

URBAN PULSE ANALYSIS USING BIG DATA

Galo Reinoso and Bilal Farooq, Polytechnique Montreal

1 Introduction

Cities have seen a vast amount of transformations in the past few years. Smartphones have proliferated. Reliable Internet connection is now available in many places. Car-sharing and bike-sharing are getting more popular. Social interactions keep on evolving with services like Facebook, Twitter and Instagram. Smartphone ownership and social networks are the foundation of location based services which enable real time information exchange. While users publicly share their informations, companies are doing the same by opening their datasets to the public.

This is in stark contrast with the methods in travel surveys that have changed more slowly. Contacting the people by phone is still the main method of gathering data. Moreover, they are costly to do and they do not give an up-to-date portrait of the current situation since they are done once or twice a decade.

The usage of these new datasets as a solution to current limitations of travel surveys is the main objective of the current research. As an initial step, using big data from Bixi and Twitter, the goal is to describe the travel and activity patterns of people living in urban centers and their evolution in time and space. The challenge rests in automating the data collection process and in assessing the reliability of such uses for these datasets. An explanation of the data gathering and processing is presented. The datasets are first analyzed separately and then jointly.

2 Literature Review

2.1 Geolocalized Mobile Data

Social Media is nowadays widely used among most North Americans. This phenomenon put at the forefront User Generated Content (UGC). This type of content needs to respect three criterias: a) it has to be published on the internet and be accessible to at least a group of people b) it need to be original content c) it needs to have been created outside of professionn constraints such as an advertisement or by a business (Kaplan and Haenlein, 2010). Twitter enables this type of original information by having a unique space and time component in its metadata even though the text is retweeted.

In fields other than Transportation, Tweets have already been used in order to detect earthquakes events by using the location attribute and detecting relevant text in them (Sakaki et al., 2010). In the entertainment industry, text analysis from tweets distributed over time was used to forecast box-office revenues (Asur and Huberman, 2010).

Using primarily call detail records from mobile phones in Boston and Rio de Janeiro, Colak et al. manages to generate origin-destination (OD) trips for those regions indicating the reliability of the patterns observed in the dataset to assess regular commuting trips (Colak et al., 2015).

2.2 Bike-sharing

Bike-sharing systems are becoming increasingly popular in urban centers. This system, which first appeared in 1965 in Amsterdam, is now used in many North American and European cities. The actual iteration of the system is different from previous ones mostly through the implementation of smart tools: smartcards, mobile phone access, electronically locking docks and telecommunications to and from the docks. The impacts of this service range from increasing the bike mode share, to increasing the usage of transit and reducing the number of vehicle trips (DeMaio, 2009).

In the case of Montreal Bixi system, an analysis made with data gathered in 2009 shows that 67.8% of the trips are made by members. During weekdays, they made 72.6% of the trips. A number which decreases on weekends. The average age was 34 years old with a men to women

ratio of 1.73:1. This consists of an over-representation of the 25-34 years old group and a greater one for men, among members. Like other bike-sharing systems, Bixi suffers from a logistical problem that arises from this service: the rebalancing of bikes along the network. Therefore, bikes are loaded and unloaded by employees of the service whenever the need arises. In some stations, 80% of the incoming bikes can be attributed to an employee manipulation (Morency et al., 2011). As a solution to this, a method has been developed by Imani et al. (2014) with data from Bixi collected in 2012.

2.3 Open Data

The previous paradigm in data distribution was to first define the needs of the dataset and then the data was created in order to fulfill them. While this "top-down" model scheme is desirable for the data manager, it is less adapted for the constantly evolving information available online (Auer et al., 2007). This dynamic particularity pushes towards a new model for data structure and management. Since the needs change rapidly, providing limited and prefiltered datasets are no longer relevant. More and more organizations decide to simply make the raw data available and let people organize and mold it as they prefer.

