The geography of taste: analyzing cell-phone mobility and social events

Citation: Calabrese, Francesco et al. "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events." Pervasive Computing. Ed. Patrik Floréen, Antonio Krüger, & Mirjana Spasojevic. LNCS Vol. 6030. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. 22–37.

As Published: http://dx.doi.org/10.1007/978-3-642-12654-3_2

Publisher: Springer-Verlag

Persistent URL: http://hdl.handle.net/1721.1/77153

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

senseable city lab:.::

Francesco Calabrese
Francisco C. Pereira
Giusy Di Lorenzo
Liang Liu
Carlo Ratti

The geography of taste: analyzing cell-phone mobility and social events

The geography of taste: analyzing cell-phone mobility and social events

Francesco Calabrese[1], Francisco C. Pereira[1,2], Giusy Di Lorenzo[1], Liang Liu[1], Carlo Ratti[1]

[1]MIT Senseable City Laboratory, Cambridge, MA
[2]Centro de Informatica e Sistemas da Universidade de Coimbra, Coimbra, Portugal
fcalabre@mit.edu, camara@dei.uc.pt, giusy@mit.edu, liuliang@mit.edu, ratti@mit.edu

Abstract. This paper deals with the analysis of crowd mobility during special events. We analyze nearly 1 million cell-phone traces and associate their destinations with social events. We show that the origins of people attending an event are strongly correlated to the type of event, with implications in city management, since the knowledge of additive flows can be a critical information on which to take decisions about events management and congestion mitigation.

1 Introduction

Being able to understand and predict crowded events is a challenge that any urban manager faces regularly, particularly in big cities. When it is not possible to determine the exact numbers (e.g., from ticket sales), the typical approach is based on intuition and experience. Even when the exact number of event attendees is known, it is still difficult to predict their effect on the city systems when traveling to and from the event. During the last years, the Pervasive Computing community has developed technologies that now allow us to face the challenge in new ways. Due to their ubiquity, GSM, bluetooth or WiFi localization technologies such as in [1–3] can now be explored at a large scale.

The development of methodologies that allow for an accurate characterization of events from anonymized and aggregated location information has further potential implications for Pervasive Computing research, namely enhancing the context awareness. Location based services can be imagined that take into account the predicted effect of events in the city. For example, navigation systems that try to avoid the predicted congested areas, social applications that lead people to (or away from) the "crowds" or interactive displays that adapt to the expected presence of people. Other applications could include inference of points of interest or emergency response planning.

In this paper, we present our work on the combination of analysis of anonymized traces from the Boston metropolitan area with a number of selected events that happened in the city attracting considerably sized crowds. The objective is to characterize the relationship between events and its attendees, more specifically of their home area. The hypothesis is that different kinds of events bring people from different areas of

the city according to distribution patterns that maintain some degree of constancy. The rationale is that people maintain regular patterns of preferences throughout time (e.g., a sports fan will often go watch games; a family that has children will often go to family events). While we make no assumptions on the distributions of "types of people" among areas of a city, it is reasonable to assume that aggregate patterns of "types of neighborhoods" will emerge.
The next section is dedicated to further understanding the motivation and context of this work, followed by a review of related work. The explanation of the data involved in this study is then made in section 4 while the core of the paper is presented in section 5, where we present our methodology and experimental results.

2 Motivation

In 2008, a study from the U.S. Federal Highway Administration [4] was dedicated to investigate the economic and congestion effects of large planned special events (PSEs) on a national level. The clearer understanding of the scale of PSEs and their economic influence is essential to achieve a more efficient transportation planning and management of traffic logistics of such events. In that study, the authors find that there are approximately 24,000 PSEs annually with over 10,000 in attendance across USA, or approximately 470 per week. These numbers, possibly similar in other parts of the world, call for application of efficient techniques of crowd analysis. From the point of view of Pervasive Computing, besides the very task of analyzing *digital footprints* obtained from ubiquitous devices, which lies in the crux of this research, other questions arise that transcend this area.
One question is understanding the stability of crowd patterns in medium to large scale events. If regularity is confidently demonstrated, then pattern sensitive services can be developed that improve the events experience (e.g. providing mobility advisory for evacuation after the event). The converse question is also relevant, namely the characterization of different neighborhoods by knowing what kinds of events their residents prefer to attend. This would allow for the construction of emotional/hobby maps of each block, becoming in turn contextual information about space, adding value to location aware systems.
Perhaps the most obvious problems at the local scale and those that we will illustrate in this paper comprise one-off spatial events which involve the movement of large numbers of people over short periods of time. These largely fall within the sphere of entertainment although some of them relate to work, but all of them involve issues of mobility and interaction between objects or agents which generate non trivial problems of planning, management, and control. The classic example is the football match but rock concerts, street parades, sudden entry or exit of crowds from airports, stations, subway trains, and high buildings could be included. Particularly these types of event, however, have tended to resist scientific inquiry, and have never been thought to be significant in terms of their impact on spatial structure, or to be worthy of theory.

