Do We Tweet Where We Ride?

Alysha Alloway[1¶] and Lauren Strug[1¶]

[1]Department of Geography, Environment, and Society, University of Minnesota - Twin Cities, Minnesota, United States of America

[¶]These authors contributed equally to this work.

Abstract

The field of geographic information science has over the last decade embraced Twitter data as a rich source of large-scale individual-level data, because of its broad availability for analytic use, specificity of location information, applicability to a variety of research objectives, and diversity of user base, among other reasons. Geolocated tweets are a particular focus in our field. For a tweet to be geolocated that means exact coordinate information, a point representing location based on latitude and longitude on the earth's surface, is linked to the tweet based on where the Twitter user tweeted from. This coordinate information has been used to infer human movement patterns and real-world interactions with our environments. This project analyzes spatial clustering of individuals using Twitter users' geolocated tweets history and transit users' origins and destinations, comparing the clusters of each user group to discuss whether Twitter users can be considered a relatively reliable proxy for the Twin Cities population in general, or the Twin Cities population using public transit in particular. Here we show that in the case of the Twin Cities, Twitter users do not represent the same pattern of clustering as transit users' origins or destinations, leading us to question what Twitter users geolocated tweets do represent, and whether geolocated tweets are as useful as a proxy for population movement as some past research has claimed. With this work, we hope to continue the discussion of how Twitter data can best be used in our field and in other fields, calling other data scientists to question the assumptions that are made when using geolocated tweets as robust portions of a research work's data analysis.

.

Introduction

The use of social media data in general, and Twitter data in particular, has been a popular topic of discussion in the field of geographic information science (GIS) in recent years. Academics have used Twitter data to expand or explore the ways in which humans interact with our environment [cite with some authors we cite]. Twitter is a popular microblogging platform which allows users to share short messages called Tweets. Many tweets hold geographic information indicating where it was sent from. These tweets are either "geo-tagged" or "geolocated." Geo-tagged tweets are associated with an aggregated place, often a city, but no coordinate information is associated with geo-tagged tweets. Geo-tagged tweets are more common than geolocated tweets. Geolocated tweets require that

the Twitter user turn on location information for their tweets. Geolocated tweets have coordinate information linked to the tweet and thus have a finer resolution of place and can be used to look at individual-level movement. It is important to note that Twitter data cannot be considered to represent the whole population of any place, in that only 22% of adults use Twitter and most of them are younger adults [8]. The 2018 Pew Center Survey shows that 44% of adults age 18 to 24 use Twitter and 7% of adults over the age of 65 use Twitter [8].

This project analyzes spatial clustering of individuals using Twitter users' geolocated tweets history and transit users' origins and destinations, comparing the clusters of each user group to discuss whether Twitter users can be considered a relatively reliable proxy for the Twin Cities population using public transit. Transit users' origin, meaning where a rider started their transit journey and destination meaning where the rider was headed using transit. The transit origin and destination information was accessed from the 2016 On-Board Travel Behavior Inventory survey, made freely available to the public by Metro Transit, the Twin Cities' transit authority.

This work addresses whether Twitter data is a useful point of data comparison for understanding urban mobility, whether Twitter data can be used in conjunction with other transit data to support a clearer or more nuanced understanding of urban mobility patterns, and whether the methods used by other researchers in the realm of geolocated social media data can be applied to our local Twin Cities context, given the lower usage of Twitter and lower urban density.

In this paper we will discuss the theories and methodologies of other researchers using Twitter data to analyze urban human movement, followed by our case study of clustering Twin Cities transit users' origin and destination nodes and Twin Cities Twitter users' geolocated tweets, and our results.

Related work

Social media has made the use of Twitter data in geography, geographic information science, and geocomputing popular in recent years. Considering that roughly a quarter of the adult population uses Twitter, there is a question of whether using geolocated tweets and Twitter user data contributes to a valid and representational dataset. As it relates to urban mobility studies, comparing geolocated tweets and Twitter user data with more traditional data sources is a common methodology [3, 11]. Twitter users may not be a complete representative of the

public and geolocated tweets are a small part of all tweets that the Twitter Streaming API can collect which is a smaller part of the whole Twitter dataset. As such, the discovered knowledge will only reflect the human activity and mobility patterns for a portion of the total population [4]. We apply the same considerations to our study. For example, geolocated tweets in the Twin Cities might correlate with tourist destinations, area colleges, or the international airport.