2.4 Clustering

Many methods exist to cluster datasets based on features of various types (Berkhin, 2006). However, few of them are focused on the density of spatial data. These density-based algorithms create clusters arbitrarily shaped which contain data that is similarly spread in space. The DB-SCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is designed to achieve this kind of regrouping with low dimensional data. It needs two input parameters: Epsilon and MinPts. The former states the maximal distance between two points to be in the same cluster. The latter is the minimum number of points needed to form a cluster. Otherwise, they are considered noise. Considering the data available, this algorithm is selected for the clustering process.

3 Data Gathering and Processing

3.1 Collection

The data was collected directly from both Bixi and Twitter. Each service allows for a connection to be established and data to be collected whenever it is publicly available. This process was automated to allow for a continuous data collection process. Moreover, two computers were collecting data to avoid data loss. Each source has its own file format, which was parsed to collect relevant information [Table 1].

Source	Format	Frequency
Twitter	JSON	Continuous Stream
Bixi	XML	On Network Change

Table 1: *Data sources*

The information collected had space and time attributes. Depending on the data provider, the coordinates and the timestamps of the data change for different reasons. In order to analyze network activity, a metric for each dataset was selected. For Twitter, a single tweet represent a unit of activity. Meanwhile, for Bixi, the data source does not allow the tracking of a single user. A more complete dataset would have been ideal. However, each station gives information on the number of bikes available and empty docks. Therefore, the departed bikes are estimated by computing the change in number of bikes available.

Source	Coordinates	Time Stamp	Metric
Twitter	Provided by the mobile phone, at the moment the tweet is sent	Whenever the tweet is sent	A tweet
Bixi	Station location	When the station usage changed	Computed departures

Table 2: *Data Description*

3.2 Processing

An exploratory analysis showcased irregularities in both datasets. The tweets dataset was striped of data that was not on the Montreal island or

that did not have coordinates but was assigned to Montreal by Twitter itself through the user expressed location. To account for Bixi relocating activities, the data set was parsed to remove any unusual spike in usage for a station. Inspired by the method in Imani et al. (2014), the data is aggregated every two minutes. The assumption is then made that a relocating operation is made when the number of departures is greater than the 95th percentile for the same station. When it is the case, a departure rate is instead computed for the station. This rate is the average of the previous five two minutes periods. To avoid any redundancy, both datasets were checked for any duplicate. The final set for the analysis in this paper includes a little more than 125,000 tweets from May 7 2014 to May 31 2014 and a little more than 470,000 status changes for all the Bixi stations.

4 Data Analysis

4.1 Bixi

The processed datasets give a good idea of the usage patterns of the bike-sharing system with the computed departures for each station [Figure 1]. On weekdays there are clearly two usage spikes, one for the morning commute and another one for the afternoon. On weekends there is a single peak which is gradually increasing during the day then peaks in the afternoon. Moreover, as expected, the system is less used at night, both during weekdays and weekends.

On weekends the daily usage remains the same, with an increase of only 4% in comparison to an average weekday [Table3]. This suggests that the network usage is not heavily skewed to one type of day and trip purpose. In fact, this is an indication of the two types of users of this service: the commuters and the recreationals. The commuters use the service on weekdays to go to work or school. The recreational users, however, use the service on weekends as a relaxing way to travel without having the time constraints a worker or student has. Hence, taking into account this increase in users on the weekend and considering that most weekday users will still use the service on weekends, it is expected to see the same amount of users or even more on weekends.

There are also two very apparent drops in the departures during week-

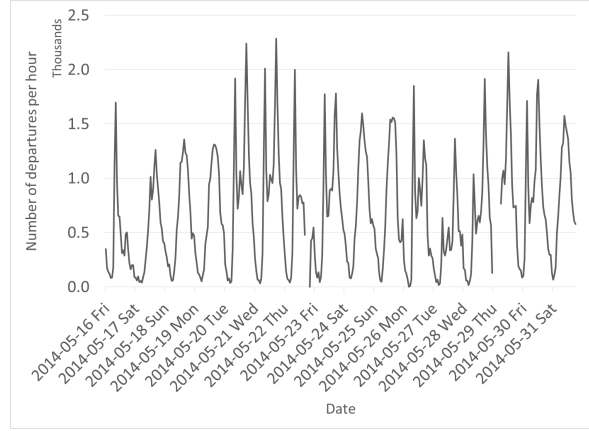


Figure 1: *Bixi departures aggregated by hour*

Day	Departures
weekday	15 530
weekend	16 175

Table 3: *Average usage per type of day*

days [Figure 2]. The data collection relies solely on the Internet. Therefore, a lost connection means that it is either the client or the server that is at fault. Since the data collection was made using two computers and no apparent interruption of the data collection process was registered, the disconnection was on the server side. This raises questions about the reliability of the data collection in this manner.

