TweetCount: Urban Insights by Counting Tweets

John Krumm
Microsoft Research AI
Redmond, WA 98052, USA
jckrumm@microsoft.com

Andrew L. Kun
ECE Department
University of New Hampshire
Durham, NH 03824, USA
Andrew.Kun@unh.edu

Petra Varsányi
ECE Department
University of New Hampshire
Durham, NH 03824, USA
petra.varsanyi@gmail.com

UbiComp/ISWC '17 Adjunct, September 11-15, 2017, Hawaii, USA.

ACM 978-1-4503-3575-1/15/09...$15.00.
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

Abstract
This paper characterizes an urban region using time series of geotagged tweet counts. Time series are constructed for each cell in a rectangular grid. We show how simple, anonymous tweet counts in the cells can be used to classify the cells into urban land use profiles based on the number of residences and businesses. We discover that Twitter activity for a certain short time of day is especially indicative of a region's profile. We go on to analyze the cells and profiles in a novel way by looking at their ability to predict tweet counts in other parts of the region.

Author Keywords
Urban computing; Twitter; location; clustering; prediction; maps; geography; New York City.

ACM Classification Keywords
J.4 Social and Behavior Sciences.

Introduction
An analysis of activity data from people in urban areas can lead to interesting insights about where they live. For instance, Handy *et al*. explored the relationship between the built environment and human behavior in urban areas [1]. Ewing *et al*. discussed the connections between urban sprawl, physical activity, and health [2]. While intentional data-gathering works well for focused

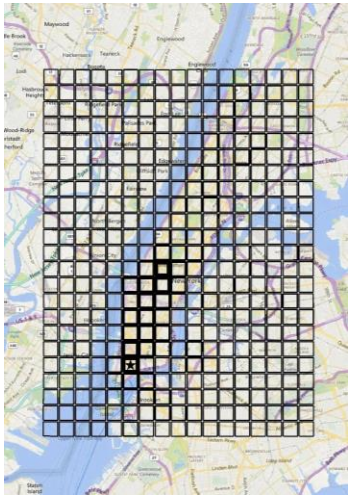


Figure 1: This grid around New York City shows the geographic extent of our tweets and the 391 1km X 1km cells we used for discretization. Thicker cell borders indicate more total tweets from that cell, with the largest counts concentrated in the south part of Manhattan. The cell with the most geotagged tweets is shown with a star.

urban studies, the ubiquity of mobile devices has led to a rich source of raw human behavior data ripe for exploration. As of 2015, nearly two-thirds of Americans owned smartphones [3].

One particularly common and informative type of mobile data is time-stamped location. At an individual level, location tracks can be used to infer a person's preferences [4] and the locations of their important places, like home and work [5]. At a collective level, logs of location can be aggregated to study travel propensity between cities [6] and classify land use [7].

This paper describes how we use location data from geotagged tweets to understand the makeup of an urban area and the relationships between the different parts. We show how simple, anonymous time series of location data from tweets can be used to classify parts of an urban area into land use profiles and how they can reveal predictive connections between different parts of a city. For each of our quantitative results, we attempt to draw conclusions about what they mean in terms of human behavior.

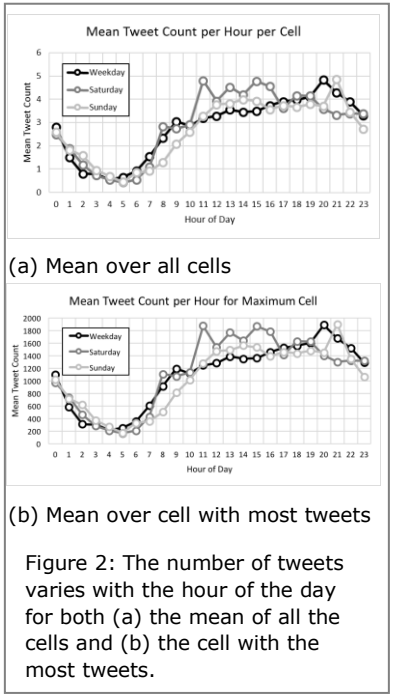
Related Work

Many social media posts come with an associated latitude/longitude measurement, which researchers have used to characterize locations. One of these characterizations is activity, such as drinking [8], restaurant health violations [9], and disease transmission [10]. There have also been efforts to automatically detect local events, such as EvenTweet [11], which exploits both the text and location of tweets, and Eyewitness [12], whose detection depends on localized, anomalous spikes in tweet activity.

