The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Characterization of Users by Using Hourly and Daily Spatio-Temporal Patterns Extracted from GPS Trajectories

Marcello Tomasini,[‡] Carmelo Bastos-Filho,[†] Ronaldo Menezes[‡]
[‡]BioComplex Lab, School of Computing, Florida Institute of Technology, Melbourne, Florida, USA
[†]Computer Engineering Department, University of Pernambuco, Recife, Brazil
mtomasini@my.fit.edu, carmelofilho@ieee.org, rmenezes@cs.fit.edu

Abstract

The research on human mobility has mostly focused on the scaling laws of movement. While important to construct better models, such relationships hardly tell anything about individuals. Furthermore, there is a growing evidence that the observed scaling relationships are actually the result of the convolution of multiple distributions originated by the heterogeneity of the users. Hence, it is important to build tools that are able to disentangle aggregated patterns and provide information at a finer granularity. In this work, we introduce a framework for the analysis of users' GPS trajectories. We show that hourly and daily features extracted from GPS trajectories can be used for an unsupervised characterization of the users. We found the existence of distinct classes of users who exhibit substantially different spatio-temporal patterns.

Introduction

The study of human mobility has experienced a massive growth with the recent availability of cell phone Call Detail Records (CDRs) and large localized datasets from social networking sites. Large mobile phone Call Detail Records (CDRs) data were instrumental in identifying and modeling fundamental patterns of human mobility. For instance: the regular schedules and travel displacement (Gonzalez, Hidalgo, and Barabasi 2008), the tendency of people to go back to previously-visited locations (Song et al. 2010a), the high predictability of people movements (Song et al. 2010b), the tendency of people to go back to recently-visited locations (Barbosa et al. 2015), and the connection between mobility and individuals' social networks (Karamshuk et al. 2011; Toole et al. 2015). Furthermore, part of the human mobility research tried to classify people in groups based on the characteristics of their movement such as the characteristic displacement (Pappalardo et al. 2015) or the similarity of the user's trajectories (Xiao et al. 2014).

The data from Location Based Social Networks (LBSNs) have the advantage, compared to CDRs, to have contextual information associated with geographical positions. Scientists have been able to identify users' information such as relevant locations of individual users (Mamei, Colonna, and Galassi 2016) to give recommendations on future locations to visit based on location history (Zheng and Xie 2010; Yang et al. 2013). The trajectories augmented with information related to the activity performed or the purpose of movement allowed researchers to get better modeling of intra-urban mobility (Wu et al. 2014) and to identify urban mobility patterns and anomalies (Gabrielli et al. 2014). LBSNs also include data on the social network of the users and thus they allow the exploration of how social ties and human mobility are intertwined (Cheng et al. 2011; Scellato et al. 2011).

Yet, there are two gaps in the current literature. First, while identifying the fundamental scaling laws of human mobility is important to build better models, such relationships hardly tell anything about single individuals or group of individuals. Furthermore, there is a growing evidence that the observed scaling relationships are actually the result of the convolution of multiple distributions originated by the heterogeneity of the users (Gonzalez, Hidalgo, and Barabasi 2008; Xiao et al. 2014; Toole et al. 2015; Pappalardo et al. 2015). Second, most techniques use additional data from multiple sources in order to annotate trajectories, which inherently limits the applicability of such methods as those sources might not be available everywhere and/or with the same accuracy. In this work, we introduce a framework to disentangle aggregated mobility and reveal specific patterns through the use of hourly and daily spatio-temporal features (*popular-hours*, *popular-days*, *hourly* and *daily entropy*) extracted from users' trajectories collected from LBSNs and online social networking sites (*e.g.* Twitter). The methods and techniques presented do not rely on any additional data and can be consistently and reliably applied to any source of geo-located trajectories data.