Moreover at the beginning of the 2014 season, the temperature was not bike-friendly. The cold and rainy weather reduced the occasions to use the bike-sharing system in an enjoyable manner for the users. In fact, the effect of temperature as explained in Miranda-Moreno and Nosal (2011) is visualized in Figure 3. The cumulated departures for the day are plotted against the distribution of the temperature for the same period. The weekends are highlighted and May 19 was a holiday (National Patriots' Day). Displayed this way, the two drops of usage are now easier to explain. They both correspond to a weekly minima of average temperature. They also both happen on days where the temperature drops by 5 °C to 10 °C in comparison with the previous days.

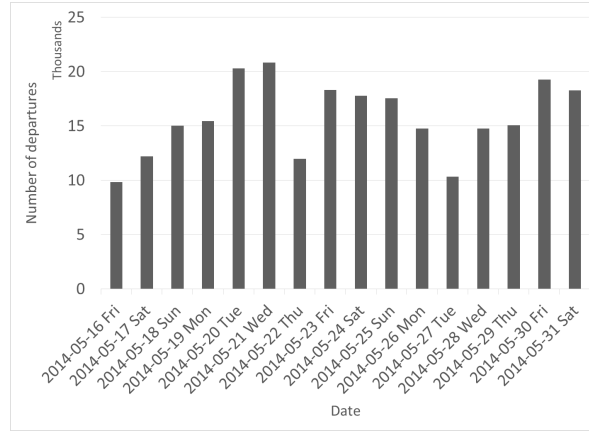


Figure 2: Bixi departures aggregated by day

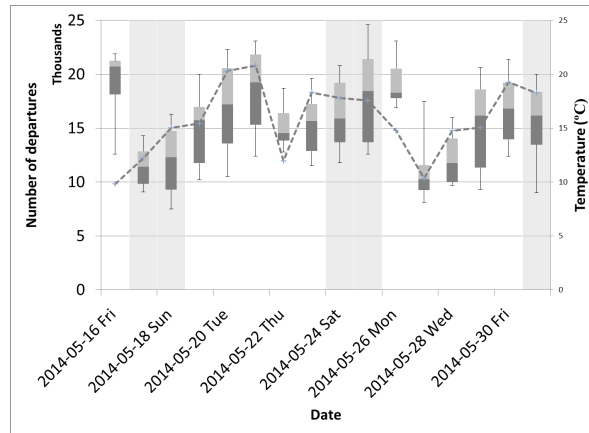


Figure 3: Departures of bixis and weather distribution

When plotted separately, the weekdays among them have a different usage rate even though the overall distributions are similar [Figure 4]. The morning and afternoon commute of a typical workday are both distinguishable. There is also an apparent usage surge around noon. This, in accordance with the average work day, implies the usage of the system in order to enjoy lunchtime by biking to get a meal or to enjoy it at a certain place. There is a noticeable difference among weekdays

where the curve for Friday presents a bigger usage in the morning than the afternoon one. It can be explained by users who take the system as a more relaxing way to get to work on the last day of the weekday.

