Personal Activity Centres and Geosocial Data Analysis: Combining big data with small data

Colin Robertson[1], Rob Feick[2] Martin Sykora[3], Ketan Shankardass[4], and Krystelle Shaughnessy[5]

[1]Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, Canada. crobertson@wlu.ca

[2]School of Planning, University of Waterloo, Waterloo, Canada. robert.feick@uwaterloo.ca

[3]Centre for Information Management, School of Business and Economics, Loughborough University, UK. M.D.Sykora@lboro.ac.uk

[4]Department of Health Sciences, Wilfrid Laurier University, Waterloo, Canada. kshankardass@wlu.ca

[5]Department of Psychology, University of Ottawa, Ottawa, Canada. Krystelle.Shaughnessy@uottawa.ca

Abstract

Understanding how people move and interact within urban settings has been greatly facilitated by the expansion of personal computing and mobile studies. Geosocial data derived from social media applications have the potential to both document how large segments of urban populations move about and use space, as well as how they interact with their environments. In this paper we examine spatial and temporal clustering of individuals' geosocial messages as a way to derive personal activity centres for a subset of Twitter users in the City of Toronto. We compare the two types of clustering, and for a subset of users, compare to actual self-reported activity centres. Our analysis reveals that home locations were detected within 500 m for up to 53 percent of users using simple spatial clustering methods based on a sample of 16 users. Work locations were detected within 500 m for 33 percent of users. Additionally, we find that the broader pattern of geosocial footprints indicated that 35 percent of users have only one activity centre, 30 percent have two activity centres, and 14 percent have three activity centres. Tweets about environment were more likely sent from locations other than work and home, and when not directed to another user. These findings indicate activity centres defined from Twitter do relate to general spatial activities, but the limited degree of spatial variability on an individual level limits the applications of geosocial footprints for more detailed analyses of movement patterns in the city.

Introduction

The proliferation of Internet and communications technologies (ICTs) and their associated information infrastructures are having transformative impacts on how people use, perceive, and co-develop urban spaces and places. Massive data streams are providing new ways to monitor, deliver, and analyze a variety of urban services (e.g. water and electricity consumption), community resources (e.g. transit and greenspace) and human activities (e.g. financial transactions and commuting flows) (Miller 2010). For researchers, new data streams with embedded geographic coordinates produce digital traces of individuals' interactions with each other and their surroundings and provide new opportunities to understand how urban communities operate and evolve, often at much finer spatial and temporal resolutions than previously possible (Batty et al. 2012).

Many of these digital traces result from user-generated and geosocial media which consist largely of photos, videos, text messages and tags, along with metadata related to locational references, time stamps and links to users' profiles. There are clear challenges to using these data, as data quality, coverage, locational accuracy and thematic relevance may be uneven, unknown, or limited (e.g., Robertson & Feick 2015). However, these data are often the only source of information available that can describe routine activities, patterns of movement, and observations of events and surroundings for large numbers of people (Goodchild 2007; Poorthuis et al. 2016). In this light, it is not surprising that a sizeable body of research has coalesced around using geosocial media from Twitter, Flickr, and Foursquare, for example, to gain new insights on topics as varied as fine-grained mobility patterns, place-sensing, vernacular geographies, and public sentiment, among others (e.g., Hollenstein & Purves 2013; Crampton et al. 2013; Mitchell et al. 2013).

In this paper, we explore the use of geosocial data for analysis of personal activity of individuals through the development of individual spatial and temporal clusters of granular geosocial media traces. An individual-based approach to spatial analysis of geosocial data is in contrast to more commonly used analyses of spatial aggregate patterns of social media activity. Aggregate approaches can highlight areas that display comparatively high or low levels of personal and work-related social communication (e.g. business, tourism and entertainment districts of large cities) as illustrated in Figure 1 below). However, aggregate approaches can also obscure our understanding of how individuals' use of urban space varies and contributes to overall patterns. As well, individual-level patterns can be aggregated to examine broader-scale patterns more common to 'big data' analyses. The goal of this work is to examine the use of spatial and temporal clusterings for exploring place-use within urban complexes and identifying locations that are of personal or functional significance to individuals. To demonstrate the possible value of this approach, we apply it to operationalize the concepts of home

(primary), work (secondary) and "third" places as a way of delineating locations of social and functional importance to individuals and groups and untangling these spaces from global patterns (Hickman 2013; Oldenburg & Brissett 1982; Soukup 2006). We use the term "activity centre" to mean the spatial expression of an individual's most important locations that they interact with on a regular basis (Golledge & Stimson 1997).

Given the increasingly varied work-life arrangements of urban populations as they respond to, and spur, changing economic, social and technological conditions, traditional place- and time-specific divisions between home, work and leisure time become more heterogeneous. Growing numbers of people are engaged in telecommuting, are "always connected" to work via mobile devices, and augment face-to-face socialization with digital alternatives (e.g. Facebook, online gaming, etc.) (Steinkuehler & Williams 2006; Sykora et al. 2015). Such changes have implications for planners and researchers alike. For example, there are public health implications of shifting activity centres in relation to exposure to environmental hazards as well as opportunities for interventions. Geosocial data offer possibilities to examine these developments in ways that are not possible with traditional data sources, such as censuses and surveys that are spatially and temporally coarse.