3 State of the Art

Before describing the related work, we bring some definitions that collect relative agreement in the literature. Within the topic of *crowd analysis*, we consider *event inference* and *crowd modeling*. The detection of an existence of a crowd given available data (e.g. images about a place, aggregated communications) is the objective of event inference. Such event may or may not be predictable or correspond to an actual public *special event*. The task of crowd modeling consists of building patterns or descriptions of (a) crowd(s) that enable prediction or simulation of crowd behaviour. A successful crowd model allows for useful applications such as predicting the use of a space, planning accessibility, preventing dangerous situations or planning an emergency evacuation, for example. Following [5, 6] we propose to organize crowd modeling according to three levels: microscopic, macroscopic, mesoscopic. At the microscopic level, the individual is the object of study, while at the macroscopic level, we work with groups. The mesoscopic model combines the properties of the previous two, either keeping a crowd as a homogeneous mass but considering an internal force or keeping the characters of the individuals while maintaining a general view of the entire crowd [6].

From the point of view of data collection, the traditional approach consists of aggregating data from control points (e.g. number of tickets sold; nights in hotels, number of people per room; counting people) as well as from surveys provided to randomly chosen individuals (e.g. [7]). During the nineties, research from computer vision brought alternative (and non-intrusive) methods that allowed to extract crowd related features, namely on detecting density (quantity of people over space), location, speed and shape (e.g. [8]). Although such properties allow for useful analysis, they are restricted to the space of study (or spaces of study, depending on the number of cameras available).

The often mentioned outburst of mobile phones during late 20th century accompanied by the more recent trend of sensors and advanced communication systems (e.g. GPS, digital cameras, Bluetooth, WiFi) allow for unforeseen amounts of data from urban areas through which to study both groups [9–11] , individuals [12] or both [3].

The afore mentioned technologies present different challenges and potential regarding event inference. The traditional methods are slow and precise when the event is controlled in space but with little precision in the opposite case (e.g. [7]). Computer vision allows for automatic inference of events also providing some properties such as those referred above but limited to areas with visual data (e.g. [8]). Using digital footprints such as communication or GPS traces, we can reach wider areas but with lower precision in comparison to these methods. In [13], the authors analyse the presence of tourists in a wide area (Lower Manhattan) during a public art installation (the "NYC waterfalls") for 4 months using cell-phone activity. In the Reality Mining project, 100 students from the MIT campus carried smart-phones over 9 months and their social and individual behaviours were analysed using Cell ID and Bluetooth [3]. In a case study of tourism loyalty in Estonia, Ahas et al [14] show that the sampling and analysis of passive mobile positioning data is a

promising resource for tourism research and management. They show that this type of aggregated data is highly correlated with accommodation statistics in urban touristic areas. In a case study in Tawaf during the Hajj, Koshak and Fouda [11] verified how GPS and GIS data can be utilized to perform tempo-spatial analysis of human walking behavior in an architectural or urban open space.
In terms of level of detail, traditional methods are generally adequate for macroscopic detail (unless individualized data is collected), computer vision allows for any of the levels but is particularly suited to macro- and mesoscopic analysis while digital footprints can be useful for any of the levels discussed, namely microscopic when individual privacy is properly protected. Of course, the precision is dependent on the penetration rate of the technology of study (e.g. number of cell-phone users in the crowd).
As for modeling of crowd behaviour, related work can be found at several distinct fields. In computer vision, crowd models are built as representations of recurrent behaviours by analysing video data of the crowd through vision methods. In physics, many approaches have been built inspired by using fluid dynamics [15], swarms [16, 17] or cellular automata [18]. In literature, there is no characterization of particular "special events" bounded in time and space and in general their goals are at the mesoscopic level (model group from aggregated individual modeling). Also, these studies of digital footprints have used aggregated information of people, rarely reaching the (anonymized) individual detail.

4 Data description

The data analyzed corresponds to an area of 15×15 kilometers within Boston, as shown in Figure 1. This area includes the main event venues in the state of Massachusetts and some of the most densely populated residential areas of Greater Boston. We analyzed cellphone mobility and events happening in that area for the period from July 30th to September 12th of 2009, as we describe next.