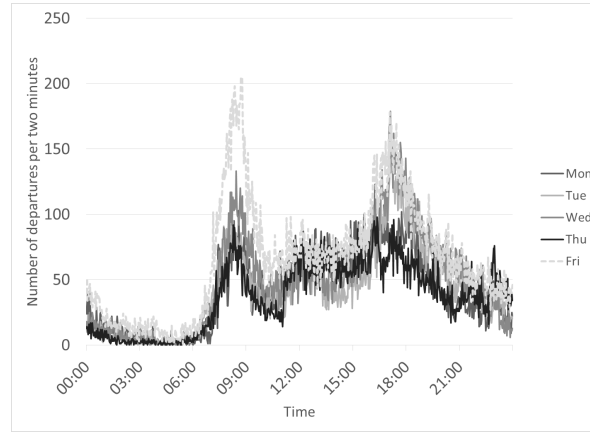


Figure 4: *Bixi departures on weekdays*

On weekends the behavior is different there is no concept of commuting [Figure 5]. After reaching a low usage between 5:00 and 6:00, it gradually increases to reach its maximum at around 15:00. Moreover, Saturday is a busier day than Sunday. This type of transportation pattern is normal since the purpose of the travel has changed for most people from working and studying to a recreational one. Therefore, the usage is spread throughout the day with a slight bump and not a sharp peak. In the 2008 Montreal OD survey, there is a very small sample of the people who biked during the months of September to December 2008. This sample claims to be representative of the average behavior over a long period of time. Meanwhile, the Bixi data has the complete usage of the system for the last two weeks of May 2014. While the bicycle usage is not similar in term of absolute numbers between the OD survey and the Bixi data, a normalization by the average usage aggregated for every 15 minutes for a weekday makes it possible to compare both datasets [Figure 6]. For the OD data, a moving average of the five previous points is used to eliminate fluctuation caused by the departure times that are frequently assigned at 15 minutes intervals.

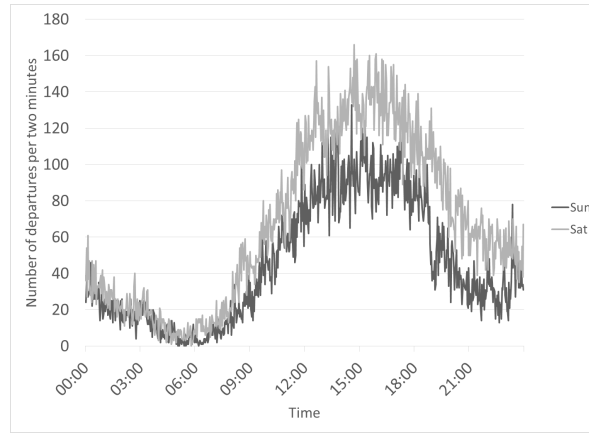


Figure 5: *Bixi departures on weekends*

Two differences arise from this figure. First, they do not capture the morning commute the same way. For Bixi, this is either because there are fewer bikers in the morning or there are more of them during the rest of the day, in comparison with the bikers from the OD survey. The latter assumption is preferred because bike-sharing users are also recreational users, not only commuters.

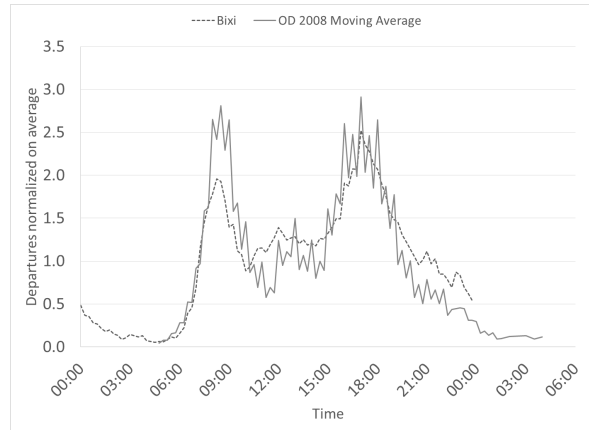


Figure 6: *Bixi departures vs OD bike trips normalized by average value*

4.2 Twitter

The processed dataset collected from Twitter is plotted on a time scale in Figure 7. There is no clear definition between the weekend and the weekdays in term of usage. In fact, the usage patterns correspond to moments where people want to express something they feel it is worth sharing. This could be because of an activity they are taking part of or an event that happened unexpectedly. For example, on the scale of the Montreal area, events of such magnitude during the last two weeks of May 2014 include the Montreal Canadiens participating in the hockey playoffs. Surge of the number of tweets are seen on both days i.e. May 19 and 22 where the Canadiens played the New York Rangers. While this do not explain all the differences between days of high and low Twitter usage, it showcases the connection between change in activities and number of tweets. The need for a more thorough investigation on the impact of important activities such as concerts, festivals, sporting events, etc. taking place in urban centers could indicate a correlation with the number of tweets.