The use of user-generated and geosocial media in the social sciences has not been without criticism and challenges. Common to other forms of "big" geodata, early analyses often conflated data set size and frequency with objective truth and assumed that simple mapping of thousands or even millions of geosocial data points would shed light on broader social and urban processes (Crampton et al. 2013). As Kitchin (2014) notes, many forms of big data are by-products or "the exhaust" of specific activities. In contrast, small data are assembled based on carefully designed sample frames and variable selections. This has raised important questions related to data quality in multi-authored data sets, how the availability of massive data streams influences research foci, and the role of theory in analysis (Miller & Goodchild 2015).

More balanced and critical approaches to data-driven research using geosocial media have emerged recently in response to these criticisms. The representativeness of geosocial media is now recognized as being highly variable since its use and creation is dominated by relatively few advantaged groups (Haklay 2010; Miller & Goodchild 2015) and is concentrated spatially in core urban areas and places of widespread popularity (Li, Goodchild & Xu 2013).

A growing suite of papers describe activity centre analysis from geolocated Tweets. Huang and Wong (2016) used the DBSCAN method to generate spatial clusters for users, and then inferred activity zone types from urban land use data. Huang (2014) took a similar approach using DBSCAN to derive spatial clusters and then auxiliary data and metadata to infer activity space details. One of the challenges of relying on social media data to derive functional activity spaces of individuals is the huge uncertainty in the specificity of detection and how that uncertainty is distributed geographically and by demographics. In this study, we con-

front this challenge by adopting a bifurcated approach to geosocial activity centres. Firstly, we explore the use of DBSCAN to generate both spatial and temporal clusters for individuals at the population (e.g., big data) scale. Secondly, we examine a small sample of individuals who were recruited directly to provide information to characterize the relationships between our cluster-based methods and the true locations of participants' home, work, and other locations. Our specific research objectives in this paper are to: 1) provide a comparative analysis of two clustering approaches to generate personal activity centres (PACs) from geosocial data (big data), 2) compare the outputs of these algorithms to personally defined activity centres for a small subset of user-reported data (small data), and 3) examine whether participants' activity on social media are directly related to their surroundings at the time. We see this as a first step towards developing a robust methodology with known providence that allows individuals' routine use of urban space to be examined independently. As well, we aim to provide some degree of validation and contextual enrichment of 'big data' approaches by integrating reports from individual participants captured within such data streams.

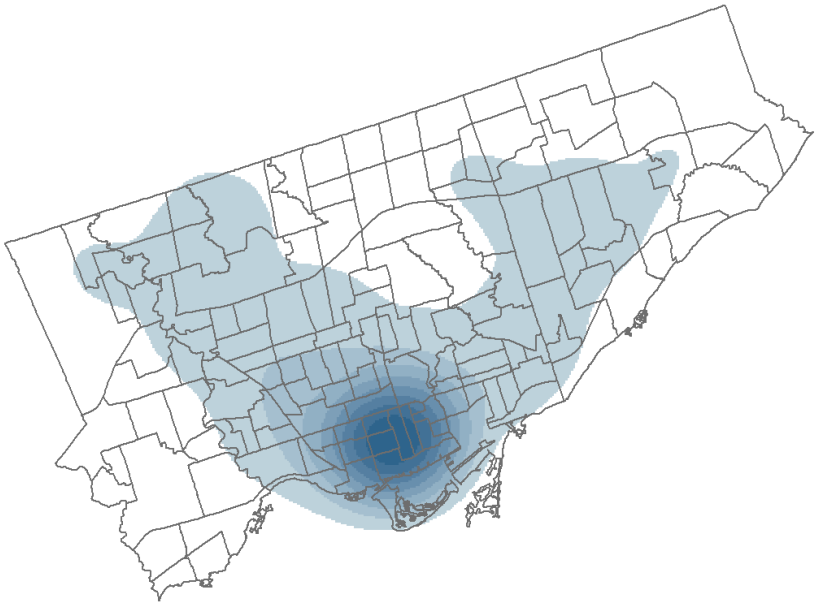

Fig. 1 Kernel density heatmap for 2.6 million geotagged Tweets in Toronto, Canada

2. Methods

2.1 Data Sources

Data were obtained from the public Twitter Streaming API during the year of September 2013 to August 2014 within the boundaries of the city of Toronto, Canada. There were a total of 2.6 million geocoded messages and more than 99,000 unique users in the dataset after duplicates were removed. As we were interested in highly active users and given the skewed distributions of user-generated content in general, we constrained our analysis to users with a minimum of twenty-five messages over the year. Users with extremely high numbers of tweets (more than 2500) were also removed from the database since inspection revealed these users represented automated accounts ("bots") and/or businesses. This reduced our database to a final dataset of 2 million messages distributed across just over 16,000 unique users.