4.1 Cellphone mobility data

The dataset used in this project consists of anonymous cellular phone signaling data collected by AirSage[1], which turns this signaling data into anonymous locations over time for cellular devices. This aggregated and anonymous cellular device information is used to correlate, model, evaluate and analyze the location, movement and flow of people in the city. The dataset consists of 130 millions of anonymous location estimations - latitude and longitude - from close to 1 million devices (corresponding to a share of approximately 20% of the population, equally spread over space) which are generated each time the device connects to the cellular network, including:

- when a call is placed or received (both at the beginning and end of a call);
- when a short message is sent or received;

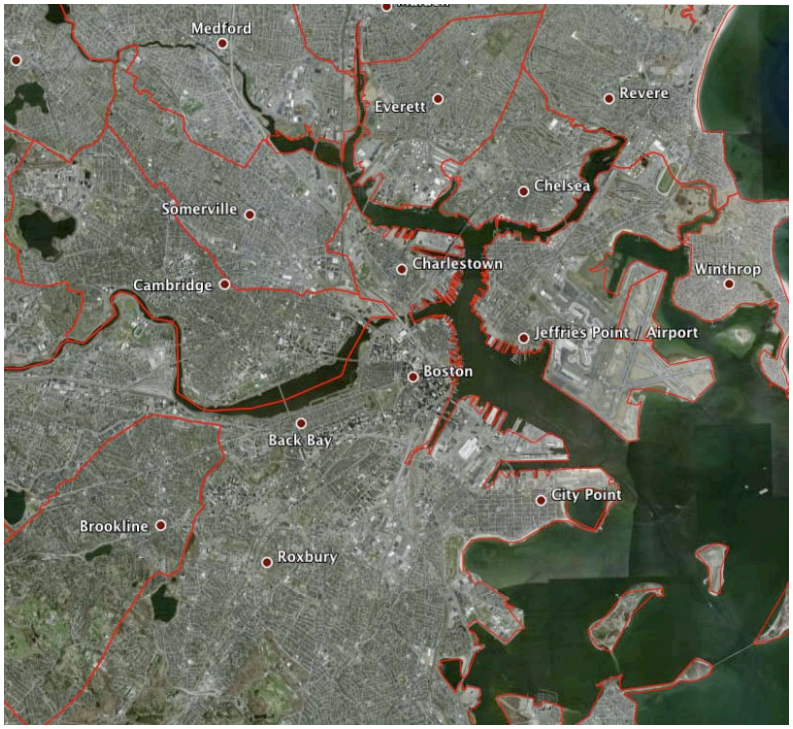


Fig. 1. Study area

– when the user connects to the internet (e.g. to browse the web, or through email programs that periodically check the mail server).

Since the location measurements are generated based on signaling events, i.e. when the cellphone communicates with the cell network, the resulting traces are far from regularly sampled. Besides, cellphone-derived location data has a greater uncertainty range than GPS data, with an average of 320 meters and median of 220 meters as reported by AirSage [1] based on internal and independent tests.

4.2 Events data

Events in the Boston metropolitan area were selected to evaluate whether people from different areas of the city chose to attend different types of events. For the selection of events, it was important to find the largest set that occurs during the time window of the study and that complies with a number of requirements:

– The attendance should have relevant size in order to allow for a significant number of identified users.

– Be isolated in space with respect to neighboring events. To avoid ambiguity in the interpretation of results, we decided to give a minimum margin of one kilometer in any direction to any other large size simultaneous event.
– The venue of the event should correspond to a well defined area with considerable dimensions. It is also important to minimize the potential to misinterpret people staying in other places for event attendees (e.g. staying in a restaurant nearby).
– Be isolated in time to any other big event (i.e. not be in the same day). For a proper analysis, it is also important to guarantee that the statistics of presence (or absence) of people in the events is minimally dependent on external events as this would lead to erroneous conclusions.
– Have a duration of at least 2 hours. The assumption is that attendees are at the venue specifically for the event. With small time durations, it becomes difficult to distinguish occasional stops from actual attendance.

Our goal was to reduce the influence of dependencies between different events and the ambiguity in determining whether a person is attending an event or simply staying in a place near. Another concern was to select events from a variety of categories, namely Performance Arts, Sports events, Family events, Music and Outdoor Cinema.

We analyzed the Boston Globe event website [19] and selected 6 different venues, corresponding to a total of 52 events. We also contacted the organizers of some events in order to get their attendance estimations. In Table 1, we show a summary of the events.

Venue	**Events**	**Type**	**Date**	**Time**
Fenway Park	11 Red Sox games (baseball)	Sports	10, 11, 12, 25 and 26 Aug, 8, 9 September	7-10pm
Agganis Arena	Cirque du Soleil Alegria (2 times)	Performance Arts	26, 27 of Aug.	7:30-10pm
DCR Hatch Shell	Friday flicks (5)	Cinema	31 July, 7, 14, 21 and 28 August	8-10pm
DCR Hatch Shell	Summer concerts (5)	Music	5, 12, 29 and 26 August, 2 September	7-9pm
Museum of Science	Friday nights (7)	Cinema	31 July, 7, 14, 21 and 28 August, 4 and 11 Sep.	5-9pm
Boston Common	Shakespeare on the Boston Common (15)	Performance Arts	31 July, 1, 2, 4-9, 11-16 August	8-10pm
Children's museum	Target fridays (7)	Family	31 July, 7, 14, 21 and 28 August, 4 and 11 Sep.	5-9pm

Table 1. Event list.