Given that "a large collection of tweets, when viewed through the lens of individual-level human mobility patterns, can be simplified to a series of key locations for each user" we assume tweets will cluster around an individual's home, work, or other common locations such as school [11]. This same assumption for transit data is that origins and destinations show the beginning of a trip (home or place spent during most of the day) and the end of a trip (work, school, home, or other common destination) in an aggregated form.

Research supports that geolocated tweets can capture features of human mobility for individuals within and between cities [1]. Jurdak et. al. provides "solid evidence that Twitter can indeed be a useful proxy for tracking and predicting human movement" [1]. Previous research demonstrates patterns and mobility of Twitter users by analyzing the spatial and temporal dynamics in their tweets. There are many studies which use Twitter data to analyze human mobility at a large-scale [5, 9, 12]. Yin et. al. provides an example of large-scale geolocated Twitter data for which shows clear delineation of administrative and urban boundaries at small and large scales [12].

There are also many examples of research using Twitter data to analyze mobility patterns for individual cities [2, 3, 4, 6, 7]. Lenormand et. al found that users tweeting geolocated tweets in Barcelona and Madrid correlated with mobility patterns observed in the cities using cell phone user data and census data on population density [3]. Li et. al. provides research that aims to discover the patterns and mobility of Twitter users by analyzing the spatial and temporal dynamics in their tweets [4]. Their study collected geo-tagged tweets from four college cities [4]. Our work applies clustering to a dataset of geolocated tweets within the boundaries of the Twin Cities Metro Transit region for comparison with the clustering patterns of Twin Cities daily travel activity. Like previous research we are seeking to find a comparison between the relative clustering of these two small-scale data sources under the assumption that geolocated tweets and transit origin and destination points represent a broad view of

human mobility in the Twin Cities. The goal of our clustering method is to explore if the origin and destination of common Metro Transit trips appear in Twitter geolocated tweets from the same bounding box as the Transit data.

Methods and materials

In this research we used two primary datasets, one of Twitter users that fit our criteria of having tweeted at least one geolocated tweet within the Twin Cities Metro Area (as defined for the purposes of this work as within the Metro Transit route network) and one of transit users' origins and destinations for one trip on the day they were surveyed. Both datasets present point information: the point at which a geolocated tweet was tweeted, the point of origin where a user started their transit journey, and the point of destination for the transit user. Each point is a coordinate, latitude and longitude pair for location information, that we can plot and then cluster together to look at overall location patterns for tweets, origins, and destinations. We worked with our data in Python using Google Collaboratory to create a reproducible Jupyter Notebook of our project. We performed K-means clustering using sklearn and scipy python modules and gathered and formatted twitter data with the Twitter REST API, json, csv, and pandas python modules.

Additionally, we aggregated our raw point data for tweets, transit origin points, and transit destination points to census tracts. This gave us an alternate way to view our data and spatial patterns. We normalized the census tracts by 2010 population. To further see the key changes between our twitter and transit datasets, we calculated z-score for each aggregated census tract dataset. We are then able to compare the z-scores of tweets with origins and tweets with destinations.

Datasets

We received a list of users from the socio-environmental data explorer (SEDE) database at the University of Minnesota [10]. 339,602 user records (including duplicates) were queried from the database from over a billion tweets that tweeted in our bounding box. Our list of users resulted in 22,387 unique users with geolocated tweets. Our bounding box in lat/long is (44.717953,-93.795776,45.344424,-92.886658). It includes all Twin Cities Metro

Area bus stops included in the Travel Behavior Inventory 2016 data. We used the Twitter REST API to gather the timelines of the 22,387 users with geolocated tweets within our bounding box.

The Travel Behavior Inventory (TBI) is a study of household demographics, daily travel activities, and typical transportation patterns throughout the greater Twin Cities region, sponsored by Metro Transit and the Metropolitan Council, the Twin Cities' regional planning authority, in partnership with the Minnesota Department of Transportation and the Wisconsin Department of Transportation. For our study, we focus on the coordinate information recorded for individual participants' origins and destinations within the Twin Cities metro area.