A secondary data source was individual survey responses from a subset of active Twitter users who resided in the study area and also had records within the larger database of tweets. We recruited participants through direct messaging and public posts on Twitter and other social media networks as part of a larger study investigating geolocated sentiment analysis and urban form. For the analysis reported here, survey data for a total of 16 participants were used. Details of the methodology for these data are reported in Sykora et al. (2015). In short, participants filled out a short entry survey upon initiation into the project that provided a baseline demographic profile including home and work locations. Over a series of weeks, each participant received short follow-up surveys that were triggered by their social media activity (e.g., posting a message to Twitter). Variables collected in these surveys included their location at the time of Tweeting (home, work, other) and the activity they were engaged in at the time of Tweeting (working, relaxing, etc.). The participant data cover a period from August 2015 to November 2016, slightly after the larger database described above. The temporal displacement between the small and big datasets used here is due to technical problems with data storage which limited our ability to collect Tweets during the concurrent period. However, we consider the impacts of this misalignment to be marginal as we collected length-of-residence data for participants in the survey, which indicated the majority had not changed residence in the period since the collection of Tweets from the API.

2.2 Analysis Methods

Recognizing some of the issues inherent in aggregate spatial analysis of geolocated social media data described above, we aimed to detect spatial areas of significance to individual users in the dataset. Two simple methods were used to derive clusters from individual social media users' geosocial data – clustering by space, and clustering by time. Our intuition is that user-activity may be similar in space (e.g., Tweeting from home or work) or similar in time (e.g., Tweeting at lunch

hour from the same or from different locations), and these may be used as indicators of how people are behaving. With granular activity detection, scaled up to the population level, we may be able to investigate interesting questions about the use of space, and link expressions and meaning derived from social media content to their geographic context in a meaningful way for individuals.

2.2.1 Personal Activity Centres and Big Data Analysis of Georeferenced Tweets

The clustering algorithm used was the density-based clustering of applications with noise (DBSCAN) method (Ester et al. 1996), one of the more widely used methods for simple clustering of points. The purpose of DBSCAN is to find spatial clusters of high density and to identify points in low-density regions as outliers (or noise). The density in DBSCAN is defined by two key parameters, the neighbourhood size and the minimum number of points belonging to a cluster. Together, these parameters define the types of clusters that will be found by the algorithm. To find irregularly shaped clusters, the algorithm distinguishes between core points and border points through the concept of whether points are density-reachable, such that point p in a spatial pattern is density-reachable from point q if there is a chain of points connecting them with density above the threshold determined by the two parameters.

We operationalized the DBSCAN algorithm to find spatial clusters with a minimum number of five points, and a maximum neighbourhood size of 100 m. We set these parameters after exploratory analysis and consideration for GPS error and local mobility within the same functional place within an urban setting (e.g. movement within a single property).

For temporal clustering we set the neighbourhood size to 30 minutes, and again the minimum number of points to five points per cluster. Connected segments of Twitter activity within 30-minute intervals would then be connected into clusters of minimum density. To capture clusters occurring over midnight, we transformed the hour of the Tweet to two dimensions (cosine and sine transforms) and used these as input into the DBSCAN algorithm. Note that because clusters are defined exclusively at the individual level, clusters can overlap spatially across users. Varying of parameters and re-running these analysis did not significantly change our overall results.

Derived clusters were ranked for each user based on spatial density. For each user, we took the set of points belonging to cluster K for user i, and computed the maximum distance separating the points. The density for user i and cluster k was computed as:

$$d_{i,k} = \frac{N_{i,k}}{Max\|P_n - P_m\| \forall P \in k}$$

and ranked such that,

$$PAC_{i,1} = P_i \in k$$

where,

$$d_{i,k} > d_{i,k+1} > d_{i,k+n}$$

Thus the set PAC_{i1} is the highest density clustering of points for user i, followed by PAC_{i2} and so on up to the number of clusters for user i. This ranking of individual-level clusters therefore maps onto our sociological-derived notions of space-use based on function: home, work, and other. Here, we focused mostly on the analysis of the first two orders (i.e., highest density locations). We hypothesize here that the densities will follow from highest density Tweeting at home, second highest at work, and third and higher order rankings at third places.

Note that for both spatial and temporal clusters we use the spatial distance in the denominator, as we are ultimately interested in functional areas of the city at the individual level. To distinguish between cluster types, spatial PACs are referred to here as PAC-Ss, while temporal clusters are noted as PAC-Ts. Correspondence between rank orders of PAC-Ss across the dataset therefore is seen to indicate equivalent types of spatial areas at the individual level and similarly for PAC-Ts. We use this framework to investigate spatial patterns of activity centres, and to compare to the true functional areas for the smaller subset of study participants.