Methods and Data

Data Preprocessing

GPS data do not tell the nature of the location visited by users. Therefore to appropriately recreate user's trajectories for each dataset, we extracted the Points of Interest (POIs) from the geo-located records by applying HDBSCAN clustering to the GPS coordinates of the data (Jenson et al.). Moreover, when analyzing users' trajectories we must carefully curate the data (see Figure 1). In fact, there are at least two sources of population-level heterogeneities in the data which can significantly impact results: the characteristic ac-

tivity of the specific data source (*e.g.* Twitter) and the users' mobility. We want to eliminate the effects of specific sources of data, hence we selected only those users with a level of activity which allows for a reliable analysis of the trajectory properties.

We removed users with anomalous movement speeds (over 80m/s) and anomalous number of records. Then we selected users who have generated records in at least 2 distinct locations (*e.g.* home and work), so that the trajectory entropy is non zero. We also impose an average activity of one record per day, so users have a comparable level of activity. Furthermore, we re-sampled the trajectories in 30-minutes intervals to mitigate the burstiness in human dynamics (Myers and Leskovec 2014). The re-sampling process also avoids artificially lowering the entropy and skewing the radius of gyration of the trajectory disproportionately towards those locations where the bursty behavior happened. We also did a basic interpolation by carrying over the location identifier of the location visited in the previous half hour if we did not have data for an immediately consequent interval. This step assumes it is very likely the user did not move to another location in such a short time span. Finally, we removed users who were present in more than one dataset, retaining only the trajectories from where the users appeared the first time; this step avoids relating trajectories that were collected temporally far apart.

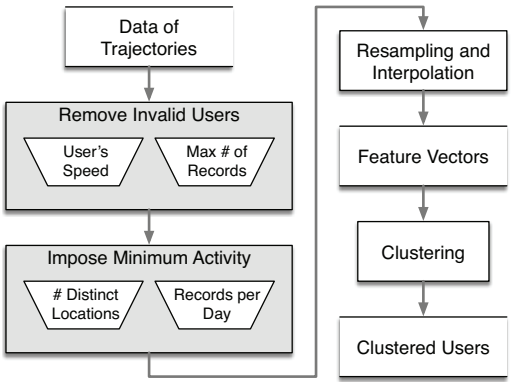

Figure 1: The data of the trajectories need to be processed to remove invalid users (*e.g.* robots) and select those trajectories with enough data to guarantee a reliable analysis. The re-sampling and interpolation step is also important to mitigate the bursty behavior of human activities. Once the datasets are curated, we can extract the feature vectors to cluster and characterize the users.

Datasets

We used 3 datasets of geo-located tweets collected over the Manhattan (NYC) area, henceforth called $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$ (Table 1). Each tweet represents a record of data with GPS coordinates and timestamp. $\mathcal{D}_1$ was collected between August 28 and September 29, 2014 and contains 108,341 users. $\mathcal{D}_2$ is comprised of 154,198 users collected between June 17 and November 4, 2016. $\mathcal{D}_3$ is comprised of 149,944 users collected between April 5 and September 11, 2017. After the preprocessing step, we ended up with 4,185 users in $\mathcal{D}_1$, 861 users in $\mathcal{D}_2$, and 933 users in $\mathcal{D}_3$ who met the aforementioned criteria. Hence we conducted the users analysis on a total of 5,979 unique users.

Table 1: Mobility datasets used for the user characterization.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$
# of users	108,341	152,671	147,942
# of tweets	1,022,286	1,068,939	1,179,389
# clustered users	4,185	861	933
Date range	Aug 28–Sep 29	Jun 17–Nov 4	Apr 5–Sep 11
Year	2014	2016	2017

Feature Extraction and User Clustering

The possible classes of users is not known *a priori*, hence it is not possible to assign labels to the users for a supervised learning process. Instead, we rely on similarity metrics to discover and group users in similar classes. We applied the spectral clustering algorithm on the users' feature vectors extracted from the GPS trajectories. We then used the Bayesian Information Criteria (BIC) and silhouette score to identify the optimal number of clusters. Furthermore, we analyze only those clusters which contain at least 5% of the users in order to be able to draw reliable conclusions as a small sample would be too sensitive to noise. One aspect that needs to be defined is how to represent the users through some descriptors (*i.e.* the feature vectors) which are able to discriminate different types of users. We decided to focus our attention on two types of features: popular-times and popular-days histograms and hourly and daily entropies.