Twitter data preparation

User timelines were filtered using the Twitter REST API to capture only geolocated tweets that fall within our bounding box. We gathered a list of users who fit our profile: having tweeted geolocated tweets and being within our bounding area of the Twin Cities metro area. For each gathered tweet with our bounding box we extracted coordinate information for the clustering analysis.

Although we started with a large number of users, many of those users were identical, meaning they had tweeted more than one geolocated tweet from within our bounding box. A challenge with Tweet gathering and our reliance on the SEDE database of tweets is that many users change their Twitter usernames, which we used to gather timeline tweets, or the account was deleted. This explains why our number of users decreased from 339,602 to 22, 387. Our final number of users is less than 22,387 because some geolocated tweets that exist within the SEDE database are too far back in time to capture with the Twitter REST API. Our final number of users with recent geolocated tweets within our bounding box is 3,585 users and our final number of tweets is 8,723 total. Users had to have one or more geolocated tweets to be included.

Clustering: K-Means Methodology

We used a K-Means clustering model validated by the Elbow Method to start looking for the presence of a spatial relationship between our data sources. K-Means clustering is a method of vector quantization which aims to partition n-observations into k-number of clusters. Each observation belongs to the cluster with the nearest mean.

This method results in the data partitioned into Voronoi cells, which tells us not only about relative density, but also about areas that are most important to a portion of the population's individual movement patterns, either coming from, going to, or tweeting from that location. Relative density can be differentiated by smaller cells of tightly grouped points versus larger, more dispersed cells of points.

The Elbow Method is a way to validate the correct number of significant clusters within a dataset, so-called because it resembles the crook of a human arm, and the elbow of the metaphorical arm is the correct number of clusters to use in a K-Means model [Fig 1].

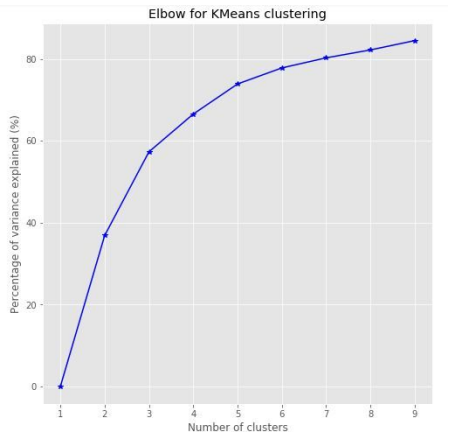

Figure 1. Elbow method for geolocated tweet data

Clustering Results

We hypothesized that tweet clusters might share similarities with both transit origin and transit destination clusters, but would likely not match up exactly with either. Generally, it could be assumed that Twitter users would either tweet from their home location or a location where they spend the majority of their time. For the transit dataset, this is assumed to be a transit users' origin location. Alternatively, a Twitter user might tweet from a location that is a common destination in the Twin Cities, which we can assume would likely appear in the transit users' destinations.

As we can see in the cluster figures below [Figs. 2-4] the number of clusters has a significant impact on how each dataset appears to the viewer. The Elbow Method for the Twitter data showed that the significant number of clusters was five clusters, while for transit origin and destination the significant number of clusters was seven. Several locations of interest can be seen in each of these clusters. Downtown Minneapolis is apparent in all these data, represented at the center of the points in green and red for the origin data and purple and green for the destination data where K=7 [Figs. 2, 3.] For the geolocated tweets where K= 5, downtown Minneapolis is not as closely defined as it is in the transit data, but it is also its own significant cluster, shown in blue [Fig 4.] All figures show the same cluster for St. Paul, the right-most cluster in purple [K=7, Fig. 2], brown [K=7, Fig. 3] and green

[K=5, Fig. 4]. Since St. Paul is far less dense than Minneapolis, the clustering results also might support that with lower population densities clustering is a less applicable method for looking at individual-level urban movement.

Comparing between datasets, Twitter data [Fig. 4] looks more similar to the transit user destination locations [Fig. 3] than it does to the transit user origin locations [Fig. 2]. There are a number of ways to explain this, and we will discuss this further as we conclude. What is also interesting is how the overall dispersal pattern of data points between transit and Twitter data differs, which we will turn to in our discussion section.