Other work has attempted to characterize urban ground use. Steiger *et al*. found tweets that indicate work locations and showed that these correspond well with work locations in UK census data [13]. Looking at 24-hour profiles of tweet counts, Arribas-Bel *et al*. created clusters of similar profiles and showed how they correspond to different ground use categories [14]. Cranshaw and Yano used spectral clustering of Foursquare venue types to segment urban areas into natural neighborhoods and then found the most indicative topics mentioned in check-ins for these neighborhoods [15].

Closer to our work are projects that use social media to look at the dynamics of people moving from place to place. For example, building on Cranshaw and Yano, Cranshaw *et al*. created Livehoods, which found urban neighborhoods that are additionally sensitive to repeat visits, helping to model and detect the character of different regions of a city [16]. Noulas *et al*. also used spectral clustering from Foursquare check-ins, with features including types of venues and check-in counts [17]. They clustered users based on where they visited and found correspondingly similar regions between New York City and London. Finally, Ferrari *et al*. looked at geotagged tweets and user IDs to find crowd clusters and identify common movement patterns between clusters using a topic model that is normally used to analyze documents [18].

Our work extends previous work by looking at urban land use and location dynamics. Specifically, the analysis in this paper differs from previous work in these ways:



(a) Mean over all cells

(b) Mean over cell with most tweets

Figure 2: The number of tweets varies with the hour of the day for both (a) the mean of all the cells and (b) the cell with the most tweets.

• We look at only aggregated tweet counts over time, not the content of the tweets and not the IDs of the users. This ensures the privacy of the users, and it shows how to take advantage of similar data that could come from other anonymous sources like cell tower connections and cameras.

• Instead of trying to find land use profiles based on social media, we instead find meaningful, intuitive profiles based on known ground use data and then show that tweet counts can serve to classify a region into one of these profiles.

• For the profiles above, we show that tweet counts in one location are temporally predictive of tweet counts in another location.

Prior to describing our analysis, we summarize the data we used.

Twitter Data

Twitter, the popular microblogging site, supports an optional latitude/longitude field for each tweet. These geotags typically come from a location sensor on the mobile device from which a user is tweeting. Morstatter *et al*. estimate that 1.45% of tweets from Twitter's firehose are geotagged [19], while Wantanabe *et al*. put the fraction at 0.7% [20].

We gathered approximately four million geotagged tweets spanning four months in mid-2015, bordered by the grid covering New York City shown in Figure 1. We chose this region because it covers a diverse range of urban areas, and its large population posts many geotagged tweets. Each cell is 1 km on a side. The thickness of the borders of the cells in Figure 1 indicate the relative number of tweets in each cell. The cell with the most tweets is shown with a star inside. This cell contains New York's city hall and one end of the Brooklyn Bridge. It is near, but does not include, the World Trade Center and Wall Street.

As an initial analysis, we can examine how the number of geotagged tweets varies with time. Figure 2(a) shows the mean number of tweets per hour over all the cells. The three curves show mean counts for weekdays, Saturdays, and Sundays over our four-month time period. The absolute numbers are relatively low (always less than five), because many of the cells are in water or other sparsely populated areas. As expected, there are few tweets in the early a.m. hours when people are generally sleeping. There is a gradual rise in the morning to a steady rate during the afternoon and evening, with a drop after about 8 p.m. We note that this variation of tweet counts over time is similar to the pattern found in Amsterdam in [14].

Figure 2(b) shows the mean number of tweets in the cell with the maximum number of tweets (shown with a star in Figure 1). In this cell there are many more tweets than the overall mean cell in Figure 2(a). Despite the magnitude difference, the shapes of the profiles for the mean cell and maximum cell are similar.

The following sections present a more quantitative analysis of what we can learn by looking at how tweet counts change over time and location.

Land Use Classification with Tweet Counts

We are interested in what tweet counts can tell us about the land use profile of an urban region as well as what features of tweet counts are most informative

I	%-ile Range	Resident Count	Business Count
0	0.0–10^{-4}	0-19	0-21
1	10^{-4} - 0.2	20-9299	22-3061
2	0.2-0.8	9300-31,043	3062-47,854
3	0.8–1.0	31,044-47,360	47,855-70,497

Table 1: Residences and businesses in land use profiles.