The *popular-times* and *popular-days* histograms are the amount of records generated by the user when her trajectory is aggregated by hour of the day and by day of the week over the observation period. The two histograms are concatenated and standardized to form a 31-dimensional $(24 + 7)$ feature vector for each user. We applied a one-dimensional Gaussian filter over the histograms to smooth the distribution and fill in missing data and normalized by applying standardization. The analysis of the spectral clustering of the popular-times and popular-days feature vectors returned $k = 8$ clusters with at least 5% of the users represented. The assumption is that people have strong habits (Neal, Wood, and Quinn 2006) and such behavior property implies that the effect accumulates over time reinforcing the hours and days at which a person is mostly active. For example, someone who is very active on the weekends and late hours might represent a younger person profile when compared with someone who follows a 9-to-5 routine; consequently, even the type of locations and the movement patterns should differ substantially.

The *hourly entropy* and the *daily entropy* (Figure 2) are similar to the popular-times and popular-days histograms where the quantity computed on the trajectory aggregated per hour or per day is the Lempel-Ziv estimate S^{est} of the entropy of the sequence of locations visited by the user (Song et al. 2010b)

$$S^{\text{est}} = \left(\frac{1}{N} \sum_i \Lambda_i \right)^{-1} \ln N, \tag{1}$$

where N is the number of distinct locations visited by a user and Λ_i is the length of the shortest substring starting at position i which does not appear in positions 1 to $i-1$. Unlike Shannon entropy, S^{est} does not depend on the trajectory length (as long as it is long enough to provide a good estimate of the entropy), but only on the sequence itself, and therefore it is possible to compare sequences with a different length (Amigó et al. 2004; Zhang et al. 2009). The 31-dimensional feature vectors are then normalized and clustered. The analysis of spectral clustering of the hourly and daily entropy feature vectors returned $k = 4$ clusters with at least a 5% users' representation. This type of features is es-

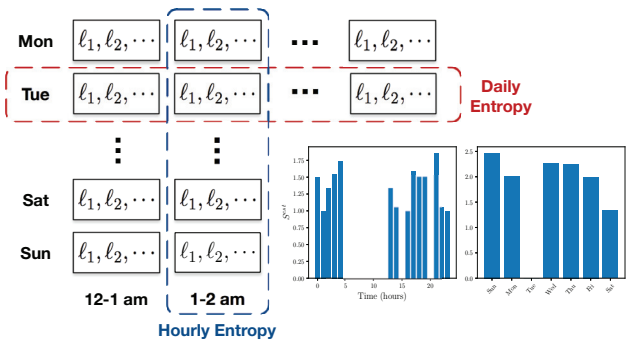

Figure 2: The hourly entropy is computed on all the locations visited by the user, in order of appearance, per hour bin over the whole dataset. The daily entropy is computed similarly, but aggregating per day of the week.

pecially powerful because it captures at the same time both the spatio-temporal activity of the user's movement per unit of aggregation and the longitudinal recurrent patterns. For instance, if we observe a user exhibits *every* Monday a sequence of the type {Home $\rightarrow$ Coffee shop $\rightarrow$ Work}, her entropy on Monday will be lower than someone who does not show recurrent habits (both in the same day and across the day of the week over the period of observation). Finally, it should be noted that we do not need to know the exact type of the locations, just that they are the same locations (as per the clustering applied to identify the POI), and therefore the methodology offers great generality.