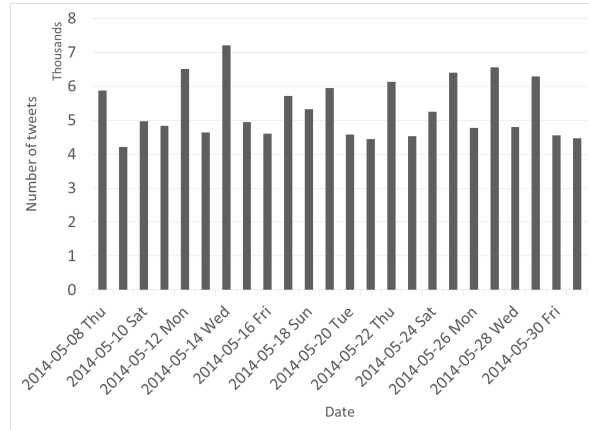


Figure 7: *Twitter activity by day*

On weekdays, starting at around 5:00, the rate increases to reach a maximum at around noon, decreases until the late afternoon/early evening, increases again until late evening then decreases gradually throughout the night [Figure 8]. The increases occur at the same time when a

change in activity worth sharing happens. At noon, people go from working to eating, sharing picture of their food, commenting about the place they are eating at or about the people they are with. While in the evening, either at home or in another place, people engage in other types of activities. More so, the spike in late evening occurs before most people go to sleep. In fact, where there are normally two peaks corresponding to commuting periods in trip patterns, here these are rather two low points of tweet activities.

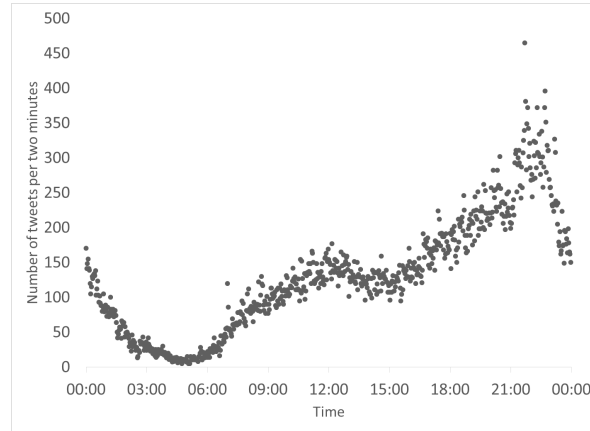


Figure 8: *Tweet frequency on weekdays*

The same patterns are generally observed during separate days of the week. When the patterns are different, it is not because of the day of the week but rather because a particular event happened on that day.

During the weekend, the pattern is different from the weekdays on two points [Figure 9]. The first maximum is reached at a much later time i.e. 15:00 instead of 12:00. People wake up later on weekends and start doing activities later too. In the late evening, there is no sudden increase in Twitter activity level. However, this last increase is sustained for a longer period of time while users are still awake.

In other words, both weekday and weekend patterns clearly show that Twitter usage is associated with recreational and non-work activities.

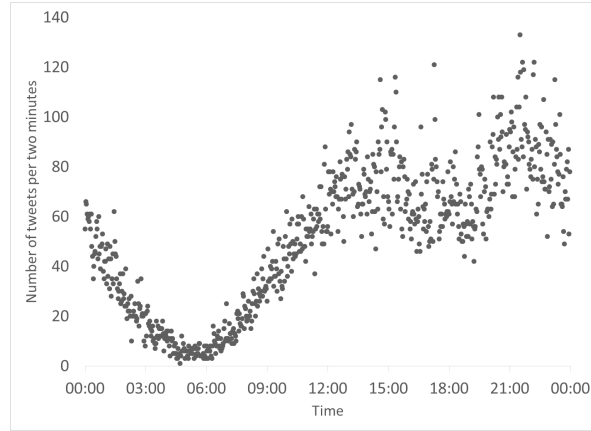


Figure 9: *Tweet frequency on weekends*

4.2.1 Clustering

On top of the time attribute of a tweet, the addition of its coordinates allow for spatial analysis over a period of time. Since the goal is to follow the pulse of the city, a natural way of displaying this behavior is by regrouping the data into groups that share similar characteristics. These are not constrained by political delimitations but rather by the level of Twitter activity and the density of activity levels.