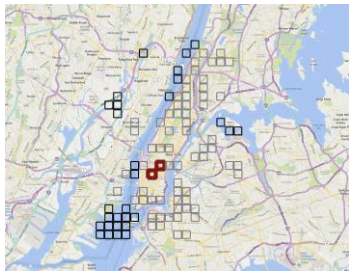


Figure 3: These are three of the distinct land use profiles. The black cells show R0-B0, with very few residences and businesses. The gray cells are R2-B2, which have a medium number of both types. The two cells with thick outlines are R3-B3, which have the most residences and businesses.

about this profile. For instance, we suspect that tweet counts from a primarily residential region will differ in magnitude and timing from a primarily business region, and that the magnitude of tweet counts at night after normal business hours may be particularly informative at making the distinction between business and residential regions.

Land Use Profiles from Residences and Business Counts
The land use profile of cells in our grid can be roughly characterized by the number of residences and businesses contained in each one. Using a dataset of residence and business locations from Bing Maps, we computed these numbers for each cell.

We divided the numbers of residences and businesses into percentile ranges to create discrete land use profiles. Our four percentiles are shown in Table 1. We chose these percentiles to highlight different types of regions in our study area. The first percentile is very narrow, and it is meant to specifically cover regions that are essentially uninhabited, such as open water and land unsuitable for building. There are still a few residences and businesses assigned to these areas from our database, primarily due to noise and mistakes. The second percentile covers approximately the first 20% of the cumulative distribution, with a thin portion removed from the lower end to account for near-zero noise of the first percentile. This lower 20% represents sparse urban land use. The third percentile represents the middle 60% of the distribution, accounting for the most common land use regions. Finally, the fourth percentile covers the top 20%, indicating the highest density land use. We use these percentiles as a simple, intuitive split of the distributions, although the percentiles could be adjusted depending on the application.

With four percentiles each for residences and businesses, there are 16 possible sets of cells. A set of cells whose number of residences is in the second percentile and whose number of businesses is in the third percentile would be named R1-B2, and its cells would have a relatively low number of residences (R1) and a medium number of businesses (B2). Cells in R0-B0 have nearly zero residences and businesses, and they are primarily those cells that cover nothing but water. In Figure 3 we show the cells that make up three of the 16 possible land use profiles, including R0-B0 (mostly water), R2-B2 (medium number of residences and businesses), and R3-B3 (large number of residences and businesses). We refer to each Ri-Bj as a separate urban land use profile. Figure 4 shows the number of cells in each profile. Of the 16 possible profiles, 7 have greater than 1% of the cells, and these 7 cells account for 98% of the total cells.

Tweet Counts and Land Use Profiles
We are interested in the relationship between tweet counts and land use profiles. To explore this relationship, we used tweet counts as classification features for the seven profiles that had at least 1% of the geographic cells. Successful classification means that tweets counts are indicative of land use.

For each of these seven profiles, we built a one-vs.-all binary classifier to distinguish the cells with that profile from all the other cells. Each classifier used the same features: mean tweet counts for every hour of the day for weekdays (24 features), Saturdays (24 features), and Sundays (24 features). An example of these features for one cell is shown as a plot in Figure 2(b).

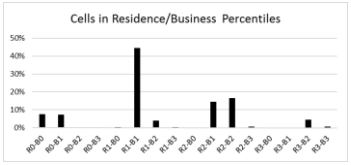


Figure 4: Seven of the 16 land use profiles have a significant fraction of cells.

The seven binary classifiers were each a FastRank decision tree, which is an efficient version of the MART gradient boosting algorithm. It learns an ensemble of decision trees, where the next tree in the ensemble is designed to correct the mistakes of earlier trees [21]. We evaluated performance using two-fold cross validation.

The performance results of the seven binary classifiers is shown in Figure 5. Classification was generally better for profiles with more cells. Weighted by the number of cells in the profiles, the overall weighted classification accuracy was 81%, weighted positive precision was 56%, and weighted positive recall was 49%. Thus, tweet counts work as distinguishing features for land use profiles.

We can speculate on why tweet counts are indicative of our profiles. It is likely that the overall magnitude of counts is broadly indicative of land use, with small counts indicating smaller numbers of residences and businesses, and similarly for larger counts. The time distributions of tweet counts may be associated with the opening and closing hours of businesses, with some businesses open only during the day and others being more active at night.