User Type Characterization

Since the labels assigned by a clustering algorithm do not have any real-world equivalent, we need to characterize those classes indirectly. For this purpose we study several properties for each class both from a statistical and descriptive point of view. There are two metrics that are especially important in human mobility: the radius of gyration, which considers only the spatial aspect of the trajectory, and the entropy, which captures temporal correlations. The radius of gyration r_g is defined as

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{cm})^2},$$

where $\mathbf{r}_i$ represents the $i = 1, \ldots, N$ positions recorded for the user and $\mathbf{r}_{cm} = 1/N \sum_{i=1}^{N} \mathbf{r}_i$ is the center of mass of the user's trajectory. r_g has been found to follow a truncated power law distribution (Gonzalez, Hidalgo, and Barabasi 2008). Such broad distribution is the result of the heterogeneity of users' mobility and it is a strong indicator of multiple classes of users being aggregated together. The entropy S of a user trajectory can be estimated using Equation (1) and is strongly related to the user's trajectory max predictability $\Pi^{\max}$

$$\begin{aligned} S = -\,[&\Pi^{\max}\log_2\Pi^{\max} + \\ &(1-\Pi^{\max})\log_2(1-\Pi^{\max})] + \\ &(1-\Pi^{\max})\log_2(N-1). \end{aligned} \tag{2}$$

The lower the entropy, the higher the predictability of a user. At the same time, the larger the radius of gyration, the more likely a user has visited more *distinct* locations and consequently the higher the entropy.

We tested the radius of gyration, the entropy, and the predictability distributions of different classes in two ways. First, by using the two-sample Kolmogorov-Smirnov (KS) test (Young 1977) with the underlying hypothesis that different classes should exhibit different distributions for the considered trajectory properties. The KS statistic can be used to test whether two underlying one-dimensional probability distributions differ, *i.e.* it checks the null hypothesis H_0 that the two data samples come from the same distribution. Given two samples of size n and m, the two sample KS statistic is defined as

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|\,,$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample. The null hypothesis H_0 is rejected at a confidence p if $D_{n,m} > c(p)\sqrt{\frac{n+m}{nm}}$ where $c(p) = \sqrt{-1/2\ln(p/2)}$ equals to 1.36 for a 95% confidence. To study each cluster we run this test for all the possible combinations of cluster labels (which for n clusters, they are the combinations of two elements of the first n integers. The number of such combinations is $\binom{n}{2}$). We further test that the distributions of the quantities under study are effectively not the same by using the k-sample Anderson-Darling (AD) test (Scholz and Stephens 1987). It tests the null hypothesis H_0 that k samples are drawn from the same population without having to specify the distribution function of that population.

Next, we characterized the classes using the content of the tweets posted by the users to associate a qualitative description, a label, to specific spatio-temporal properties of the user's trajectories. We focused on the study of the profile of locations visited by the users inside each cluster, that is the distribution of the type of locations. Location types are obtained by parsing the text of the tweets for a Foursquare check-in from which we extract the type of the location. Foursquare and Twitter actively prevents "crawling" of their website, therefore we extracted location only from a random sample of 20% of the tweets from each class. We collected a total of 1,987 check-ins. We also consolidated the name of similar categories following the hierarchy used by Foursquare[1]. The final categories are: Restaurant

[1]https://developer.foursquare.com/docs/resources/categories

& Food, Nightlife Spot (*e.g.* bars, pubs, nightclubs), Coffee Shop (originally it was listed under restaurant & food, but we kept it separately as it is usually popular at any time of the day), Arts & Entertainment (*e.g.* events, concerts, museums, monuments), Outdoors & Recreation (*e.g.* gyms, parks, scenic views), Shop & Service (*e.g.* stores, markets), Travel & Transport (*e.g.* hotel, motel, bus/train/metro stations), Professional & Other Places (*e.g.* offices, schools, hospitals, private homes). The distribution of the check-ins categories shows the restaurant and food category being the most popular, as expected (Table 2).

