Urban partitioning through Twitter activity timeseries

Philipp Kats
New York University (NYU)

I. Abstract

With the popularity of different social media platforms, the amounts of data recorded might be seen as an opportunity to fulfill needs of city management. Strugglyng multiple challenges, city gouverneurs are eager to get real-time data in order to understand what is happening in the city and reinforce their decisions. In our study, we investigate how Twitter stream might be used to characterize urban areas and predict their socioeconomic properties. As result, several directions of potential implementation were defined, including event detection and evaluation, areas partitioning and profiling, and prediction of areas socio-economic properties.

II. Introduction

As digital technologies are becoming more and more widespread, big data created by recording the digital traces left behind human activities become a powerful mean to study various aspects of human behavior. Many of those aspects can be described with social media feeds - data, generated and shared by people though multiple global social media platforms [1]. At the same time, the increasing urbanization of the world's population and great diversity of new urban population deeply affect urban environment. Solving many challenges of modern cities, including crime, illegal construction, tax regulation, emergencies, and many others, require frequent updates and prioritization, and, therefore, large quantities of highly granular incoming data with frequent updates for analysis.

This need might be resolved by records of modern communication systems, and social media in particular. Focusing on records aggregated on spatial locations rather than on individuals, new approaches have been initiated different types of communication might be used for many purposes, from urban landscape description [2] [3] [4], to regional delineation [5] [6], population density estimation, land use classification [7] [8] and identification of social groups and events [9] . As many channels and platforms of communication have unique sets of features, it is crucial to develop theoretical frameworks and a real-time monitoring systems, as it is required to understand how the individual dynamics shape the structure of our cities in order to make better tactical decisions and general strategies for city governance [1].

We use Twitter stream for 3 years, generated within New York city as a main source of data in our study. Twitter source was selected for it's accessibility, rich data, including geographical location, and large user base within the city. Also, while Twitter by itself represents a wide range of activities, it also provides "app signature" — credentials of application, that initiated particular tweet, — along with each message. Therefore, some part of the tweets can be interpreted as very specific activity. Our choice of geographical boundaries was defined by large penetration of Social Media, huge population, and multiple complimentary datasets available

Further we provide a comparative study of twitter stream for New York City. The focus of our study is on demonstrating that temporal "signature" of an area can be interpreted and used to explore relationships between urban areas, used for event detection and evaluation, and even represent certain sociodemographic characteristics of the area.

III. Materials and Methods of Aquisition

A. Twitter data gathering and preprocessing

A feed of Twitter data was collected through official API using ensemble of custom scrapers. Data then being processed, unnecessary attributes were removed, and time adjusting to EST.

B. Spam Filtering

Several applications generate tweets automatically and do not represent any human activity. Many of them generate significant quantities of messages and do so in the specific short range of time. Therefore, they can potentially add noise to our data, and should be removed.

As many of those applications have abnormal (in comparison with whole dataset) percentage of tweets per user: most of them, in fact, have less than 10 users, while generating thousands of messages. Therefore, to filter most of spam messages, we decided to drop all the tweets, generated by applications, for which more than 10% of total tweets belongs to one user. As such, we remove all tweets from 12 applications, including *NYC_511* road traffic bot, *NYC job offer* bot, and others. In terms of messages, roughly about .2% of tweets were removed.

C. Data overview

Final dataset consists of more than 23 millions of tweets from about 6 hundred of thousands unique users, published withing the geographical boundaries of New York City, from June 2013 and until June 2016 (with few minor gaps). Tweets were generated using 603 different applications, though almost 60% were generated by *Twitter for IOS*, 16% - *Instagram* and 14% - *Twitter for Android*.