It is notable that two of the cases violate one or more of the requirements, namely indoor cinema in the Museum of Science at the same time as the cinema sessions in the Hatch Shell and with an intersection with the Children's museum event. The Cirque du Soleil event also conflicts with the summer concerts at the Hatch Shell. The reason is that, since the venues are far apart and only one has space for very large crowds (Hatch Shell), the overall results should not be affected. In figure 2, we show the event locations.

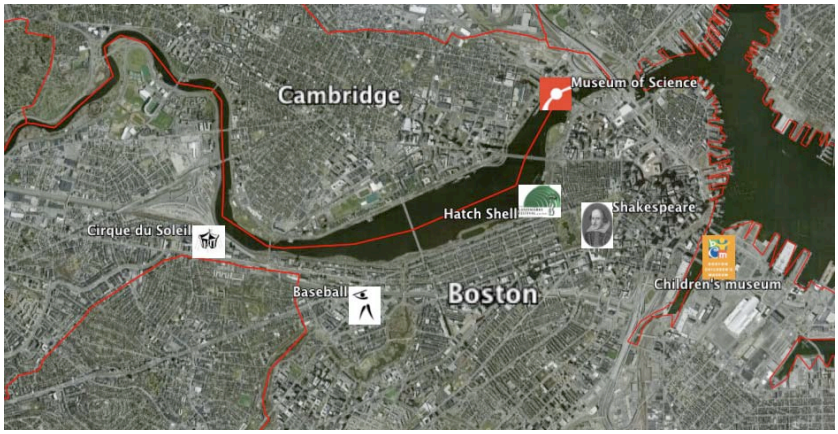


Fig. 2. Event locations

4.3 Data preparation

The data as provided does not directly allow determining mobility traces of users. We then applied a process to perform an estimation of the mobility choices each user takes over time. The process involves two steps:

- Inferring what we call *stops*: places in which a person has stopped for a sufficiently long time.
- Inferring the home location of each user.
- Performing a spatio-temporal analysis of the sequence of stops to detect which users are attending a given event.

In order to infer the sequence of stops that each user makes, we first characterized the individual calling activity and verified whether that was frequent enough to allow monitoring the user's movement over time with fine enough temporal resolution. As we said in the section 4.1. each location measurement m_i, collected for every cellphone, is characterized by a position p_i, expressed in latitude and longitude, and a timestamp,

t_i. For each user we measured the interevent time i.e. the time interval between two consecutive network connections (similar to what was measured in [20]). The average interevent time measured for the whole population is 260 minutes, much lower than the one found in [20]. Since the distribution of interevent times for an individual spans over several decades, we further characterized each calling activity distribution by its first and third quantile and the median. Fig. 3 shows the distribution of the first and third quantile and the median for the whole population. The arithmetic average of the medians is 84 minutes (the geometric average of the medians is 10.3 minutes) which results small enough to be able to detect changes of location where the user stops as low as 1.5 hours (time comparable to the average length of the considered social events).

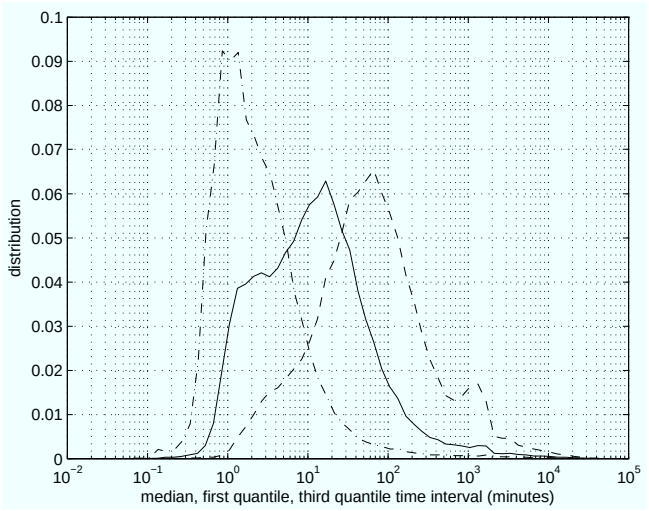


Fig. 3. Characterization of individual calling activity for the whole population. Median (solid line), first quantile (dash-dotted line) and third quantile (dashed line) of individual interevent time.

