is the median age, is not influential on the number of shared tweets while the income appears to be statistically significant to be considered as an effective characteristic. Therefore, it can be inferred that the areas with higher income can probably have more users sharing their contents. From the education level viewpoint, the number of tweets shows no relationship with the number of students, whereas it exhibits an inverse trend with the number of people having university degrees, meaning that in regions with higher education level less content is probably tweeted. Also, the results demonstrate a statistically significant effect of the job type on the tweeting activity. Moran's local index was used to show the spatial correlation in order to better understand how Twitter activities interacted with income and qualification parameters. For this purpose, 645 population blocks including the number of tweets, income per capita and the number of people with a university degree were used and after 999 times Monte Carlo simulations, clusters with statistically significant level were identified.

In Figures 3 and 4, areas of high-high show the census blocks with high tweets as well as high-income levels and a high number of people with college degrees respectively. As can be seen in these figures in downtown Boston, in addition to high the number of Twitter users, the income and education levels of individuals are also higher than the other areas of the city. Figures 3 and 4 provide valuable information for researchers by identifying the spatial clusters of Twitter data and the demographic characteristics of people in those areas. Spatial clusters of tweets and demographics information can be helpful in research, such as analyzing human behavior in elections and marketing, by identifying Twitter target individuals.

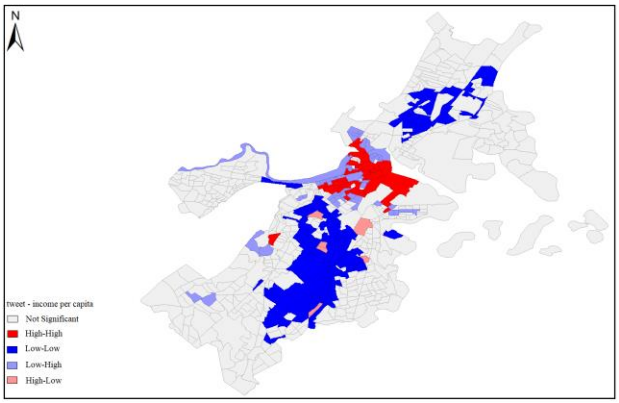

Figure 3. Clusters Showing the Spatial Correlation of the Number of Tweets and Income per capita

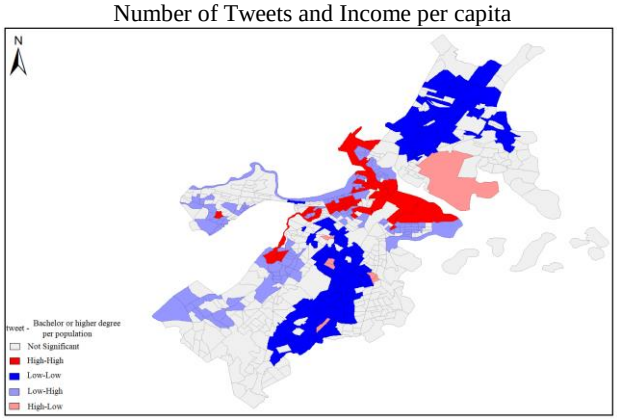

Figure 4. Clusters Showing the Spatial Correlation of the Number of Tweets and Number of People with a University Degree

4. CONCLUSION

Twitter is one of the most popular social network services which is used to share messages between users. Geo-tagged tweets, by sharing location alongside the content makes it possible to analyze the twitter data regarding a spatial point of view. In this paper, a 4-month geo-tagged twitter data of Boston city was used, and by utilizing the geographic locations of the tweets, the hotspot maps were produced that visualizes the areas in which the number of tweets was higher than their neighboring areas. Then, employing a multiple linear regression model, the environmental characteristics which are influential on the number of tweets in each census block were identified and it was investigated how these parameters affect the response variable. This study shows that the number of tweets is influenced by income, having a university degree and the type of job in a neighborhood. These results can yield knowledge and insight on understanding twitter users and the parameters affecting their activity which can be beneficial in applications of many sorts.

REFERENCES

Allen, C., Tsou, M.H., Aslam, A., Nagel, A. and Gawron, J.M., 2016. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PloS one, 11(7).

Blank, G., 2017. The digital divide among Twitter users and its implications for social research. Social Science Computer Review, 35(6), pp.679-697.

Curini, L., Iacus, S. and Canova, L., 2015. Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. Social Indicators Research, 121(2), pp.525-542.

Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R. and de L. Arcanjo, F., 2011. Inferring the location of twitter messages based on user relationships. Transactions in GIS, 15(6), pp.735-751.

Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A. and Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PloS one, 6(12), p.e26752.

Funayama, T., Yamamoto, Y., Tomita, M., Uchida, O. and Kajita, Y., 2014, November. Disaster mitigation support system using Twitter and GIS. In 2014 Twelfth International Conference on ICT and Knowledge Engineering (pp. 18-23). IEEE.

Gayo-Avello, D., 2012. " I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data. arXiv preprint arXiv:1204.6441.

Gayo-Avello, D., Metaxas, P.T., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P., Kalampokis, E., Tambouris, E. and Tarabanis, K., 2013. Understanding the predictive power of social media. Internet Research.

Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P. and Rosenquist, J.N., 2011, July. Understanding the demographics of Twitter users. In the fifth international AAAI conference on weblogs and social media.

Pennacchiotti, M. and Popescu, A.M., 2011, August. Democrats, republicans and starbucks afficionados: user classification in twitter. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 430-438). ACM.

Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M., 2010, October. Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

Singh, J.P., Dwivedi, Y.K., Rana, N.P., Kumar, A. and Kapoor, K.K., 2017. Event classification and location prediction from tweets during disasters. Annals of Operations Research, pp.1-21.

Sloan, L. and Morgan, J., 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PloS one, 10(11).

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P. and Rana, O., 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological research online, 18(3), pp.1-11.

Yang, W. and Mu, L., 2015. GIS analysis of depression among Twitter users. Applied Geography, 60, pp.217-223.