We queried our seven binary classifiers for the relative number of times each feature was used in their respective decision trees. Each weekday feature, on average, was used 17% of the time in the decision trees, while Saturday and Sunday features were used 8% and 7%, respectively. The most important two features are weekday tweet counts at hours zero and one (midnight to 1 a.m. and 1 a.m. to 2 a.m. in local time), each used about 52% of the time in the decision

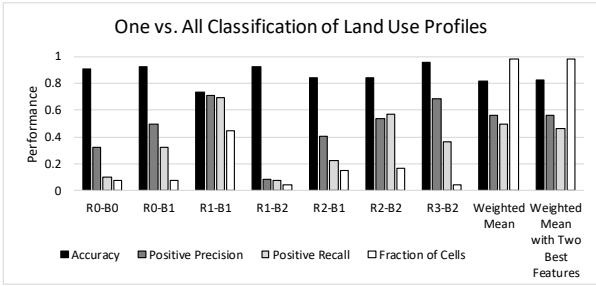


Figure 5: Classification performance for tweet counts was generally good, with larger regions performing better. The weighted mean shows overall performance weighted by the fraction of cells with each of the seven profiles.

trees. Rerunning the two-fold cross validation tests using only these two features resulted in weighted average performance that was nearly indistinguishable from the original analysis using all 72 features. (See last entries in plot of Figure 5.) This surprising result shows the discriminative power of weekday tweet counts from midnight to 2 a.m. This implies that doing the same classification task using webcams or aerial imagery should include features that can be derived from images taken after sundown, such as vehicle headlights and urban illumination.

Interactions Between Land Use Profiles

We are interested in assessing how tweet counts change over time in different land use profiles, and how these changes are related between different profiles. We look at these interactions in terms of the predictive power of the profiles, measuring how well past counts can predict future counts. This is similar in spirit to work on predicting road traffic [22] and internet traffic [23] based on correlations between measurements of the same variable at different times and places.

In order to explore this question we use vector autoregression (VAR) [24]. A VAR model of order N describes the values of an $M\times1$ dimensional vector $y[n] = [y_1\ y_2\ \dots y_M]^T$ at time step n as a weighted sum of the values of vector $y[n]$ at N previous time steps:

$$y[n] = c + A_1 y[n-1] + \cdots + A_N y[n-N] \qquad (1)$$

In equation 1, the vectors c, and $y[n-i], i \in \{0, 1, 2, \dots, N\}$, are $M\times1$ dimensional vectors, where each dimension represents a different variable. In our case each dimension of $y[n]$ represents the tweet counts in a specific profile's set of cells. The $M\times M$ dimensional matrices $A_i, i \in \{1, 2, \dots, N\}$ represent the influence of each delayed vector $y[n-i]$ on the current vector $y[n]$. These matrices can be estimated using tools such as multivariate least-squares [24].

We created sequences of tweet counts for each of the seven urban profiles introduced in the previous section. Each tweet count represents the number of tweets in a 30-minute-long window. We applied a median filter with window size 3 to the tweet counts, in order to eliminate spikes in tweet counts, which might be due to Twitterbots. The result was a sequence of 7×1 dimensional vectors $y[n] = [y_1\ y_2\ \dots y_7]^T$, where an increment in the value of n represents an increment of 30 minutes of time. Our data covers 123 days between May 1 and August 31, 2015. Thus, we have a total of $123\ days \times 24\frac{hours}{day}\times2\frac{tweet\ counts}{hour} = 5{,}904$ tweet counts for each cluster. We selected $N = 5$, which means that our VAR model uses 5 prior values (2.5 hours) of $y[n]$ to predict the current value of $y[n]$. This value of N means that our VAR model will rely on the rapid changes in tweets, and not on the periodic 24-hour changes.

Next, we asked the following question: Which profiles are best at predicting tweet counts among the seven profiles? We wanted to first find the answer to this question for the case when we use only one profile to predict another profile's tweet counts. We did this by using a simplified version of the VAR model, where each profile's tweet count is predicted as a weighted sum of prior tweet counts from only one of the seven profiles. Such a simplified model would be called auto-regressive (AR) rather than full vector auto-regressive (VAR).

The results of this analysis are summarized in Table 2, where individual cells represent the root-mean-square (RMS) error of predicting tweets in a target profile using only one of the seven possible profiles. The results indicate that each profile is its own best predictor. The size of this effect is considerable: using the second-best profile to predict the target profile on average increases the RMS error by a factor of 1.8. This finding is consistent with our results that indicate that simple tweet counts serve as good discriminator functions for land use profiles. However, the tweet count result for classification (not prediction) only indicates that there are times in the 24-hour daily cycle when average Twitter activity is different and informative among the profiles. The AR analysis does more – it indicates that the dynamics of Twitter activity is determined to a large extent by underlying differences in the land use profiles. Each profile is somewhat independent in how it behaves, at least in terms of tweet counts.