The analysis above tells us that the cellphone data can be used to extract users' movements as it changes over the course of the day. To extract the sequence of stops, we first extracted trajectories from the individual location measurements. A trajectory is a sequence of chronological locations visited by a user.

$$Traj = \{p_1 \rightarrow p_2 \rightarrow .... \rightarrow p_n\}$$

A sub-trajectory is obtained by segmenting the trajectory with a spatial threshold ΔS, where $distance(p_i, p_{i+1}) > \Delta S, i = 1..n$. The segmentation aims at removing spatial gaps between two recorded points (p_i, p_{i+1})

of more than ΔS. If a gap is found, p_i becomes the end point of the last sub-trajectory, and p_{i+1} becomes the starting point of the new sub-trajectory. Once sub-trajectories are detected, we first resampled with a constant sampling time T_c and then applied to them a low pass filter in order to eliminate some measurement noise contained in the data (as done in [21] [22]). For each sub-trajectory we determined the time at which the user stops traveling, and call the location stop s.
The extraction of a stop depends on two parameters: time distance threshold (T_{th}) and a spatial distance threshold (S_{th}). Therefore, a single stop s can be regarded as a virtual location characterized by a group of consecutive location points

$$P = \{p_s, p_{s+1}, ..., p_m\},$$

where $\forall s \leq i, j \leq m$, $max(distance(p_i, p_j)) < S_{th}$ and $t_m - t_s > T_{th}$.
Once the stops have been extracted, the home location of each user is then estimated as the most frequent stop during the night hours.

The information about the stops and home location allows us to derive the mobility choices of users, and detect whether they are attending an event, and the origin of the trip to attend the event.
Hence, we first grouped together users that live close in space (their home location is close), creating a grid in space where the side of each cell is 500 meters. Then, to understand if a user is attending an event we checked the following assumptions: i) the user stops in the same cell of the event location, ii) the stop overlaps at least 70 percent with the duration of the event, and iii) the user's home location is different from the event location. The Figure 4 shows the idea behind these assumptions. We do not require a full overlap to take into account the fact that we are not able to detect locations of users with a very high frequency, and so might not consider users just because they do not connect to the network at the beginning and end of the event.

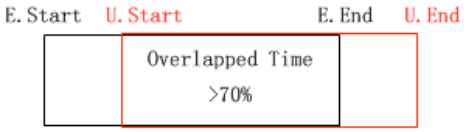


Fig. 4. Audience detection algorithm: if intersection of duration of user stop and duration of the event is greater than 70 percent and user's home is not the same as the event location, then we mark the user as audience of the event

Finally, the mobility choices are derived by inferring the spatial origins' distribution of the people that attempt to the events. Given an event, for each cell of the grid we count the number of people attending to that event and whose home location falls inside that cell. This spatial distribution can then be plot on a map to show the areas of the city which are more interested in attending the event. Examples of such map are shown in the following section.

5 Methodology

Our methodology for describing events through mobility choices is based on the use of the estimated origins of people attending to the events. Figure 5 shows some examples of spatial variation of the estimated origins of people attending different events.

Sport events such as baseball games (Figure 5(a)) attract about double the number of people which normally live in the Fenway Park area. Moreover, those people seem to be predominantly attended by people living in the surrounding of the baseball stadium, as well as the south Boston area (Figure 5(b)).

Performing arts events such as the "Shakespeare on the Boston Common" (Figure 5(c) and 5(d)) which his held yearly, attract people from the whole Boston metropolitan area, and very strongly people which live in the immediate surroundings of the Boston Common (average distance lower than 500 meters). The number of people attending the event is instead about 1.5 times greaten than what it is usually found in the Boston Common.

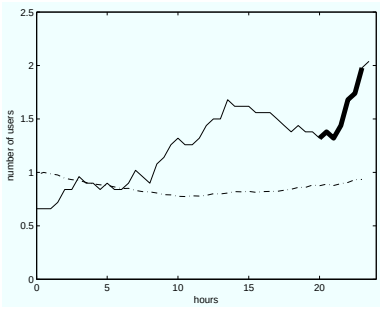
By comparing the two images in Figures 5(b) and 5(d) it is easy to understand that most of the people attending to one type of event are most probably not attending the other type of events, showing a complementary role of sports and arts events in attracting different categories of people.

Finally, Figures 5(e), 5(f), 5(g), 5(h) show the spatial distribution of origins of people for two events (movie screening) happening almost at the same time in two very close areas in Boston (DCR Hatch Shell and Museum of Science).

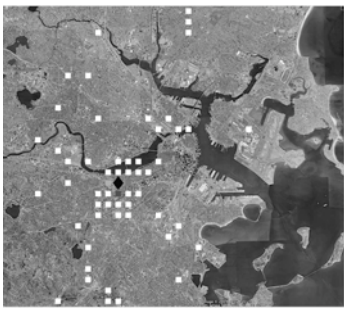
Since the origins of people attending an event are strictly related to the location and type of events, we argue that by using just this information we would be able to predict the type of event. If a relationship between origin of people and type of event is found, it would be possible to determine the abnormal and additive travel demand due to a planned event by just considering the type of that event. It would then be possible to provide a city with critical information on which to take decisions about changes in the transportation management, e.g. increasing the number of bus lines connecting certain areas of the city to the venue of the event.