Spatial and temporal clustering methods were compared using the variance of information (VI) distance for comparing clustering methods (Meila 2007). The VI statistic is based on the difference in entropy introduced by the different clustering methods. The entropy of a clustering C of k clusters,is defined as:

$$H(C) = -\sum_{k=1}^{k} P(k) log P(k)$$

where $P(K)$ is the proportion of points in cluster k. Given two clustering methods, we can compute the joint-entropy, otherwise known as the mutual information of clustering C and clustering C' as:

$$I(C, C') = \sum_{k=1}^{k} \sum_{k'=1}^{k'} P(k, k') \, log \frac{P(k, k')}{P(k)P'(k')}$$

which is the amount of information common to the two types of clusters. We can therefore measure the similarity between two clustering methods using the following statistic as defined by Meila (2007);

$$VI(C, C') = [H(C) - I(C, C')] + [H(C') - I(C, C')]$$

The VI is combined entropy, minus the mutual information (i.e., the entropy common to both clustering methods). As such the VI is a metric that measures how dissimilar cluster sets are, and has a value of zero when the two cluster sets are identical. We will use the VI to measure the degree to which spatial and temporal clusterings of social media activity in Toronto are similar, and how similarities vary across space. As all analysis was done at an individual level, cluster comparisons were made for individual users, yielding a distribution of cluster distances for the two clustering methods across the approximately 16,000 users represented in the dataset. We examined the magnitude, distribution, and spatial patterns of cluster distances.

2.2.2 Comparison of clustering methods and validation data

For the subset of 16 users who participated in the validation study, we had both spatial and temporal clusters (PACs) calculated from the larger Twitter database (n=3738), as well as data obtained from their participation in the wider study of social media use in Toronto (n=125). Using our PAC ordering framework described earlier, we compared home, work, and other locations as reported by participants, to PAC orders as derived from clusters in space and time.

Deriving clusters for users from the Twitter database allowed us to map hypothesized functional areas for each individual in the wider population of users. We examined the proportion of cluster types in relation to the total number of Tweets measured in each neighbourhood of Toronto.

2.2.3 Examining Tweets pertaining to the environment

Participants who were sent a survey in response to their Twitter activity were asked to report if their Tweet related to their immediate environment. The purpose of this question was in order to better understand the explicit linkage between the content shared on Twitter and their surroundings at the time of Tweeting. Social media have been postulated as a potential tool for researching person-place linkages, especially in the context of health and epidemiological research. To explore this idea further, we estimated the probability that a submitted Tweet was about the environment in relation to the place it was sent from, whether it was a message directed to someone, whether the user used Twitter for professional or personal purposes, and the age of the user. To explore these relationships, a logistic regression model was constructed with a random effect for users. In this subset of the data, there were 59 unique users and 772 messages.

Results

A total of 2,090,637 messages by 16,793 unique users were analyzed for spatial and temporal clustering. In general, the number of clusters per user was higher for temporal clustering than for spatial clustering (Table 1), and temporal clusters had higher densities and higher numbers of Tweets. In terms of distribution, only

1.9% of users did not have temporal clusters, which signifies no dominant time periods in which these users posted messages. Of the remaining users, 16.9% had only one temporal cluster, 23.7% had two and 20.3% had three as their highest order temporal cluster.

Table 1 Descriptive statistics of spatial and temporal clusters of Twitter messages

Metric	Spatial	Temporal
Number of clusters total	42,606	52,351
Avg. number of points per cluster	35.80	30.42
Mean density	0.19	1.41
Mean number of clusters per user	2.54	3.23

In contrast to the PAC-T findings, only 2.5% of users had zero spatial clusters. Some 35.1% had a maximum of one spatial cluster, 29.8% had only two, and 14.1% had three PAC-Ss. Overall, 18.5% and 37.0% of users had four or more spatial and temporal clusters respectively. In general, users' clusters tended to be found at only a handful of discrete locations, whether defined spatially or temporally. This finding is in line with research in the urban sociological, geography and planning fields that has found consistency in urban space use due to stability in home, work and often social activity spaces (Oldenburg & Brissett, 1982). Note that the spatial footprint of temporal clusters could vary significantly, as only time was used as a criterion for clustering, although density ranking was based on spatial densities. The distributions for the maximum number of clusters are shown in Figure 2.