The clustering of tweets for the whole period spurs twenty-eight clusters scattered around the island [Figure 10]. They were created using the DBSCAN algorithm. The clustering process is applied to the 129 820 tweets and 70.1% of them are assigned to a cluster, the rest being characterized as noise. Of those 90 170 tweets, 55.7% are reunited in the biggest cluster. Colored in black, it encompasses many boroughs such as Plateau-Mont-Royal, Ville-Marie, which contains the Central Business District (CBD), and Notre-Dame-de-Grâce (NDG).

The huge amount of tweets in that area is a display of the amount of activity that it generates. The dominance of this cluster over the others confirms the label of “Core of the City” to this area. In fact, the second biggest cluster only contains 8.1% of the clustered tweets [Table 4].

As the rate of the activities in the urban center changes hour after hour, so do the clusters characteristics. The variation of their size

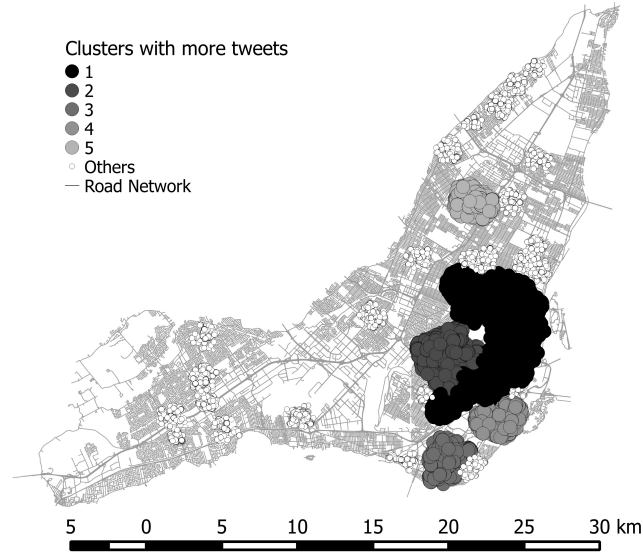


Figure 10: *Clusters for the whole period*

Cluster	Tweets	%
1	50 675	55.7
2	7 384	8.1
3	3 697	4.1
4	3 431	3.8
5	2 485	2.7
Others	23 297	25.6

Table 4: *Summary of the biggest clusters*

over time, similar to a heartbeat, enables the visualization of the urban pulse. After overlaying the five biggest clusters created for each hour and applying a transparency to each layer, the main activity centers are located [Figure 11]. The biggest one is similar to the one from Figure 10 with the four others being of similar size. The parameters for the clustering are adjusted to take into account the amount of tweets that is different hour after hour.

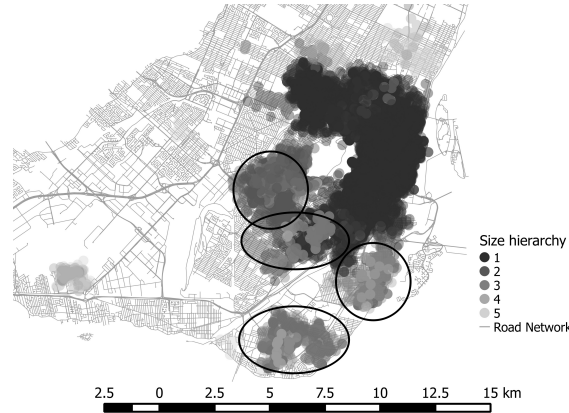


Figure 11: *Overlaying of the five biggest clusters per hour*

4.3 Joint Analysis of Bixi and Twitter

As stated earlier, Bixi pattern usage follows a typical commuter travel habit. There are two usage spikes for morning and afternoon commutes. For Twitter, the rate witnesses a peak in the evening and a smaller one during lunchtime. While both datasets evolve on different scales, a normalization by the average value enables a comparison on an equal basis [Figure 12]. This confirms the assumption that Bixi usage is higher than activity levels during both commute times. In the morning and in the afternoon, the tweet rate is steadily increasing while Bixi usage surges. During lunchtime, both patterns are similar. The graphs intersect at around 19:00, hinting at a shift of the majority towards doing activities rather than traveling. This state goes on until the next day, with a surge in the late evening at a rate three times higher than the average.