In the next section we will show 8 different models that we have developed to perform the prediction of the type of event starting from the mobility data associated with it.

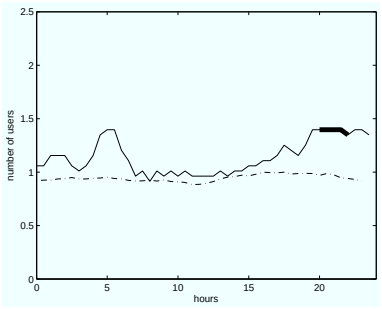
Note that the number of attendees we are able to detect is strictly related to the share of the telecom operator partnering with Airsage. We empirically tested that the number of users correspond to about 20% of the population (as reported by the latest US census) and is equally distributed over the different zipcodes. Since we selected only events with relevant size, this allowed us to detect significant numbers of users per event. We verified that those numbers are also consistent for events of the same type, proving that there is a significant and consistent number of detected attendees allowing us to perform the comparative analysis reported in the next section. Estimating the actual number of attendees



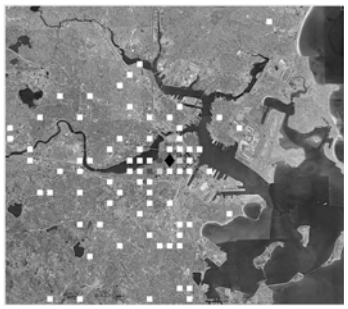
(a) Number of users over time



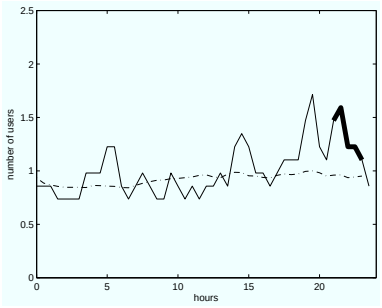
(b) Boston Red Sox vs. Baltimore Orioles at Fenway Park, 2009-9-9



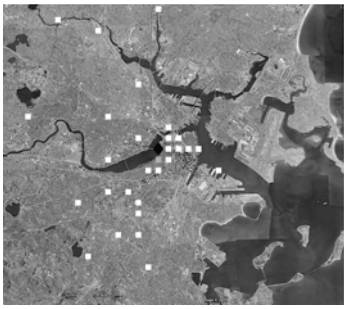
(c) Number of users over time



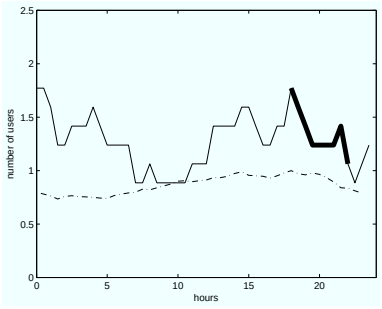
(d) Shakespeare on the Boston Common, 2009-8-13



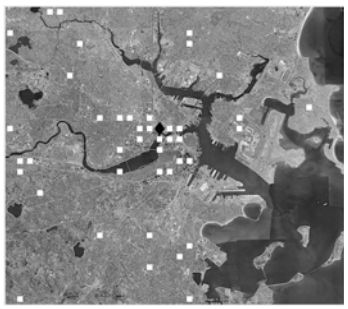
(e) Number of users over time



(f) Friday flicks at DCR Hatch Shell, 2009-8-21 21:00-23:00



(g) Number of users over time



(h) Friday night at Museum of Science, 2009-8-21 18:00-22:00

Fig. 5. Examples of events in Boston. Figures *a*, *c*, *e* and *g* show the number of users at the locations of events over the course of the day of the event (solid line) compared to an average day (dash-dotted line). Note that number of users are scaled with respect to the maximum in an average day. Figures *b*, *d*, *f* and *h* show the locations of the events (diamond) and estimated origins distribution of people attending the events: shade from light (low) to dark (high).

is still an open problem, considering also that ground truth data to validate models is sometime absent or very noisy (usually based on head counts or aerial photography).

5.1 Prediction

The task at hand is to understand the relationships between events and origins of people. Particularly, we seek for the predictive potential of events in respect to mobility phenomena. This can be seen from two perspectives: a classification task in which we want to understand how a vector of features (e.g., attendees origin distribution) predicts a classification (e.g., an event name or type); a clustering task, in which the feature vectors are distributed according to similarity among themselves. We used the Weka open source platform [23], which contains a wide range of choices for data analysis. For classification, we use a Multilayer Perceptron, with one hidden layer and the typical heuristic of $(classes + attributes)/2$ for the number of nodes. For clustering, we apply the K-Means algorithm (with $K = \#$ event types or $K = \#$ event places). In each experiment, we used 10-fold cross-validation, in which a tenth of the dataset is left aside for testing the algorithm while using the remaining for training. This train-test process is ran 10 times (one for each tenth of the dataset).