Table 2 also shows that R1-B1 and R3-B2 were consistently among the best profiles to use in predicting other profiles. R1-B1 was ranked as the 2nd or 3rd best

		RMS error (in tweets/30 minutes) when predicting target using one profile						
		R0-B0	R0-B1	R1-B1	R1-B2	R2-B1	R2-B2	R3-B2
Profile to predict	R0-B0	**1.19**	2.04	1.66	1.72	1.83	2.00	1.79
	R0-B1	2.01	**0.85**	1.96	1.96	2.00	2.02	1.96
	R1-B1	16.47	19.89	**7.45**	14.14	14.98	19.55	13.74
	R1-B2	15.58	17.26	13.42	**4.94**	15.60	16.83	15.02
	R2-B1	9.88	11.02	7.54	9.53	**4.86**	10.92	7.49
	R2-B2	62.64	72.59	47.18	53.16	54.14	**39.57**	45.10
	R3-B2	31.00	36.24	21.93	28.39	25.00	33.89	**11.89**

Table 2: Root mean square (RMS) prediction error of tweet counts in the seven land use profiles. We count tweets in 30 minute time periods, thus the RMS error is also given in units of [tweets/30 minutes]. The lowest error is bolded in green cells, the second lowest error is underlined in blue cells.

single profile to predict all the other six profiles, while R3-B2 was ranked as 2nd or 3rd for five of the six. In contrast, R0-B1 was consistently the worst (7th) profile to predict all six other profiles, while R2-B2 was ranked 6th or 7th. This implies that some land use profiles are consistently better than others at driving the dynamics of an urban region.

We also assessed the prediction error when two profiles are used to predict a target profile, moving from a simple AR model to VAR. Since we found that profiles predict themselves the best, we always predicted profiles with profile pairs that include the profile to be predicted. Adding a second profile does not reduce the prediction error by much: the improvement is about 5% on average. This again implies that the human dynamics of land use profiles are somewhat independent of each other, driven mostly by their own internal dynamics.

What is it that makes R1-B1 and R3-B2 more valuable in predicting other profiles, and in contrast why are predictions less accurate when using R0-B1 and R2-B2? R0-B0 and R0-B1 have very low tweet counts, about 2 and 1 tweets per 30 minutes, respectively. This is visualized in Figure 6, where the mean tweet count per 30-minute time increment for R0-B0 and R0-B1 is dwarfed by the counts in the other clusters. Such sparse data is likely to be insufficient to accurately predict the variability in profiles with significant Twitter activity. R2-B2, on the other hand, has the largest number of tweets, but we see in Figure 7 that the mean daily changes have prominent peaks between 3 a.m. and 5 a.m., around 2 p.m., and around 8 p.m. (local time). These local peaks are either not present in the other profiles, or are less pronounced, and as such R2-B2 cannot be used to predict the other profiles very accurately.

Figure 6 also shows why R1-B1 and R3-B2 are good at predicting other profiles. Both have a substantial number of tweets, and both are smooth functions that closely resemble the general 24-hour periodic trends of the other profiles: a gradual rise in Twitter activity that begins around 5 a.m., and a reduction in Twitter activity that begins around 8 p.m.

Conclusion

Anonymous geotagged tweets can reveal interesting insights about the aggregate land use profiles and dynamics of an urban area. Although ignoring user identifiers makes the analysis less exact, our approach shows how to extract interesting insights while still

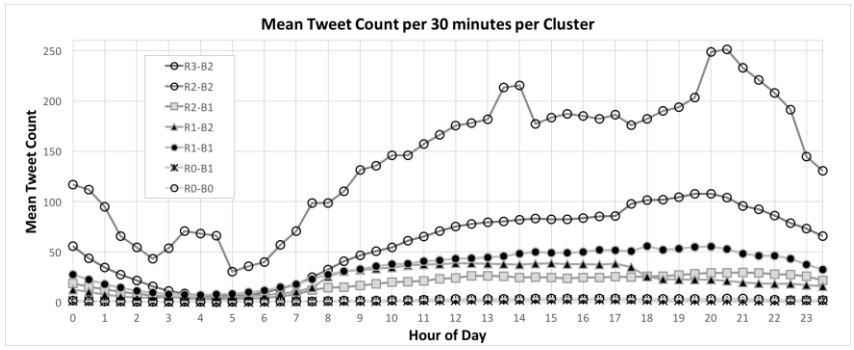


Figure 6: The number of tweets varies with the hour of the day for all seven clusters.

preserving privacy. Our examination of land use profiles showed that tweet count features can be used to classify an urban area into different regions based on their relative numbers of residences and businesses. Looking at the specific features, we found that simple, average counts on weekdays between midnight and 2 a.m. are powerful classification features. Our VAR analysis is a simple, principled way of discovering predictive relationships between the human dynamics of different parts of an urban area. We found that land use regions are best predicted by past values of their own tweet counts and that certain profiles are more predictive than others.