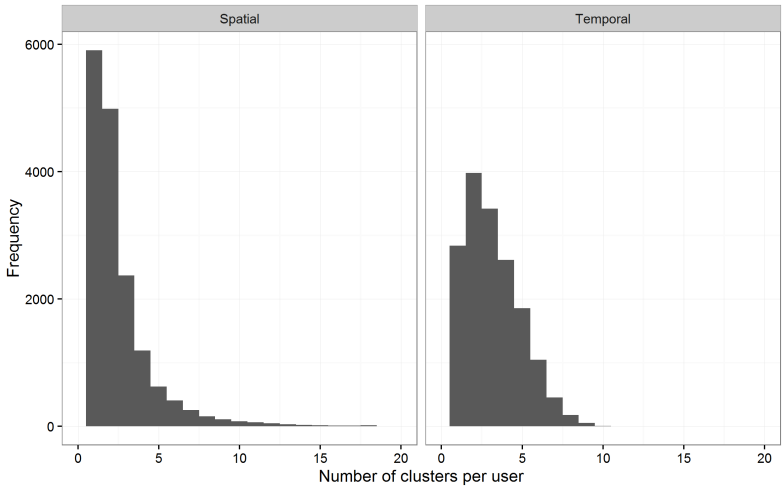

Fig. 2 Number of spatial and temporal clusters per user

The distribution of VI scores ranged from 0 to 5.14, with a mean of 2.18 and standard deviation of 0.78. Randomly sampling from the upper and lower tails of the VI distribution, we present the groupings of two similar and dissimilar PAC clusterings in Figure 3. In this figure, red points belong to first order PAC-Ss and PAC-Ts, blue points signify second order PACs, while points classed as third or higher order PACs are green. Grey points do not belong to a PAC cluster. Different behaviours are captured temporally compared to spatially. In the case of User A, the locations of the first (red) and second order (blue) clusters are reversed when their PACs are defined based on space (Fig. 3a) as opposed to time, however the points that are members of the clusters are very similar. For User B, where spatial clustering reveals two distinct clusters for order one and order two, temporal clustering captures what is likely a commuting pattern as part of the secondary cluster. In general, the more clusters that were discovered for a user, the lower their cluster agreement scores. However, by constraining comparisons between the two lower order clusters and investigating individual users, the differences and meaning behind the different clusters becomes more apparent.

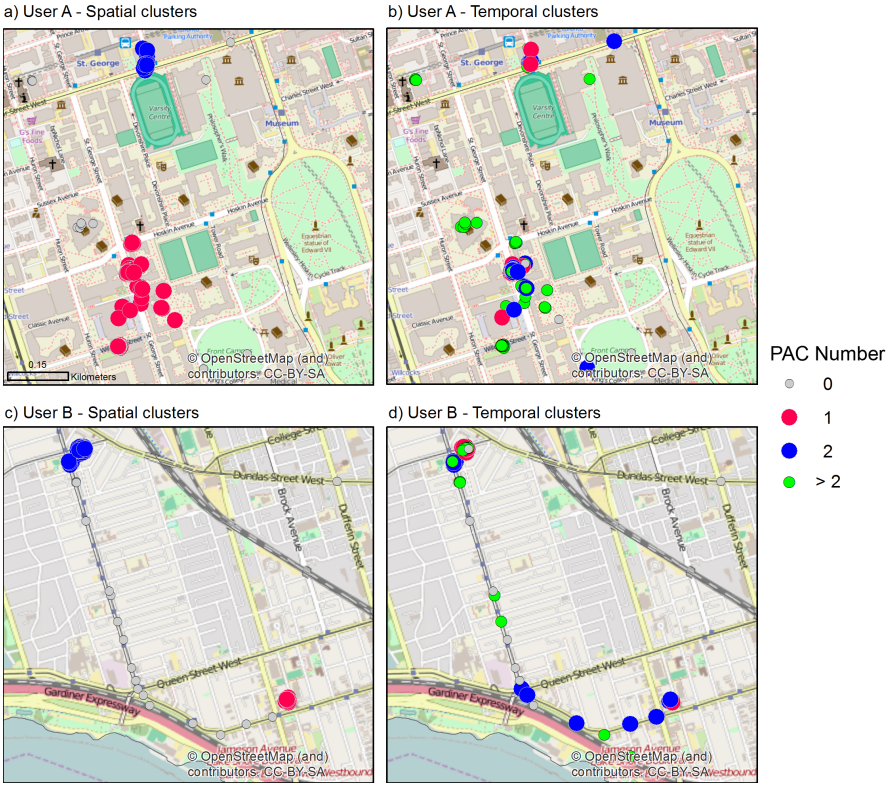

Fig. 3 Spatial and temporal clusters of Twitter messages from DBSCAN; similar (a-b) and dissimliar (c-d) clusterings.

In Figure 4, we see spatial and temporal clusters in relation to users' true home and work locations as reported by the 16 study participants. In this way, we see how the spatial expression of the activity centres differs from the functionally important areas they identified. For User C's spatial clusters (Figure 4a), we see that the tweets around their home location (red square) are spread out over a range of about 1 km in their local neighbourhood. Conversely, their highest density spatial PAC was at a distant location. Exploratory analyses of PAC-S-1 and PAC-S-2 for User C revealed these locations to be both houses in residential areas.

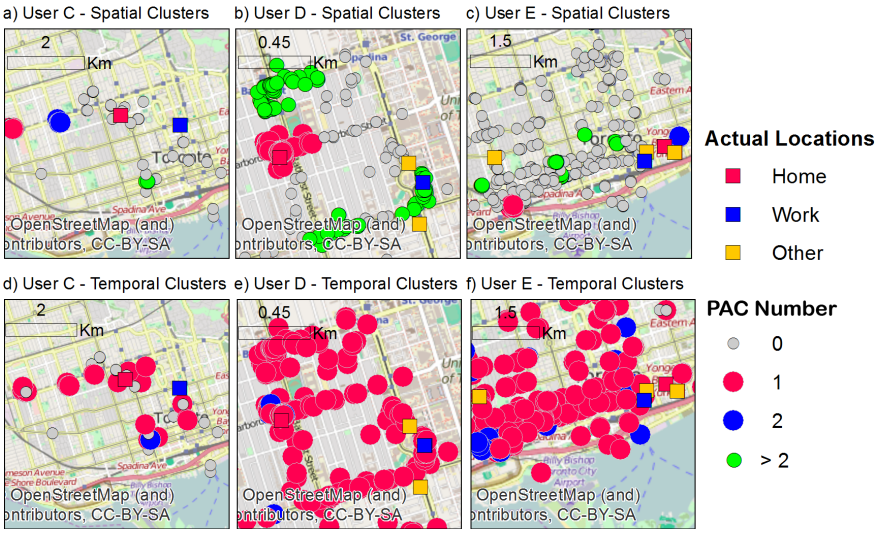

Fig. 4 Spatial (**a**-**c**) and temporal (**d**-**f**) clusters for 3 users in relation to their reported home, work and other locations.