Table 2: Foursquare Check-ins Category Distribution

Category	Size (%)
Restaurant & Food	22.395571
Nightlife Spot	16.809260
Outdoors & Recreation	16.758933
Shop & Service	11.373931
Professional & Other Places	10.317061
Travel & Transport	7.498742
Arts & Entertainment	6.894816
Coffee Shop	5.938601

The distribution of types of places is expected to vary significantly for example between locals, who have a typical home-work routine, and tourists who will visit mostly monuments and attractions. One caveat is that restaurants, bars, and coffee places are perceived as generally popular among every user category. Also, we should note that while such characterization would not be possible on a dataset of only GPS points and associated timestamps, in the context of this work the additional data extracted from the text of the tweets are solely used to get a better understanding of the efficacy of the methodology we introduced.

Experimental Results

Popular-times and Popular-days

In each of the clusters of users we can clearly observe typical hourly and daily patterns (Figure 3). For instance, there are groups of users with peaks early in the morning, at noon, at 8pm, and at 1 am. Weekly activity is more balanced with only few clusters showing a clear week-day versus weekend patterns.

The KS statistic confirms that most users' clusters exhibit significantly different distributions for the quantities of interest (compactly represented using boxplots in Figure 4). The AD statistic for the 8 samples from the clusters supports the KS statistic ($p \approx 10^{-10}$ for the radius of gyration, $p \approx 10^{-66}$ for the estimated entropy, and $p \approx 10^{-29}$ for the max predictability). However, there are pairwise exceptions for the KS statistic. The radius of gyration distribution does not vary much between users probably due to the limited area from which the data was collected. For example clusters 0 and 4 ($p \simeq 0.50$), clusters 4 and 12 ($p \sim 0.60$), and clusters 7 and 15 ($p \simeq 0.45$). There are also exceptions for the entropy, such as clusters 1 and 12 ($p \simeq 0.13$) and 1

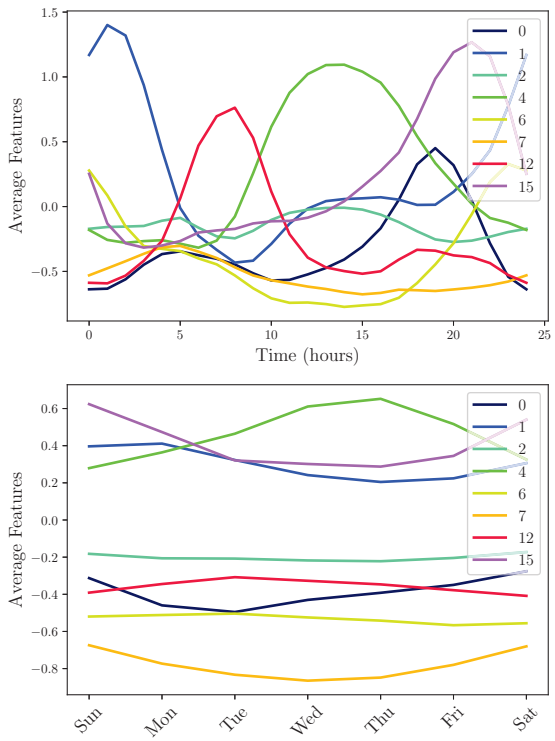

Figure 3: Average popular-times (top) and popular-days (bottom) feature vectors for each cluster of users. The labels in the legend represent the cluster ID. Distinct hourly and daily patterns are visible.

and 15 ($p \simeq 0.14$). Finally, clusters 4 and 12 ($p \simeq 0.52$) and clusters 7 and 12 ($p \simeq 0.64$) did not reject H_0 for the max predictability.

