Uncovering the presence and movements of tourists from user-generated content

Fabien Girardin[1], Francesco Calabrese[2], Filippo Dal Fiorre[2], Assaf Biderman[2], Carlo Ratti[2], Josep Blat[1]
[2]Universitat Pompeu Fabra, Barcelona, Spain
{fabien.girardin, josep.blat}@upf.edu
[2]Massachusetts Institute of Technology, SENSEable City Lab, Cambridge, MA, USA
{fcalabre, dalfiore, abider, ratti}@mit.edu

In recent years, the large deployment of mobile devices has led to a massive increase in the volume of records of where people have been and when they were there. The analysis of these spatio-temporal data can supply high-level human behavior information valuable to social scientists, urban planners and local authorities. This paper explores this hypothesis by reporting on new information revealed by this pervasive user-generated content. We present novel techniques, methods and tools we have been developing to explore the significance of these new types of data. In a case study of Rome, Italy, we showcase the ability to uncover the presence and movements of tourists from geo-referenced photos they explicitly make public, as well as from network data implicitly generated by users of mobile phones.

Introduction

Nowadays every click of every move of every user who interacts with any software may be gathered in a database and submitted to a second-degree data-mining operation. With the ubiquity of mobile and wireless devices, the logs produced become anchored to the physical space. Their accumulation can reveal human individual and social behaviors [1]. The emergence and the novel access to these digital traces coming from the "virtual world" is expected to have significant impact on the social sciences, as they could finally have access to a quantity of data that are of the same order of magnitude as that of their older sisters, the natural sciences [2].

One type of digital trace comes from people's implicit interaction with infrastructures which results, for instance, in location logs produced when carrying a mobile phone which is in constant dialogue with a wireless network. In addition to these automatically generated and implicit data, another type is generated explicitly by mobile and web users when they publicly disclose content such as photos, messages, sensor measurements or blog posts. As suggested by Goodchild [3], the most powerful sensor web is made of the 6 billion humans occupying the Earth's surface. He argues that this large collection of mobile and intelligent sensors will affect the processes of acquiring and compiling geographic information (i.e. volunteered geographic information).

This paper illustrates the potential of these digital traces to uncover the presence and movement of people in a city. Our case study focuses on tourists in Rome, because sources of digital footprints are multiple in that context: prior to their visits they generate server logs when consulting digital maps and travel web sites, during their trip they leave traces on wireless networks when using their mobile phones, and after their visits they upload online information, reviews and photos. The validation of these records with respect to existing surveys can lead to an improved understanding of different aspects of mobility and travel. In tourism and urban planning, information about who populates different parts of the city at different times can lead to the provision of customized services (or advertising), accurate timing of service provision based on demand (e.g. rescheduling of monument opening times based on the presence of tourists), and, in general, more synchronous management of service infrastructures.

Throughout this paper we show the new aspects revealed by geographically anchored digital traces; and we present novel techniques, methods and tools we have been developing to explore and reveal the significance of these new types of data to uncover people presence and movements in the city. While the accumulation and aggregation of digital footprints makes the emerging processed information more credible and reliable, we explore as well how to validate this pervasively user-generated content.

Digital footprints

In the past years, research in location sensing and tracking has been dominated by the quest for figuring out where persons and objects are in space, not without raising some issues. first, there are privacy and ethical concerns related to collecting data without the individual's consent. Second, mobile tracking solutions that persons must carry generate fatigue effects, provide snapshots that are limited in space and time, and do not always work in urban areas because of signal multipath and urban canyon obstructions. The third alternative is of distributed and fixed solutions that require expensive deployment and maintenance of sensors with onerous data transmission costs to a centralized system. These multiple issues suggest that the research community should investigate and evaluate the use of new data types as well as approaches that do not rely on the deployment of ad-hoc and costly infrastructures.

In our study, we exploit the digital footprints generated by pervasively-situated people using existing services. To do so, we have been developing tools to absorb, process, map, and visualize the vast amount of data generated by people with mobile and wireless devices such as digital cameras and mobile phones. In our previous work, we showed that explicitly-disclosed spatio-temporal data coming from open web platforms can overcome the constraints mentioned previously and provide novel insights on the dynamics of people in an urban space [4].

In this paper, we explore two types of digital traces: georeferenced photos made publicly available on the web, and aggregated records of wireless network events generated by mobile phone users making calls and sending text messages.