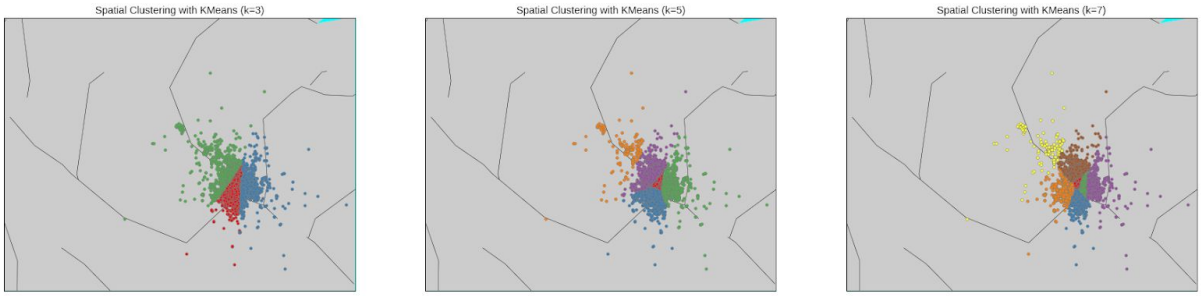

Figure 2: Origin Clusters

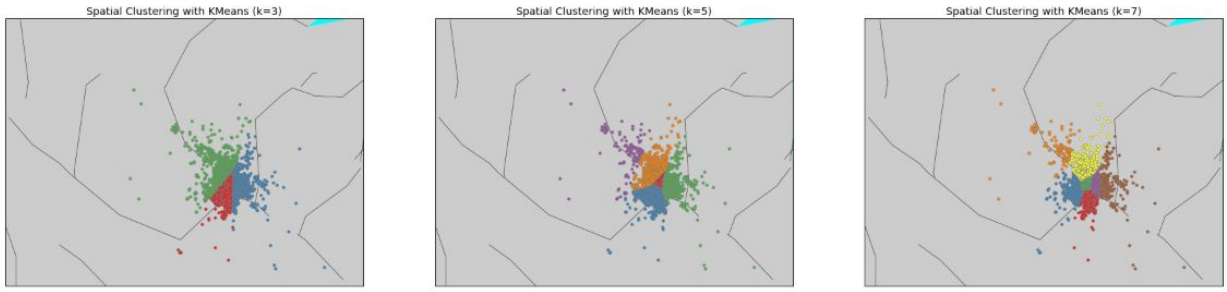

Figure 3: Destination Clusters

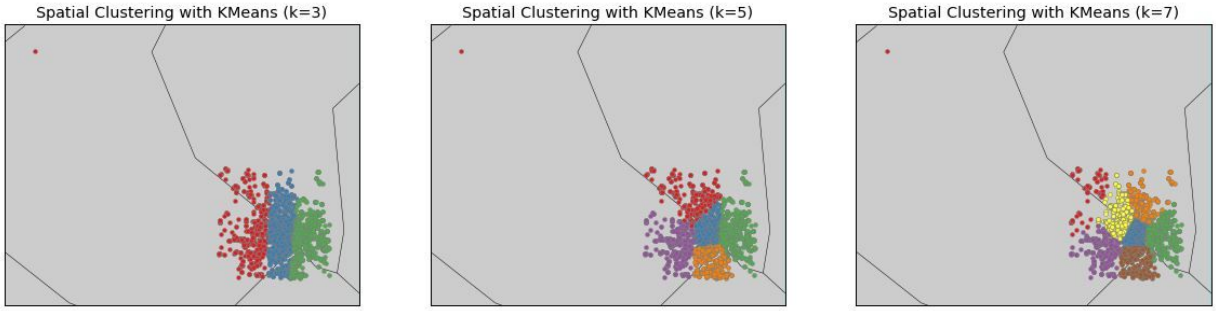

Figure 4: Tweet Clusters

Aggregation to Census Tracts and Z-Score Comparison

To further visualize our data beyond raw point counts, we aggregated tweets, origin, and destination points to census tract geography [Fig 1, Appendix A]. By aggregating the raw data points to census tracts we are also able to normalize the three datasets by the population count of each census tract. Normalizing by population counts gives us a better understanding of where people are during the day away from their homes, since census population counts rely on the nighttime (i.e. household, dorm, etc.) location of census respondents. Areas that show high concentrations of tweets, transit origins, and transit destinations with low populations can be assumed to show where people concentrate that are not where they live, such as work, special event locations, and popular destinations [Fig. 2, Appendix A].