With temporally defined clusters (Figure 4d), User C's PAC-T-1 (red) is between 4 and 6 pm, and PAC-T-2 is later in the evening around 8-10 pm (blue). The varied spatial pattern associated with these temporal clusters indicates that while this user regularly tweets at these times, they do so from different locations. For User D, the spatial cluster (Figure 4b) identifies their home location well (red), while no tweets were recorded or predicted at work. Their work location is associated with a number of tweets (green), but these did not constitute their PAC-S-2. The pattern shows several areas of high activity as well as some commuting patterns, over a very small area (~1km). For User D's temporal clusters (Figure 4e), the majority of tweets were in PAC-T-1 (red) which was a cluster of quite regular tweeting activity that spans over the entire working day from 8am until around midnight. User D had the most tweets of the users we investigated. User E had a lower density overall spatial pattern (Figure 4c) and a temporal pattern of tweeting (Figure 4e) similar to user D.

In terms of spatial clusters, for user E (Figure 4c), PAC-S-1 was located near an urban park in the west side of the city, while their home location was located 4km east in the central district. PAC-S-2 was located 500m west of their home, while their work location was 500m east of their home. This is within the range of locational accuracy provided by the postal code reference used to locate home and work locations. Temporal clustering for User E (Figure 4f) found a PAC-T-1 to be morning, between 7 and 9 am, and PAC-T-2 was evening, between 7 and 10 pm, both of which were highly dispersed at the neighbourhood scale (~3km).

The distribution of spatial PACs computed over the full dataset differed significantly from the pattern in Figure 1. Figure 5 presents the proportional mapping of Tweets by neighborhood and the proportion of PAC-S-1 tweet clusters relative to the total number of tweets. Figure 5a shows a pattern similar to Figure 1 with a familiar high concentration of messages in the downtown central core, and incrementally fewer as one moves out to suburban and non-core parts of the city. Alternatively, mapping PAC-S-1 tweets as a proportion of the total tweets in a neighbourhood shows the inverse pattern where the highest values are in outlying areas. This shows that in aggregate, the spatial PAC-1 clustering captures an intuitive demarcation of predominantly working and entertainment areas of the city and predominantly residential areas.

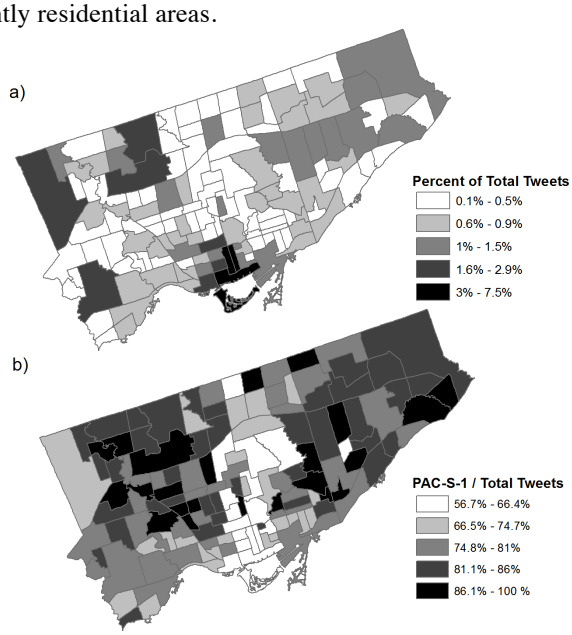

Fig. 5 Spatial distribution of: **a**) total Tweet variation, and **b**) spatial cluster order 1 (PAC-S-1), aggregated by neighbourhoods

The relationships of PAC-S-1 to home and work locations for all study participants are provided in Figure 6. The figure shows that, in general, spatial PAC-1

clusters (PAC-S-1) are very close to home locations. Over 50% of PAC-S-1s are within 500 m of the study participants' geocoded home location. For PAC-S-2, the median distance to work locations is 1.4 km, with about 20% within 500 m and 33% within 1 km. PAC-S-3 clusters show the largest median distances to both home and work locations. For temporal clustering, the PAC-T-1 was closest to home locations, with a median distance of 1.4 km, and a median distance of 2.2 km to work locations.