D. Spatial joint

As we need to aggregate tweets by place of origin, each tweet was given a Postal Code, basing on its geographical location through spatial joint and 262 Postal Code Boundaries (PCB) of New York. Postal Code boundaries were chosen as they provide, on one hand, picture detailed enough to provide valuable and interpretable results, and, at the same time, large enough to have enough tweets aggregated within

	user_id	id	%
application			
Twitter for iPhone	336590	13611475	59.059995
Instagram	275865	3706224	16.081253
Twitter for Android	94154	3237441	14.047210
foursquare	41888	624360	2.709089
Foursquare	33150	539947	2.342822
Twitter for iPad	10380	308223	1.337375
dlvr.it	26	130040	0.564242
iOS	32018	86180	0.373934
Twitter for Android Tablets	2444	81596	0.354044
Tweetbot for iOS	3250	69645	0.302189

Figure 1. 10 top popular applications for Twitter

each code to receive statistically significant results. In order to add interpretation volume, Postal codes were then associated with particular neighborhoods, basing on *ZIP Code Definitions of New York City Neighborhoods* [10] table.

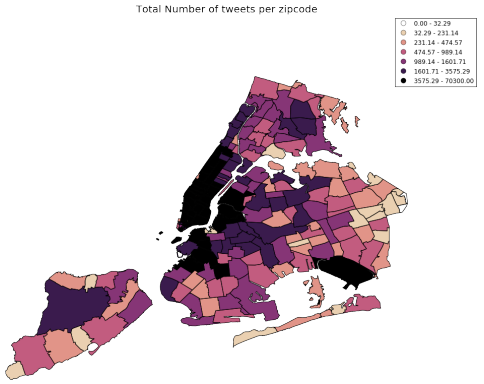

Figure 2. Replace this text with your caption

E. American Community Survey

Sociodemographical data from American Community Survey (ACS) 5-year Summary tables for year 2013 was used as a reference to establish relationships between social media stream and neighborhood characteristics. Data was collected from official source [11] . ACS contains hundreds of questions on multiple topics, from population and demographics, to household median income, commute, health and insurance, education, rent, access to the Internet, and many others. It provides many geographical levels, include postal codes.

IV. Time series

Typical week signatures were constructed in order to understand basic patterns for every location, following procedure, presented in [12]. First, all tweets were aggregated to the total number of tweets per each 15-minutes range. As there were gaps in data collection, we remove them by dropping all samples for days with zero tweets in total. An aggregation was made separately for every group of tweets: — city in total, each Postal Code, and every of the top popular applications. Each aggregation resulted in one record per group - Postal code, application and city as a whole, containing 672 (4x24x7) attributes each, and normalized by dividing each attribute by their total sum.

Plotting those average weeks, we revealed both local and universal city-vise patterns: on average, people used to tweet the most around midnight. This pattern changes slightly through the week, with pike going late on Fridays and earlier on Sundays and Mondays. At the same time, variance for the weekend is much higher. It would be reasonable to expect seasonal and weather-related dependencies.

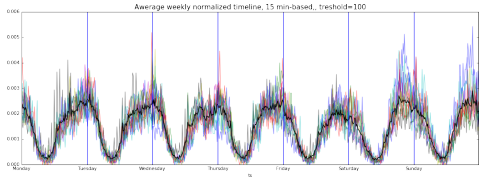

Figure 3. Averge week for each zipcode in New York. Blue vertical lines represent midnight. Black line - average week pattern for whole New York

A. Postal code Boundaries and Neighborhoods

While each Postal code has its unique week signature, many of the signatures represent are similar and clearly depends on areas functional role. For example, patterns for residential and business areas differs drastically: residential areas have large pikes in the evening, especially through weekday, while business districts are characterized by pikes in the first part of the day. As such, all zipcodes in Lower Manhattan have more tweets during the day, and less in the evening, than New York on average. On the contrary, areas such as Flatbush (Brooklyn) and Jamaica (Queens) neighborhoods have well defined pikes in the evening.