We calculated a z-score for each census tract for tweet, transit origin, and transit destination point counts. We calculated the average point count and standard deviation for each data type by census tract in Excel, and then used those mean and standard deviation values to calculate z-score. A z-score tells us how many standard deviations a record deviates from the average (mean). We visualized the z-scores for tweets, origins, and destinations using quartiles where the Q1 group are the z-score values between the lowest value and the middle number between the lowest value and the median, Q2 are z-score values between the low middle value and the median of the data, Q3 are the values between the median and the high-middle value of the dataset, and Q4 are the z-score values between the high-middle value and the highest value. Calculating z-scores helped us understand where each of the data we looked at diverge from their average value, meaning the census tracts that are significantly different from each data sources typical point count. We used these values to explore census tracts that showed a great amount of mismatch between data sources, i.e. where tweets counts were much higher than their average and transit counts were much lower than their average, and vice versa.

Clustering, aggregation to census tracts to normalize by population, and z-score calculation are all exploratory stages of analysis, marking the beginning stages of this project. Now that we have explored our data, we better understand what can be done to further analyze it, as well as the general implications and limitations of working with Twitter data, which we will now turn to in the discussion section.

General Discussion

The clustering analysis does not show an exact visual match between the transit origin and destination clusters and the geolocated tweet location clusters, but these figures [Figs. 2-4] do share some common patterns of clusters. It is important to note that Minneapolis has a higher population density, higher transit connectivity, and higher share of jobs, which could affect the patterns we see on our figures. Despite these differences, both Minneapolis and St. Paul city centers were common high density areas for all datasets discussed, supporting the well recognized assumption that people use transit to commute, but also showing that the landscape of geolocated tweets made by Twitter users follows the same general pattern of where people are located during the day.

Looking at our results, it is clear that the overall pattern between transit user data, both origin and destination, and the geolocated tweet location data differs greatly in some ways and shares commonalities in others. The tweet locations have a uniform point pattern across the Twin Cities area, while the transit user data shows clear linear-style clustering that approximates the transit network. With these patterns in mind, we can claim that people who rely on public transportation are less likely to live in certain areas, and that is not a location limitation shared with the population of people using Twitter. However, the Twin Cities urban centers and the airport and Mall of America were all outliers of high activity for both transit use and tweet location. We hesitate to claim that Twitter users and transit users are proxies for each other, in that the demographic characteristics of Twitter users and transit users in the Twin Cities are known to differ significantly.

We compared these datasets for the purpose of this project, but it is important to note the size of the twitter data after applying our research prerequisites was limited in comparison to the transit user data. These datasets also did not account for any weighting of user, Twitter or transit, based on whether they are a tourist or live locally. Some assumptions could be made that people who purposefully share their location with others might bias that location sharing to locations that have significance, for example a tourist visiting the Mall of America and tweeting about it.

As for our results and future work, while we do see some amount of visible spatial relationship between the transit origin and destination figures and the Twitter figures, more work remains to substantively combine these datasets to give us a better picture of where people are in the Twin Cities during the day. While geolocated tweets from Twitter are an exciting data source to explore, it is challenging as a researcher to gather a significant amount of

data in a short period of time with the constraints of using the Twitter API for free. To further our work, we feel a larger sample of geolocated tweets would give more significance to a comparison. Additionally, linking users and the temporal aspect of tweets to create origin-destination points for our twitter dataset would be interesting to analyze with linked origin-destination transit survey data. This project, as an exploration of working with Twitter data in Python and comparing it with a more traditional data source, provided us with a number of possible next steps. It challenged us to work with technology outside of our comfort zone, and hopefully inspires others to explore nontraditional data methods and data sources as well. This work is a jumping-off point for exploring these data sources further, and an interesting look into the popular, controversial topic of working tangibly with social media data.

Acknowledgements

Thank you to Professor Eric Shook for your assistance and allowing us access to the SEDE database. Thank you to the CyberGIS Seminar for your helpful peer review feedback. The content is solely the responsibility of the authors and does not necessarily represent the views of the University of Minnesota department of Geography, Society, and Environment.