We then conducted the analysis of the type of places visited by the users of each cluster using the Foursquare check-ins (Figure 5). We observed that every cluster with a late-hour activity peak (0, 1, 6, 15) has restaurant and nightlife spot as popular categories. Cluster 1 is an exception having peaks for travel, transport and outdoor. The anomaly of cluster 1 can be explained by the fact it is possible that late at night people tweet on their way back home. Clusters with morning or afternoon activity (4 and 12) have a high activity of shops, professional places and outdoor. This confirms that hourly and daily peak activities reflect into the type of locations visited by a user and it is strongly related to the mobility.

Hourly and Daily Entropy

We can observe distinctive hourly and daily patterns (Figure 6). Clusters 1 and 2 are associated with higher entropy in the weekends and late at night. Cluster 5 is associated with lower entropy during the weekends and a sudden increase of entropy at noon. Cluster 6 resembles the typical average behavior (notice the circadian rhythm) of the population as a whole.

Once again, most of the pairwise KS statistic reject the

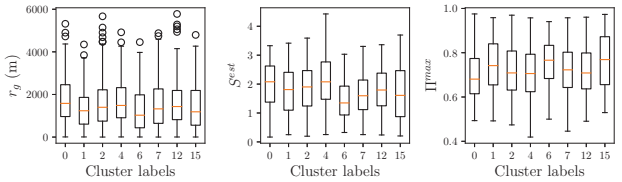

Figure 4: The distributions of the radius of gyration, entropy estimate and predictability of the trajectories among the clusters resulting from popular-times and popular-days. Some of the clusters of users exhibit significant differences pointing to potentially fundamentally different movement patterns between types of users.

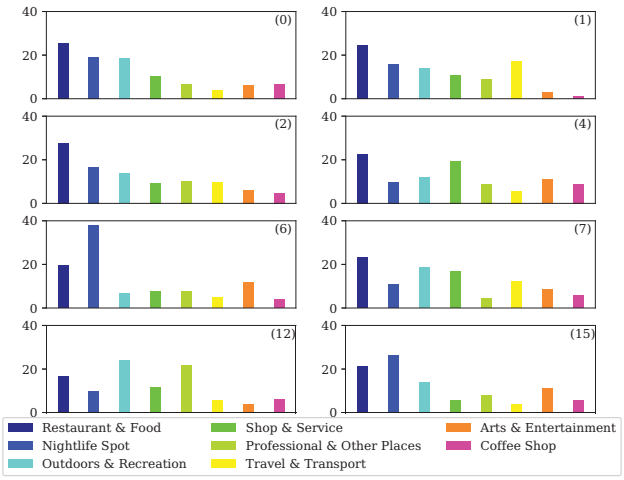

Figure 5: Each plot represents the breakdown (%) of location categories for the corresponding cluster (identified by the label in the top right corner). There are distinctive distributions of categories characterizing each cluster of users.

null hypothesis with a 95% degree of confidence, with the sole exception of the pair of clusters 1 and 5 with regard to the radius of gyration. The 4 sample AD statistic is also strongly in agreement ($p \approx 10^{-41}$ for the radius of gyration, $p \approx 0$ for the entropy, and $p \approx 10^{-161}$ for the predictability). A compact representation of the cluster distributions of the radius of gyration, the entropy, and the max predictability is represented in Figure 7. We observe cluster 1 has very low entropy with several outliers (resulting from a bimodal distribution which for lack of space cannot be represented here), and cluster 6 has a much higher entropy (lower predictability) than the other clusters.

The analysis of the distribution of the types of locations inside of the clusters could shed some light on the observed patterns (Figure 8). Cluster 1 location representation has a bimodal distribution; a high representation of nightlife and restaurant (nocturnal activities) is combined with the art and entertainment category (diurnal activity). Cluster 2 is mainly linked to activities in the evening and night. Cluster 5 shows a high representation of outdoors activities, which could explain the unusually high entropy rate in the morning, and shop service and professional places, which could explain the low entropy in the weekends. Finally cluster 6 has no clear pattern being a mix, which would explain the "average" behavior.