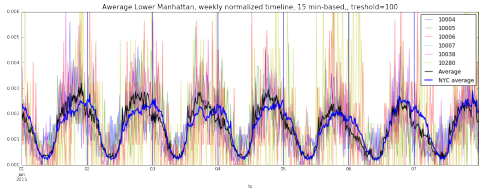

Figure 4. Average week signature for Lower Manhattan, Manhattan

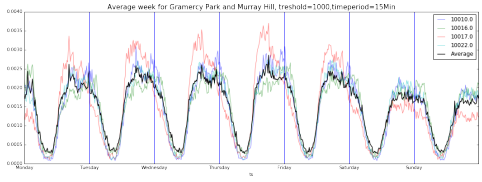

Figure 5. Average week signature for Grammercy park and Murray Hill, Manhattan

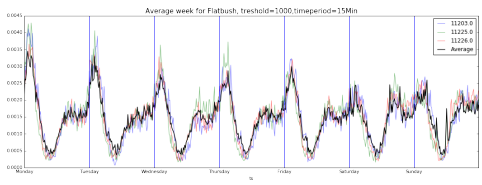

Figure 6. Average week signature for Flatbush, Brooklyn

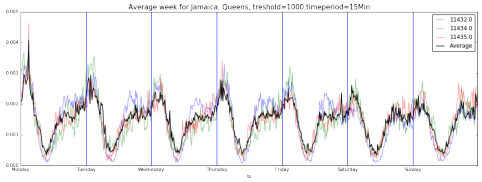

Figure 7. Average week signature for Jamaica, Queens

B. Applications

Having the application signature for each tweet, we were able to create a detailed city-wise signature per each of the top 15 popular applications. signature for each application depends heavily on its particular functionality. For example, *Dlvr.it* app, which is used to request food and grocceries delivert, have a smooth and balanced timeline. On the contrary, *Foursquare* app has pikes in the evening. Also, for some behavioral reason, *Twitter for iPad* has a narrow pike during lunch time for every workday.

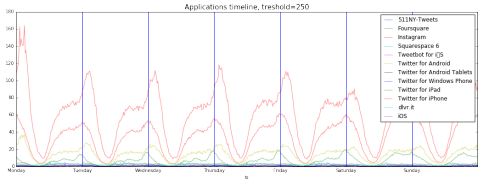

Figure 8. Average week plot for most popular applications

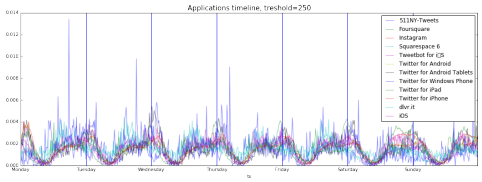

Figure 9. Average week plot for most popular applications, normalized

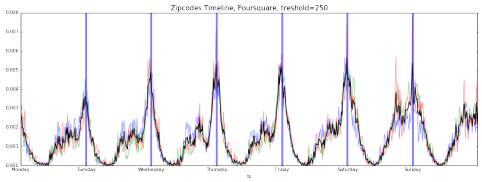

Figure 10. Average week plot for Foursquare app

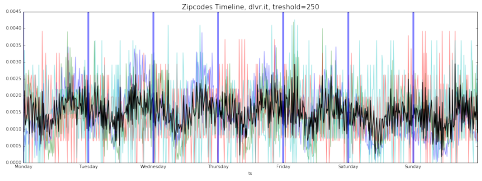

Figure 11. Average week plot for Dlvr.it app

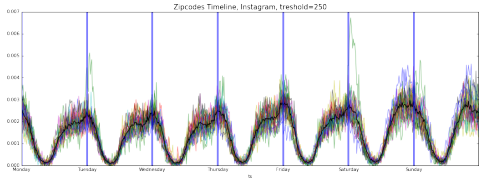

Figure 12. Average week plot for Instagram

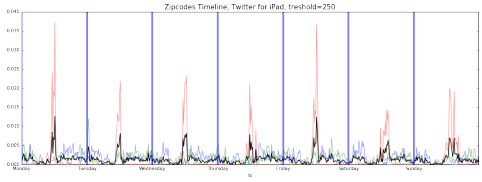

Figure 13. Average week plot for Twitter for Ipad

V. Clustering

To determine main patterns in signatures and display their spatial representation, several clustering techniques were applied. Two clustering algorithms were used: *k-mean* and *affinity propagation*. Here, in order to reduce noise, PCBs with a total number of tweets lower than certain threshold (250 tweets for average week) were not counted.