References

1. Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D., & Wu, Y. (2015). Understanding Human Mobility from Twitter. *PLoS ONE, 10*(7), E0131469.
2. Lee, J.H., Gao, S., Janowicz, K., Goulias, K.G., 2015. Can Twitter data be used to validate travel demand models?, In: IATBR., WINDSOR.
3. Lenormand M, Picornell M, Cantu´-Ros OG, Tugores A, Louail T, et al. (2014) Cross-Checking Different Sources of Mobility Information. PLoS ONE 9(8): e105184. doi:10.1371/journal.pone.0105184
4. Li, Yue, Li, Qinghua, & Shan, Jie. (2017). Discover Patterns and Mobility of Twitter Users-A Study of Four US College Cities. *ISPRS International Journal of Geo-Information, 6*(2), 42.
5. Liu J, Zhao K, Khan S, Cameron M, Jurdak R. Multi-scale population and mobility estimation with geo-tagged tweets. In: Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on. IEEE; 2015. p. 83–86.
6. Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. Environment and Planning A: Economy and Space, 47(2), 465–484.

7. Molinari, M. E. , Oxoli, D., Kilsedar, C. E., & Brovelli, M. A. . (2018). User Geolocated Content Analysis for Urban Studies: Investigating Mobility Perception and Hubs Using Twitter. *The International Archives of the Photogrammetry, 42*(4), 439-442.

8. Perrin, A., & Anderson, M., (2019). Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center.* https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

9. Rashidi, Abbasi, Maghrebi, Hasan, & Waller. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C, 75*(C), 197-211.

10. Shook, E. & Turner, V. (2015). The socio-environmental data explorer (SEDE): a social media–enhanced decision support system to explore risk perception to hazard events. *Cartography and Geographic Information Science,* 43(5), 427-441.

11. Soliman, A., Soltani, K., Yin, J., Padmanabhan, A. and Wang, S., (2017) Social sensing of urban land use based on analysis of Twitter users mobility patterns. PLoS One.

12. Yin, Junjun, Soliman, Aiman, Yin, Dandong, & Wang, Shaowen. (2017). Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science, 31*(7), 1293-1313.

Appendix A

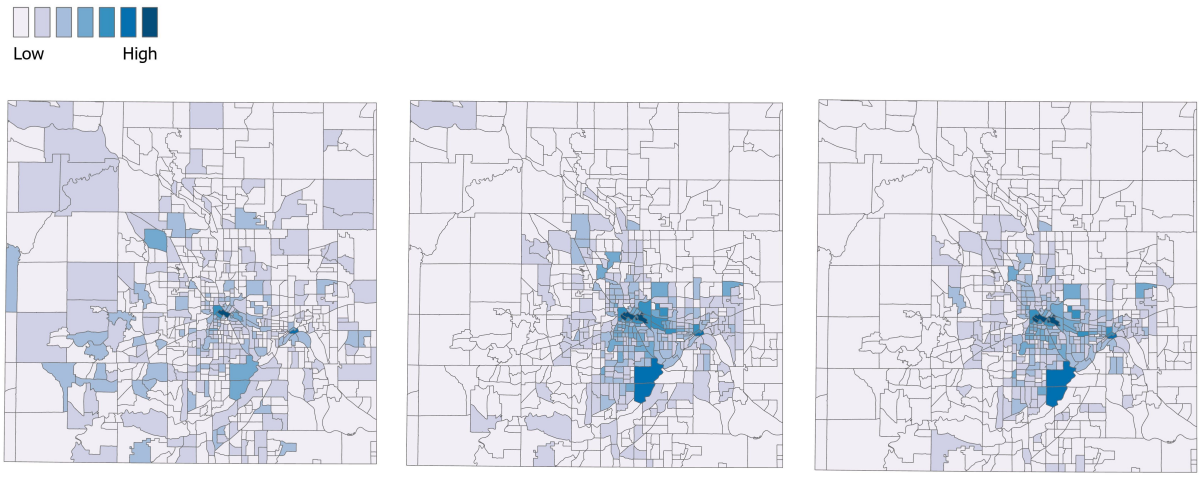

Figure 1. Tweets (left), transit origins (center), and transit destinations (right) aggregated to census tracts.