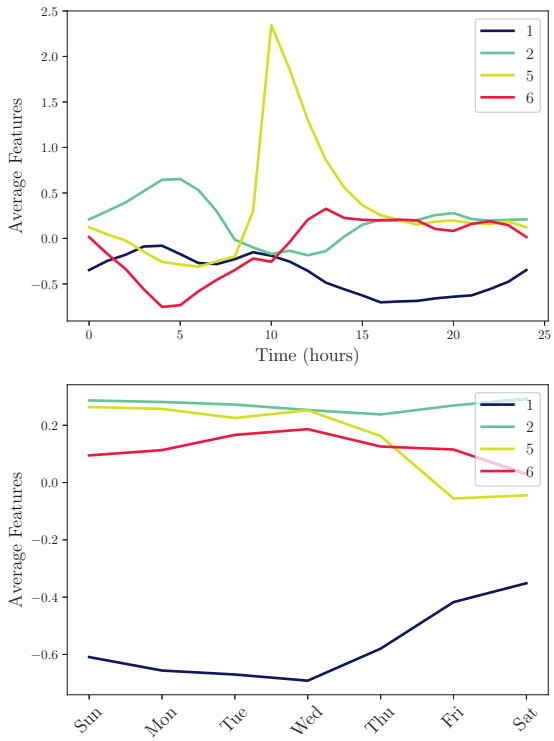

Figure 6: Average hourly entropy (top) and daily entropy (bottom) feature vectors for each cluster of users. The labels in the legend represent the cluster ID. Distinct hourly and daily patterns are visible.

Discussion and Conclusion

We have shown how the popular-times and popular-days along with the hourly and daily entropies are powerful features which can be built from simple GPS trajectories without any additional data. The benefit of such approach is its generality and applicability to datasets coming from different and heterogeneous sources. Such features made possible to identify classes of users that can then be compared using well-understood human mobility metrics and non-parametric statistical tools. The classes of users showed significantly different spatio-temporal patterns associated with the time and day of activity. The efficacy of the methods were further verified by studying the distribution of the categories of the visited locations inside each cluster.

The ability to discern specific characteristics of the users and analyze mobility patterns is important to further the understanding of human mobility and it also finds application in industry such as in location-based recommendation systems (Zheng and Xie 2010; Yang et al. 2013). One question

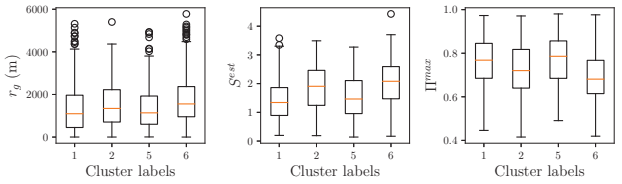

Figure 7: The distributions of the radius of gyration, entropy estimate and predictability of the trajectories among the clusters resulting from hourly and daily entropy. Some of the clusters of users exhibit significant differences pointing to potentially fundamentally different movement patterns between types of users.

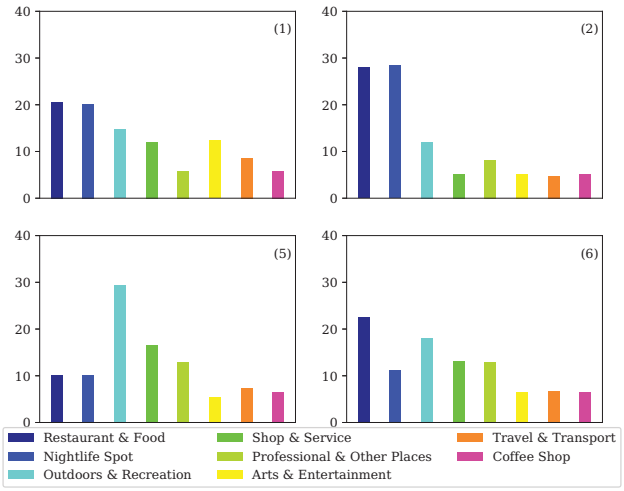

Figure 8: Each plot represents the breakdown (%) of location categories for the corresponding cluster (identified by the label in the top right corner). There are distinctive distributions of location types that characterize each cluster of users.

which remains open is how these user types can be represented by human mobility modeling tools and which impact they have on human dynamics.