6 Experiments

We aggregated attendees in terms of zipcode area and distance to event, discretized in 2000 bins. We did so because if we were to use a geographic coordinate of individuals, the resulting data would be sparse. Instead, by aggregating data geographically, we could find useful patterns. To avoid the strong bias towards attendees in the neighborhood of the event, we also remove those that live in the same area of the event (their home location falls in the same 500m x 500m cell of the event) because we would not be able to distinguish between event and home.
For each event, we created an *instance* that contains the corresponding attendee *origin pattern distribution*, evaluated at the level of the zipcode area (with average size of $4.5km^2$). For example, for one showing of the Shakespeare's "Comedy of Errors" at the Boston Common, we have 96 attendees (users monitored by the system, with a share of about 20% of the population) and then count the total number of people coming from each zipcode.
Our goal is to test whether similar events show similar geographical patterns. More specifically, given *origin pattern distribution*, the goal is to predict the type of event (as defined in Table 1).
We met this goal by testing 8 prediction models, and we measure their accuracy in terms of fraction of correctly identified event types.
Before training our algorithms, we analyzed the overall distribution of events to get the *classifier* baselines. The principle is to know the accuracy of a classifier that simply selects randomly any of the 5 event types or that always chooses the same event type, and use them as a baseline

to compare for the improvement of the quality. The average value of this baseline is 23.34% (standard deviation of 4.03) for random classification. Differently, if the classifier chooses the event with highest probability (performing arts), the accuracy will be 35%.
The first experiment was to use all vectors as just described, applied to a Multilayer Perceptron. The result is a surprising 89.36% of correctly classified events in the test set. From the clustering analysis, we see that mostly attendees come from the event's zipcode area, suggesting that people who live close to an event are preferentially attracted by it. To focus on effects other than close proximity, we created a new prediction model considering only people coming from zipcode different from the event's.
The result is 59.57%, which still indicates the recurrence of origin patterns for events of the same type. A clustering analysis brings the distributions that we can see in Figure 6.
Further analyses were made by putting a minimum threshold of at least 10 attendees for each zipcode area and by using home-event distance instead of zipcode (distance discretized in 2000 meter bins). The overall process of feature selection and attendee aggregation is the same as described above, and Table 2, shows the results. The item "Improvement" corresponds to the difference to the best baseline (fixed).

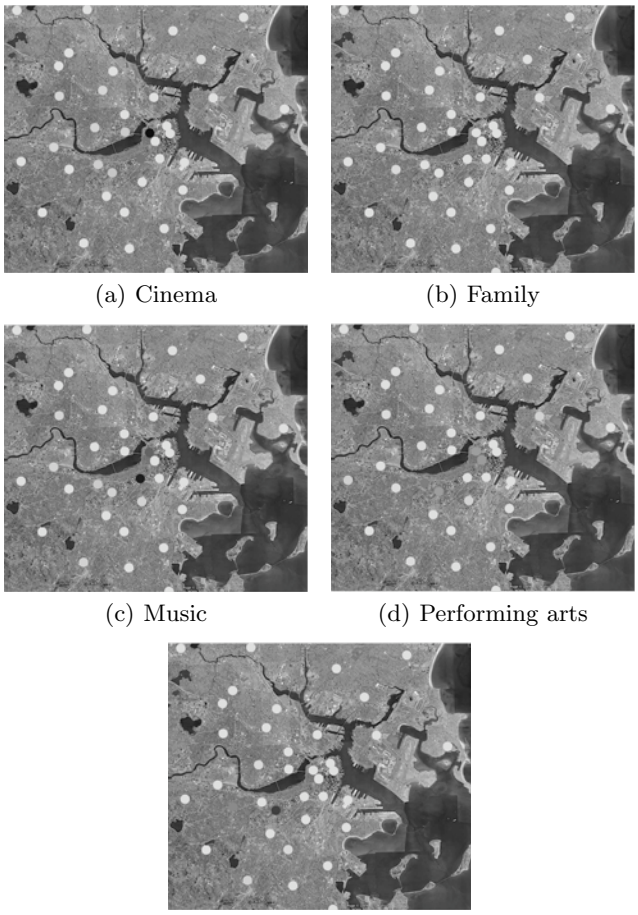
	All attendees		**Exc. event zipcode**		
Features	**Precision**	**Improv.**	**Precision**	**Improv.**	**Observation**
Fixed baseline	35%				Always choose same class
Random baseline	23.34%				Random choice
Zipcode	89.36%	54.36%	59.57%	24.57%	All attendees
	95.74%	60.74%	53.19%	18.19%	All attendees when count>10
Distance	51.06%	16.06%	48.9%	13.9%	All att. Resolution 2000m

Table 2. Summary of prediction results

A first aspect that easily comes out of the predictions performed, is the clear difference between our classifiers and the baselines, indicating a consistency in the patterns found.
By comparing the results of the two predictions made using the zipcode areas, it is clear that the improvements found are consistent, and do not depend on small number of attendees that can be found sometimes in some zipcode areas.
Interesting conclusions can be taken by comparing the improvement of the models using zipcode and distance. In fact the lower improvement shows that not only distance affects the event choices of people, but also where they live.