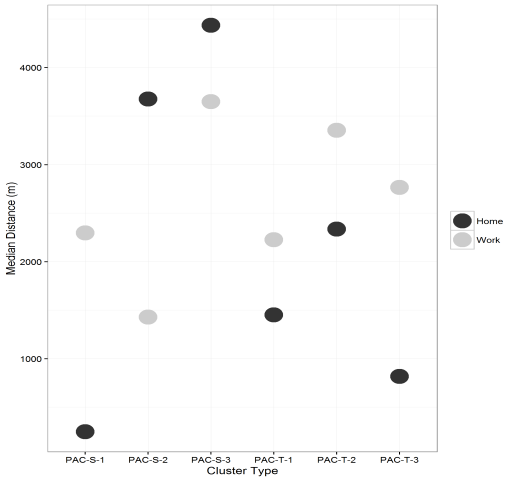

Fig. 6 Median distance between cluster types, home and work locations for all participants.

For the 59 users and 772 messages investigated here, 28.0% were reported to be about users' surroundings at the time of Tweeting. By users, the 1st and 3rd quartiles of this proportion were 7.0% and 42.9% respectively, indicating a high degree of individual variation. The results of the regression model that examined Tweets that were about users' environments are reported in Table 2. From this table we see that two variables are significant predictors of the log-odds of an environmental Tweet, whether the Tweet was directed at another user (negative association) and whether the Tweet was sent from a third place (positive association). The variance of the random effects term was 0.789 and the mixed-effects model reported here performed better than a regular logistic model based on AIC (however same effects were present plus a slight positive effect for age).

Table 2 Logistic Regression models results

Model Term	Estimate	Standard Error	P-Value
Intercept	-1.92	0.607	0.001
Use for work	0.175	0.502	0.727
Tweet directed	-1.059	0.234	<0.001
Age of User	0.022	0.017	0.195
Place-work	-0.071	0.233	0.760
Place-other	1.583	0.234	<0.001

We visualize these results in Figure 7 which shows the probability of a Tweet being about the environment as reported by users, which shows some discriminative ability of the factors reported in Table 2.

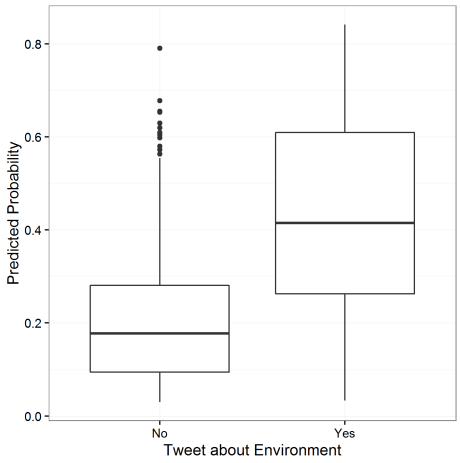

Fig. 7 Model probabilities of environment-related Tweets in participant data.

Discussion

Our analysis provides a comparative clustering of geosocial footprints for both a large database of geolocated tweets in the City of Toronto and a subset of study participants selected for more in-depth analysis. The spatial and temporal clusterings provide reasonable estimates of personal activity centres for many individuals. From the distribution of cluster orders, we see that on average Twitter users are active from only a handful of locations, usually within close proximity to their home and work locations (e.g. Figure 3 a and b). For some users, there is also evi-

dence of commuting behaviours in their geosocial footprints (Figure 3d) and temporal consistency in messaging behaviour.

Spatial clustering of individual tweet locations provided a more realistic estimate of space use at the individual level. We discovered clusters that aligned with our hypothesized activity space categories of home and work locations as evidenced in Figure 6. The hypothesis that personal activity centres derived from geosocial data can estimate real functional areas seem to be supported by the analysis. Our focus on using clustering to delimit the spatial and temporal extents of ranked PACs complements other approaches that derive individuals' activity spaces through social network connections or message content. Study participants' reported home, work and other activity locations provided a means to examine how well the calculated PACs corresponded to what may be considered true functional areas for individuals. In this way, we sought to couple the analytical advantages of big geodata, including extensive sampling of populations and unobtrusive data collection methods, with more structured small data that serve as an indicator of the validity of the clusterings to delimit meaningful locations at the individual level.

There are important limitations to this work that we are seeking to address in ongoing research. Like many types of big data, geosocial media is partial in nature. For instance, social media users are not representative of a city's sociodemographic composition, these data can only be created under certain conditions (e.g. not while driving), and only a small fraction of the data are encoded with GPS coordinates (Morstatter et al. 2013). In our study, these realities were accentuated by the small set of active participants that produced the validation data and, in recognition of this limitation, ongoing work is focused on expanding the participant pool within Toronto and selected other cities in North America. As well, even though our analyses focused on individuals, we aimed to understand both how people use social media in the city (generalizing to the broader population of Twitter users), and how individuals' digital expressions are impacted by their environment. In general however, representativeness issues related to analysis of social media may be alleviated as these technologies become more utilized and accessible by more of the population (Boyd 2014).