5 Conclusion and Future Direction

The use of publicly online available data helps to understand the dynamics of an urban center. Based solely on the Bixi data, the behavior of two typical users emerges – the commuters and the recreationals. The formers form the majority on weekdays while the latter are more present on weekends. In turn, it balances the network usage at around 16 000

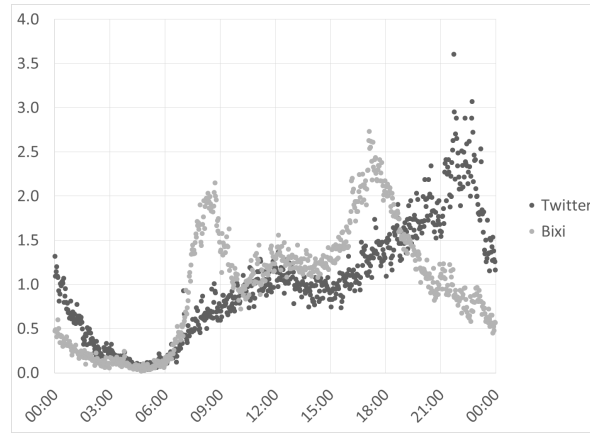


Figure 12: *Bixi network usage and Twitter activity aggregated by two-minutes then normalized by the average value of each dataset*

daily users on an average day. This number, however, is very dependent on daily temperatures.

The activity level of an urban center can be directly obtained from the rate of tweets that are shared in a certain area. It usually attains a two maximums, one around lunch time and another one late in the evening. The former being reached much later on weekends in accordance with a delayed awakening hour.

On a spatial basis, clustering all the tweets enables the detection of the core centers of activity for the city. Repeating the same process for each hour of the day and keeping the five biggest clusters creates the visualization of the pulse of the urban core. As time goes by, the clusters size and shape varies indicating a shift in activity rate and location.

The joint analysis of both datasets displays the complementarity of the behaviors they represent. When the Bixi usage is high, the activity level on Twitter is much less and vice-versa. While Bixi usage represents the trip behaviors, the tweet rate is an estimate of undergoing activities.

This exploratory analysis displayed the capacities of such datasets but further research can be done with them. In fact, now that a general overview of behaviors is made, a more thorough analysis of the most active individuals can be achieved. While these datasets are biased towards younger users, as more users join these services, the richness

of the information keeps increasing. This move to an online based massive data collection could help fill the gaps unfulfilled by traditional surveying methods.

6 Acknowledgment

We would like to gratefully acknowledge the contribution of Nicolas Saunier for his input on many aspects of this paper and Jean-Simon Bourdeau for his help with the coding and geoprocessing tools.

References

- Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer.
- Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.
- Colak, S., Alexander, L., Guatimosim Alvim, B., Mehndiretta, S. R., and Gonzalez, M. C. (2015). Analyzing cell phone location data for urban travel: Current methods, limitations and opportunities. In *Transportation Research Board 94th Annual Meeting*.
- DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4):41–56.
- Imani, A. F., Eluru, N., El-Geneidy, A. M., Rabbat, M., and Haq, U. (2014). How land-use and urban form impact bicycle flows: Evidence from the bicycle-sharing system (BIXI) in montreal.
- Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.
- Miranda-Moreno, L. F. and Nosal, T. (2011). Weather or not to cycle. *Transportation Research Record: Journal of the Transportation Research Board*, 2247(1):42–52.
- Morency, C., Trépanier, M., and Godefroy, F. (2011). Insights into Montreal’s bikesharing system. In *Transportation Research Board 90th Annual Meeting*, volume 6.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.