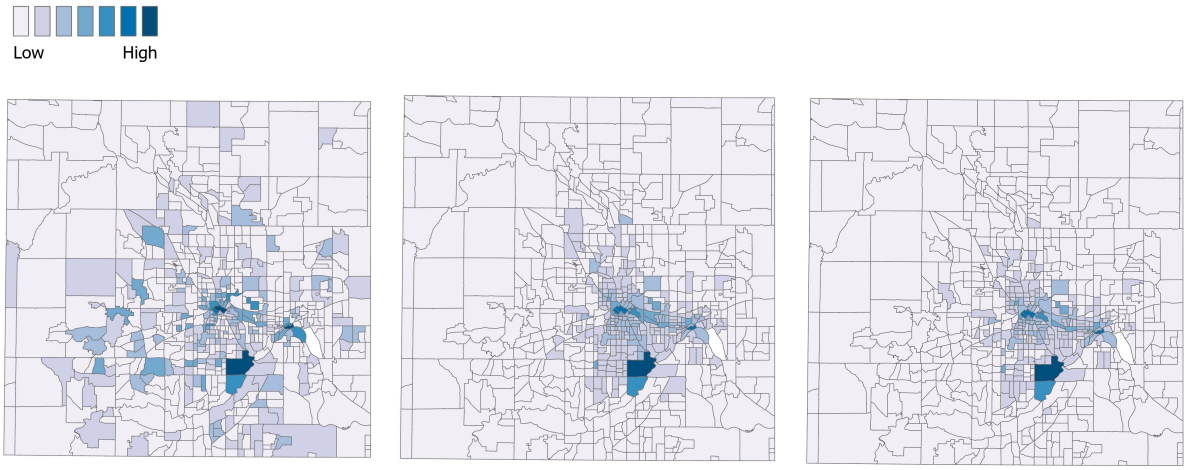

Figure 2. Tweets (left), transit origins (center), and transit destinations (right) aggregated to census tract and normalized by 2010 population count.

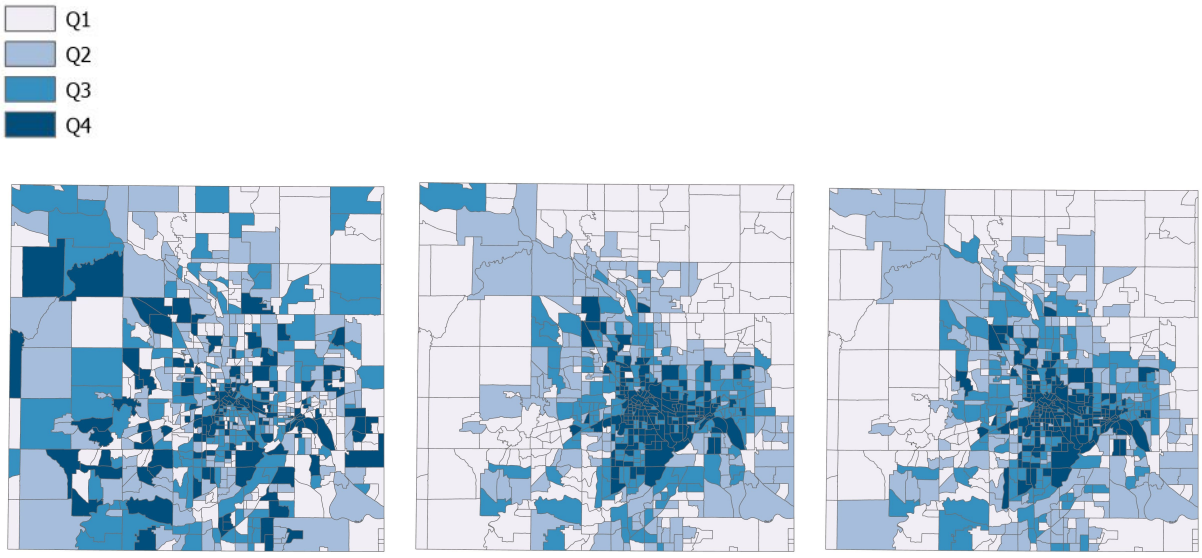

Figure 3. Tweets (left), transit origins (center), and transit destinations (right) z-score quantiles.