References

Amigó, J. M.; Szczepański, J.; Wajnryb, E.; and Sanchez-Vives, M. V. 2004. Estimating the entropy rate of spike trains via lempel-ziv complexity. *Neural Computation* 16(4):717–736.

Barbosa, H.; de Lima-Neto, F. B.; Evsukoff, A.; and Menezes, R. 2015. The effect of recency to human mobility. *EPJ Data Science* 4(1):1–14.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring millions of footprints in location sharing services. *ICWSM* 2011:81–88.

Gabrielli, L.; Rinzivillo, S.; Ronzano, F.; and Villatoro, D. 2014. From tweets to semantic trajectories: mining anomalous urban mobility patterns. In *Citizen in Sensor Networks*. Springer. 26–35.

Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Jenson, S.; Reeves, M.; Tomasini, M.; and Menezes, R. Mining location information from users' spatio-temporal data. IEEE.

Karamshuk, D.; Boldrini, C.; Conti, M.; and Passarella, A. 2011. Human mobility models for opportunistic networks. *Communications Magazine, IEEE* 49(12):157–165.

Mamei, M.; Colonna, M.; and Galassi, M. 2016. Automatic identification of relevant places from cellular network data. *Pervasive and Mobile Computing*.

Myers, S. A., and Leskovec, J. 2014. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, 913–924. ACM.

Neal, D. T.; Wood, W.; and Quinn, J. M. 2006. Habitsa repeat performance. *Current Directions in Psychological Science* 15(4):198–202.

Pappalardo, L.; Simini, F.; Rinzivillo, S.; Pedreschi, D.; Giannotti, F.; and Barabási, A.-L. 2015. Returners and explorers dichotomy in human mobility. *Nature communications* 6.

Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks. *ICWSM* 11:329–336.

Scholz, F. W., and Stephens, M. A. 1987. K-sample anderson–darling tests. *Journal of the American Statistical Association* 82(399):918–924.

Song, C.; Koren, T.; Wang, P.; and Barabási, A.-L. 2010a. Modelling the scaling properties of human mobility. *Nature Physics* 6(10):818–823.

Song, C.; Qu, Z.; Blumm, N.; and Barabási, A.-L. 2010b. Limits of predictability in human mobility. *Science* 327(5968):1018–1021.

Toole, J. L.; Herrera-Yaqüe, C.; Schneider, C. M.; and González, M. C. 2015. Coupling human mobility and social ties. *Journal of The Royal Society Interface* 12(105):20141128.

Wu, L.; Zhi, Y.; Sui, Z.; and Liu, Y. 2014. Intra-urban human mobility and activity transition: evidence from social media check-in data. *PloS one* 9(5):e97010.

Xiao, X.; Zheng, Y.; Luo, Q.; and Xie, X. 2014. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing* 5(1):3–19.

Yang, D.; Zhang, D.; Yu, Z.; and Wang, Z. 2013. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 119–128. ACM.

Young, I. T. 1977. Proof without prejudice: use of the kolmogorov-smirnov test for the analysis of histograms from flow systems and other sources. *Journal of Histochemistry & Cytochemistry* 25(7):935–941.

Zhang, Y.; Hao, J.; Zhou, C.; and Chang, K. 2009. Normalized lempel-ziv complexity and its application in bio-sequence analysis. *Journal of mathematical chemistry* 46(4):1203–1212.

Zheng, Y., and Xie, X. 2010. Learning location correlation from gps trajectories. In *Mobile Data Management (MDM), 2010 Eleventh International Conference on*, 27–32. IEEE.