There are several interesting implications of this work that merit further study. First, the approach demonstrated for deriving spatial and temporal footprints from geosocial data streams offers new information to understand individuals' use of urban space. This could be of particular value for examining the dynamics of space use in response to evolving work-life patterns, changes to the urban fabric (e.g. promotion of mixed land uses), or seasonal conditions (e.g. snow storms, heat waves). Second, an individual's PACs can be enriched with complementary data extracted from their social media user profiles (e.g. interests, profession, demographic characteristics) or from analysis of the content of their messages (e.g. sentiment analysis). This could help researchers to understand some of the reasons that underlie a person's routine activity patterns. In particular, combining detailed analysis of message content with PACs may hold potential for public health plan-

ning and evaluation in cities, examining spatial health and equity issues related to congestion, pollution, stress, and fear of crime. Finally, there is a clear need to consider methods to protect personal privacy given the growing sources of geospatial traces and new methods to rapidly derive associated outputs. This could include, for example, limited random displacement of data points prior to developing PACs or post-process randomization of data points within the convex hull of a PAC.

Figure 5 demonstrated the stark contrast in spatial pattern when mapping aggregate tweet density versus mapping aggregate patterns of PACs. There is potential to provide more nuanced spatial analysis of geosocial data by providing functional meaning to otherwise disaggregate patterns. For example, in an urban analytics setting, we could constrain an analysis of place-based issues to only those messages located within the vicinity of individuals' PAC-1s to help filter out noise in the signal and provide meaningful spatial context to other forms of public engagement tools for urban managers and planners.

Finally, model results indicate that the 'environmental content' in tweets may be limited, and importantly, may vary systemically. Significant associations with the place from which the message was sent, as well as whether it was directed could be used as filtering criterion when doing environmental analyses of Twitter data. Deeper understanding of these forms of variability in the content and intention in geosocial data is needed before such data can be used to their greater potential for understanding human activities and interactions with the environment.

Our analysis provided comparative clustering using existing algorithms to derive personal activity centres from geosocial media data. We demonstrated the effectiveness in locating home and work locations from simple spatial clustering for a majority of users investigated. Ongoing studies and validation data will provide further insight into the preliminary patterns investigated here.

ACKNOWLEDGMENTS

The authors gratefully acknowledge our study participants as well as the Social Sciences and Humanities Research Council of Canada for funding this research.

REFERENCES

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., et al. (2012). Smart cities of the future. *The European Physical Journal Special Topics*, 214(1), 481-518..

Boyd, D. (2014). *It's Complicated: The Social Lives of Networked Teens*. New Haven, CT, USA: Yale University Press.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W. & Zook, M. (2013). Beyond the Geotag: Situating "big Data" and Leveraging the Potential of the Geoweb. *Cartography and Geographic Information Science* 40 (2): 130–39.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Golledge, R.G., and Stimson, R.J. (1997). *Spatial behavior: A geographic perspective*. Guilford Press.

Goodchild, M.F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 211–221.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design* 37, 682–703

Hickman, P. (2013). "Third places" and social interaction in deprived neighbourhoods in Great Britain. *Journal of Housing and the Built Environment*, 28(2), 221-236.

Hollenstein, L., & Purves, R. (2013). "Exploring Place through User-Generated Content: Using Flickr Tags to Describe City Cores." *Journal of Spatial Information Science*, no. 1 (January): 21–48.

Huang, Q., & Wong, D.W.S. (2016). "Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?" *International Journal of Geographical Information Science*, 30, 1873–1898.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. GeoJournal, 79(1), 1–14.

Lee, J.H., Davis, A.W., & Goulias, K.G. (2016). Activity Space Estimation with Longitudinal Observations of Social Media Data, in Paper submitted for presentation at the 95th Annual Meeting of the Transportation Research Board. Washington, D.C., January 10-14, 2016.

Li, L., Goodchild, M.F., Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40, 61–77.

Meilă, M., (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis*, *98*(5), pp.873-895.

Miller, H.J. (2010). The Data Avalanche Is Here. Shouldn't We Be Digging? *Journal of Regional Science*, 50(1), 181–201.

Miller, H.J. & Goodchild, M.F. (2015). Data-driven geography. *GeoJournal*,*80*(4), pp.449-461.

Mitchell, L., Frank, M.R., Harris, K.D., Dodds, P.S. and Danforth, C.M., (2013). The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS One*, 8(5), p.e64417.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*.

Oldenburg, R. & Brissett, D. (1982). The third place. *Qualitative sociology*, 5(4), 265-284.

Poorthuis, A., Zook, M., Shelton, T., Graham, M, and Stephens, M. (2016). Using Geotagged Digital Social Data in Geographic Research. In *Key Methods in Geography*. eds. Clifford, N., French, S., Cope, M., and Gillespie, T. London: Sage. 248-269.

Robertson, C. and Feick, R. (2015). Bumps and bruises in the digital skins of cities: unevenly distributed user-generated content across US urban areas. *Cartography and Geographic Information Science*, 1-18.

Soukup, C. (2006). Computer-mediated communication as a virtual third place: building Oldenburg's great good places on the world wide web. *New Media &Society*, 8(3), 421-440

Steinkuehler, C.A. and Williams, D., (2006). Where everybody knows your (screen) name: Online games as "third places". *Journal of Computer-Mediated Communication*, *11*(4), 885-909.

Sykora, M.D., Robertson, C., Shankardass, K., Feick, R., Shaughnessy, K., Coates, B., Lawrence, H. & Jackson, T. (2015). Stresscapes: validating linkages between place and stress expression on social media. Published by CEUR Workshop Proceedings.

Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic books.