A. K-mean

K-mean is a widely used and relatively easy algorithm, that require the number of clusters to be set manually. While silhouette score may be used in order to determine the best number of clusters in our case it might make sense to start with a fairly small number of clusters in order to be able to interpret them.

As such, we started with 2 clusters, receiving expected partition of working and residential areas. Adding more clusters, we received a complete model with partitions of (as interpreted) three main areas: central business district (lower Manhattan), upper Manhattan and downtown Brooklyn, airports and all others, mostly residential areas. Both times, areas with total number of tweets below the threshold were dropped

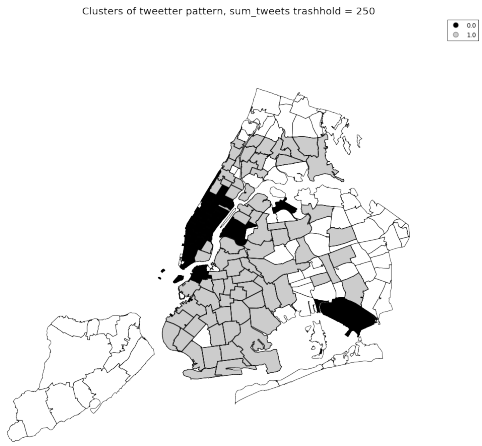

Figure 14. K-mean clustering, n=2, treshold=250

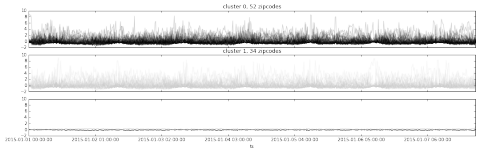

Figure 15. K-mean cluster, n=2

B. *Affinity propagation*

The second technique, Affinity propagation (AP), is particularly popular for time series. Algorithm creates clusters by sending messages between pairs of samples until convergence. Clusters are then described using most representative exemplars. AP does not require a predefined number of clusters. In practice, this algorithm is also more capable of defining unique time series as a unique singleton clusters.

For tweets time series, AP detected 12 clusters, detecting (as interpreted) clusters of Central Business District, Downtown, three clusters of residential areas, and several unique PCBs, including JFK, La Guardia, Bay Ridge, and few others. As with K-mean, areas with number of tweets below the threshold were dropped.

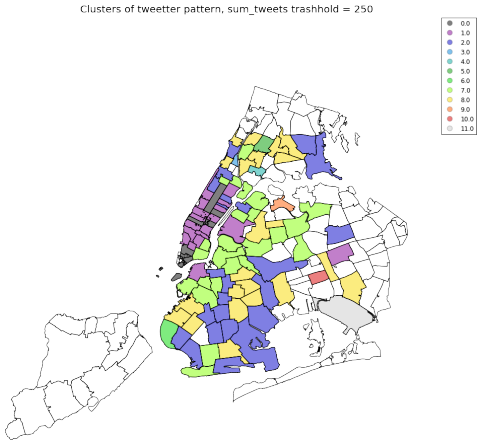

Figure 16. Affinity Propagation clustering outcomes: 12 clusters, including singleton clusters for JFK and La Guardia Airports

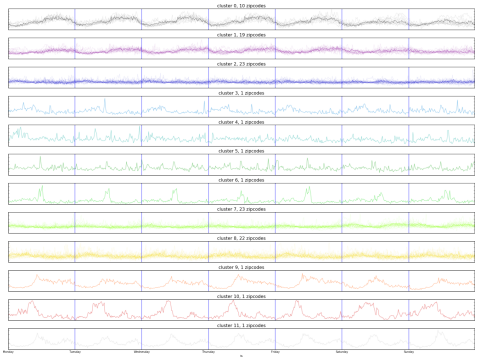

Figure 17. Affinity Propagation clustering outcomes: time Series of postal codes, grouped into 12 clusters