References

1. Handy, S.L., et al., How the built environment affects physical activity: views from urban planning. American journal of preventive medicine, 2002. 23(2): p. 64-73.
2. Ewing, R., et al., Relationship between urban sprawl and physical activity, obesity, and morbidity, in Urban Ecology. 2008, Springer. p. 567-582.
3. Smith, A., U.S. Smartphone Use in 2015, 2015, Pew Research Center.
4. Froehlich, J., et al., Voting with your feet: An investigative study of the relationship between place visit behavior and preference, in Eighth International Conference on Ubiquitous Computing (UbiComp 2006)2006, Springer: Orange County, California, USA. p. 333-350.
5. Liao, L., D. Fox, and H. Kautz, Extracting places and activities from gps traces using hierarchical conditional random fields. The International Journal of Robotics Research, 2007. 26(1): p. 119-134.
6. Krumm, J., R. Caruana, and S. Counts, Learning likely locations, in Twenty-First Conference on User Modeling, Adaptation and Personalization (UMAP 2013)2013, Springer: Rome, Italy. p. 64-76.
7. Pan, G., et al., Land-use classification using taxi GPS traces. Intelligent Transportation Systems, IEEE Transactions on, 2013. 14(1): p. 113-123.
8. Hossain, N., et al., Precise Localization of Homes and Activities: Detecting Drinking-While-Tweeting Patterns in Communities, in 10th International AAAI Conference on Web and Social Media (ICWSM-16)2016. Cologne, Germany.
9. Sadilek, A., et al. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. in Twenty-Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI 2016). 2016. Phoenix, Arizona USA.
10. Sadilek, A., H.A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. in Twenty-Sixth AAAI Conference on Artificial

Intelligence (AAAI 2012). 2012. Toronto, Ontario, Canada.

11. Abdelhaq, H., C. Sengstock, and M. Gertz, Eventweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment, 2013. 6(12): p. 1326-1329.
12. Krumm, J. and E. Horvitz, Eyewitness: identifying local events via space-time signals in Twitter feeds, in Twenty-Third International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2015)2015: Seattle, Washington USA.
13. Steiger, E., et al., Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. Computers, Environment and Urban Systems, 2015. 54: p. 255-265.
14. Arribas-Bel, D., et al., Cyber Cities: Social Media as a Tool for Understanding Cities. Applied Spatial Analysis and Policy, 2015. 8(3): p. 231-247.
15. Cranshaw, J. and T. Yano. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. in CSSWC Workshop at NIPS. 2010.
16. Cranshaw, J., et al. The livehoods project: Utilizing social media to understand the dynamics of a city. in 6th International AAAI Conference on Weblogs and Social Media. 2012. Dublin, Ireland.
17. Noulas, A., et al., Exploiting semantic annotations for clustering geographic areas and users in location-based social networks, in AAAI Workshop on the Social Mobile Web2011. p. 32-35.
18. Ferrari, L., et al. Extracting urban patterns from location-based social networks. in 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. 2011. Chicago, Illinois USA: ACM.
19. Morstatter, F., et al., Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose, in Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)2013: Boston, Massachusetts, USA.
20. Watanabe, K., et al. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. in Proceedings of the 20th ACM international conference on Information and knowledge management. 2011. ACM.
21. Friedman, J.H., Greedy function approximation: a gradient boosting machine. Annals of Statistics, 2001. 29(5): p. 1189-1232.
22. Min, W. and L. Wynter, Real-time road traffic prediction with spatio-temporal correlations. Transportation Research Part C: Emerging Technologies, 2011. 19(4): p. 606-616.
23. Barthélemy, M., B. Gondran, and E. Guichard, Large scale cross-correlations in Internet traffic. Physical Review E, 2002. 66(5): p. 056110.
24. Lütkepohl, H., New introduction to multiple time series analysis. 2005: Springer Science & Business Media.