6.1 Limitations

Our methodology has two limitations. The location data is not continuously provided but is available only when users are active (call, SMS,



(a) Cinema (b) Family

(c) Music (d) Performing arts

(e) Sports

Fig. 6. Spatial visualization of clusters centroids. The circles correspond to the zip-code areas with value greater than zero. The shade from light (low) to dark (high) is proportional to the value.

data connection). This results in narrowing down the number of users we can analyze.
Secondly, we assign origins to users' home locations regardless of where their trips start. This does not hinder our analysis because we are interested in characterizing the taste of the local communities.
Further studies considering larger datasets of events and cell-phone users should be performed to obtain more statistically significant results.

7 Conclusions

Based on our analysis of nearly 1 million cell-phone traces we correlated social events people go to with their home locations. Our results show that there is a strong correlation in that: people who live close to an event are preferentially attracted by it; events of the same type show similar spatial distribution of origins. As a consequence, we could partly predict where people will come from for future events.
In the future, we will run the same study on datasets of cities other than Boston to verify to which extent the city's individual characteristics affect the patterns found.
Explicit spatial knowledge about crowd environment could also be considered to improve the proposed model.

8 Acknowledgements

Acknowledgments go to Airsage for proving us with the data, and to Leonardo Soto, Daniele Quercia, Mauro Martino and Assaf Biderman for their general feedback.

References

1. Airsage: Airsage wise technology. http://www.airsage.com/
2. LaMarca, A.e.a.: Place lab: Device positioning using radio beacons in the wild. In: Proc. of the Intl. Conference on Pervasive Computing. (2005)
3. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal Ubiquitous Computing **10**(4) (2006) 255–268
4. Skolnik, J., Chami, R., Walker, M.: Planned Special Events - Economic Role and Congestion Effects. Federal Highway Administration, US-DOT (2008)
5. Alexiadis, V., Jeannotte, K., Chandra, A.: Traffic analysis tools primer, traffic analysis toolbox. Federal Highway Administration, US-DOT (2004)
6. Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L.Q.: Crowd analysis: a survey. Machine Vision and Applications (2008)
7. Kelly, J., Williams, P.W., Schieven, A., Dunn, I.: Toward a destination visitor attendance estimation model: Whistler, british columbia, canada. Journal of Travel Research **44** (2006)

8. Davies, A., Yin, J., Velastin, S.: Crowd monitoring using image processing. Electron. Commun. Eng. J **7**(1) (1995)
9. Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D.: Mobile landscapes: Using location data from cell-phones for urban analysis. Environment and Planning B: Planning and Design **33**(5) (2006)
10. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. IEEE Pervasive Computing **6**(3) (July-September 2007) 30–38
11. Koshak, N., Fouda, A.: Analyzing pedestrian movement in mataf using gps and gis to support space redesign. In: The 9th International Conference on Design and Decision Support Systems in Architecture and Urban Planning. (2008)
12. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453** (2008)
13. Girardin, F., Vaccari, A., Gerber, A., Ratti, C.: Quantifying urban attractiveness from the distribution and density of digital footprints. Journal of Spatial Data Infrastructures (4) (2009)
14. Ahas, R., A, A.K., Tiru, M.: Spatial and temporal variability of tourism loyalty in estonia: Mobile positioning perspective. In: Proceedings of the Nordic Geographers Meeting (NGM09). (2009)
15. : Social force model for pedestrian dynamics. Phys. Rev. E **51**(5) (1995)
16. Bellomo, N.: Modeling Crowds and Swarms:Congested and Panic Flows. In: Modeling Complex Living Systems. (2008)
17. Banarjee, S., Grosan, C., Abarha, A.: Emotional ant based modeling of crowd dynamics. In: Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC'05). (2006)
18. Bandini, S., Manzoni, S., Vizzari, G.: Crowd Behaviour Modeling: From Cellular Automata to Multi-Agent Systems. In: Multi-Agent Systems: Simulation and Applications. CRC press (2009)
19. Globe, B.: Events and things to do in boston: Website. http://calendar.boston.com (2009)
20. Gonzalez, M., Hidalgo, C., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196) (2008) 779–782
21. Calabrese, F., Ratti, C.: Real time rome. Networks and Communications Studies **20**(3-4) (2006) 247–258
22. Calabrese, F., Ratti, C., Colonna, M., Lovisolo, P., Parata, D.: A system for real-time monitoring of urban mobility: a case study in rome. IEEE Transactions on Intelligent Transportation Systems, submitted (2009)
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann (2005)