VI. Socioeconomic properties of the partition

As clusters were mostly formed by neighbor areas, it would be legitimate to wonder whatever those clusters represent similar socio-economic properties of neighborhoods. To understand that, we used data from last American Community Survey. For each cluster, a distribution of most important parameters was created, covering topics as race, poverty, commute time, median income and median household rent, and others. Comparing those distributions, we were able to establish a set of parameters, that has a significant difference in values from cluster to cluster. This success of clustering may lead to the hypothesis, that average week signatures might be used to "predict" those characteristics - both for Postal codes,and, in more practical case, on the more granular level. Moreover, the very same technique might be used to detect temporal changes for specific areas, indicating changes in it's characteristics and/or functional zoning.

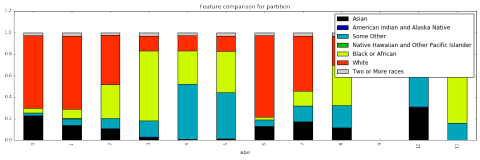

Figure 18. Racial decomposition of 12 time series clusters, defined by AP

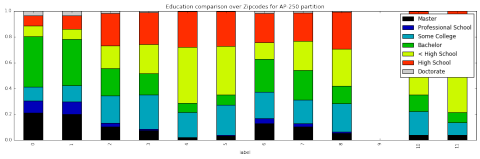

Figure 19. Highter education decomposition of 12 time series clusters, defined by AP

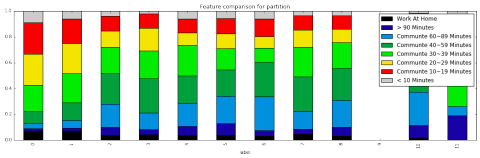

Figure 20. Commute decomposition of 12 time series clusters, defined by AP

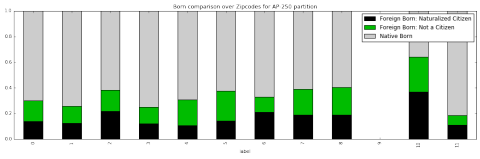

Figure 21. Foreign population decomposition of 12 time series clusters, defined by AP

VII. Prediction

VIII. Event detection: Snow storm

In January 2015, 23-31, a powerful blizzard affected the Eastern United States. On Monday, January 25, Storm approached New York. Next day, many roads were closed. As

Twitter stream represents human behavior, it would be correct to assume that certain changes from the general pattern would occur within this period. Using this great opportunity, we investigated how certain global events might be represented in social media, and if we'll be able to detect any global and local events using timeseries solely. As data for particular week is sparse, we switch to the 6-hour range for each day, and compare differences between average week and week of the storm. Indeed, we found a significant shift in patterns for two days January 25 and 26. The way pattern changes, however, differs depending on the area: for Lower Manhattan, twitter activity was significantly higher on Monday evening, and significantly lower during Tuesday work hours. On the contrary, for Flatbush neighborhood twitter activity was bigger during work hours on Monday and Tuesday evening, and had a clean spike on Tuesday morning.

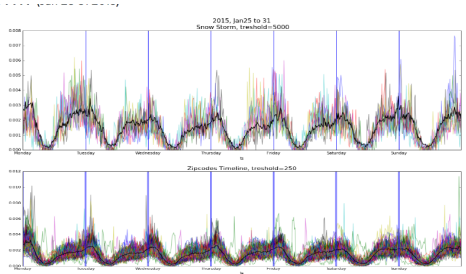

Figure 22. Snow storm week (above) versus average week (below).

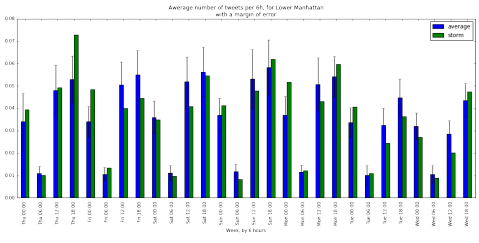

Figure 23. Average week vs Snow Storm week for Lower Manhattan

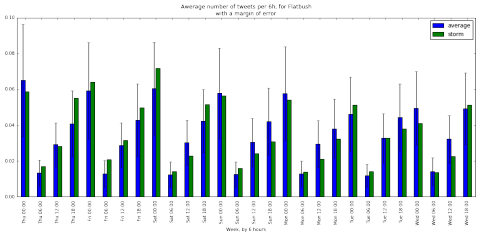

Figure 24. Average week vs Snow Storm week for Flatbush area

IX. Discussion

In this study, we explored New York areas representation in the stream of twitter social media records through building average week signatures for each area. As study shows, being cleaned and prepared, stream of social media may represent functionality as well as certain socioeconomic characteristics of urban landscape. Our results showed, that average week signature for residential postal codes differs clearly from the office/retail areas. As such, over continuous monitoring, we might outline shifts in area functionality and demographics, detect events and predict how they will influence mobility. Using clustering techniques of average week signature, we were able to group areas by their social behavior. This partition effectively represent both functional and socioeconomic properties of the neighborhoods, and, therefore, might be used to monitor changes in those properties at the detailed level. This prediction model might be improved in the future by adding signatures for specific applications. Significant events can be spotted through their twitter "fingerprints". As such, snow storm of 2015 was presented by behavioral changes both in business areas residential neighborhoods, most likely due to the problems with transportation. Any of those approaches might be turned into practical implementation, in order to reinforce decision making, resources allocation and near-real time neighborhood dynamics monitoring.

References

[1] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong." [Online]. Available: http://arxiv.org/abs/1406.4400

[2] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez, "Characterizing urban landscapes using geolocated tweets," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, pp. 239–248. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6406289

[3] C. G. W. Jacobs-Crisioni and E. Koomen, "Linking urban structure and activity dynamics using cell phone usage data," in *Workshop on complexity modeling for urban structure and dynamics, 15th AGILE international conference on Geographic Information Science, Avignon*. [Online]. Available: http://www.feweb.vu.nl/gis/publications/docs/Linking_urban_structure_and_activity_dynamics_using_cell_phone_usage_data_abstract%20AGILE2012.pdf

[4] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: using location data from cell phones for urban analysis," vol. 33, no. 5, pp. 727–748. [Online]. Available: http://epb.sagepub.com/content/33/5/727.short

[5] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti, "The Impact of Social Segregation on Human Mobility in Developing and Urbanized Regions." [Online]. Available: http://arxiv.org/abs/1401.5743

[6] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," vol. 9, no. 6, p. e96180. [Online]. Available: http://arxiv.org/abs/1311.2911

[7] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," vol. 28, no. 9, pp. 1988–2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.913794

[8] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong," in *Computational Approaches for Urban Environments*, ser. Geotechnologies and the Environment, M. Helbich, J. J. Arsanjani, and M. Leitner, Eds. Springer International Publishing, no. 13, pp. 363–387, DOI: 10.1007/978-3-319-11469-9_15. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-319-11469-9_15

[9] J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: analysing cities using the space–time structure of the mobile phone network," vol. 36, no. 5, pp. 824–836. [Online]. Available: http://epb.sagepub.com/content/36/5/824.short

[10] New York State. NYC Neighborhood ZIP Code Definitions. [Online]. Available: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

[11] U. C. Bureau. American Community Survey (ACS). [Online]. Available: https://www.census.gov/programs-surveys/acs/

[12] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, “Cellular census: Explorations in urban data collection,” no. 3, pp. 30–38. [Online]. Available: http://www.computer.org/csdl/mags/pc/2007/03/b3030-abs.html