Before going any further, let us maintain that collecting and analyzing data about the behavior of individuals raises important privacy issues; people are concerned about revealing the history of their whereabouts to an un-trusted third party. Our approach addresses these concerns at two levels: first, our dataset includes information that users explicitly disclose. They position their photos on a map and have control over who gets access to this location data. Moreover, while obtaining this public information, we applied an obfuscation algorithm to lose the relationship with the web identity of the individual. Thus, we only analyze anonymized records of digital traces publicly disclosed by individuals. Similarly, collection and analysis of aggregated mobile network usage data complies with the 2002 Directive by the European Parliament and Council on privacy. We receive no record about an individual's identity or trajectory, but only the number of people that use a mobile phone at a given area in the city, at a given moment. Individual users cannot be identified based on the data we collected and analyze.

Techniques and tools

Explicit footprints: georeferenced photos

We collected photos uploaded on the popular photo-sharing platform Flickr[1]. People use this service to share and organize photos. They also have the option to add geographical referencing. Each time a photo is virtually linked to a physical location, the Flickr system assigns longitude and latitude values and retrieves the time of capture from the Exchangeable Image File Format

---

[1] http://www.flickr.com

(EXIF) metadata embedded in the image file. The location provided by the user generally indicates where the photo was taken; but sometimes it denotes the location of the photographed object. When the user provides the location of the photo, the zoom level of the map (from 8 for region/city level to the most precise 16 for street level) is recorded to provide an accuracy attribute, completing the spatio-temporal information.

From the publicly available photos, we retrieve the coordinates, timestamp, accuracy level, and an obfuscated identifier of the owner via the Flickr API. In Rome, for the 3-year period from November 2004 to November 2007, we extracted 144501 georeferenced photos uploaded by 6019 people.

A first pre-processing step separates the visitors from residents of the region. To achieve this, we examine the photographer presence in the area over time as the discriminating factor. We divided the time into 30-day periods computed the number of periods each photographer was active in the area. If a photographer took all his/her photos within one period of 30 days, the algorithm considers him/her as a visitor, while if there is an interval greater than 30 days between two photos taken, he/she is categorized as resident. The aim of this strict threshold was to capture the real one-time tourists. Out of a population of 6019 photographers, we identified 4719 one-time visitors.

In addition, we were interested to know more about the nationalities of our photographers. Hence, we took advantage of the Flickr social function that invites users to voluntarily provide “offline bits” on their city and country of residence. We found out that 59% of the users actually disclosed this information. While it is hard to say how much of this data is true, the country of residence could be retrieved by automatically parsing the data in most cases. We had to assign manually a country in some cases, because of spelling errors or idiosyncratic names (e.g. “Big Apple” for New York).

Implicit footprints: wireless network usage

As demonstrated in previous research works, the wide diffusion of mobile phones and the widespread coverage of mobile phone wireless networks in urban areas make these technologies very interesting as means to identify presence of crowds [5]. To collect these data, we collaborated with Telecome Italia, which deployed on their network a system called LocHNESs (Localizing & Handling Network Event Systems) – a software platform that locates and stores events that occur on the mobile network (call in progress, SMS sending, call handover) through external probes. These probes analyze all the incoming signaling messages and send a notification of the detected events to the LocHNESs platform, which periodically localizes the messages and produces aggregated maps of the distribution of users that are engaged in a call, in raster format. Detailed information about the platform can be found in [6].

Telecom Italia installed the LocHNESs platform and the related probes on a group of Base Station Controller (BSC) located in Rome, covering an area of approximately $100Km^2$ in the north-east of the city. The area was divided into a grid of squares with an area of 250 x 250 m^2, and various types of statistical information were produced every 5 minutes. For the scope of this paper we use raster maps that were developed to derive the spatial distribution of foreigners, calculated by grouping the location of those mobile phones whose IMSI (International Mobile Subscriber Identity) numbers were of foreign mobile network operators[2].

---

[2] The first three digits of the IMSI are the Mobile Country Code (MCC) and identify the nationality of the mobile network operator.

Processing and visualization

To reveal the meaning of digital footprints, our approach takes advantage of open and freely-available resources and combines them using de facto standards often based on the eXtensible Markup Language (XML). We use Google Earth for interactive visual synthesis of encodings generated by using a combination of MySQL (for data storage and querying to select and aggregate) and a software we have been developing to access, process, transform, aggregate, cluster, sample, and filter the raw data stored in the database and to generate outputs [4].

Dykes et al., 2007 [7], suggest that this large type of data can only make sense through geovisualization. Thus, after collecting the data, we map and visualize them to explain user behaviors. Tools such as Google Earth provide quick support for visual synthesis and preliminary investigation of digital traces.

Novel aspects revealed by analysis of digital footprints

The application of these techniques and tools to process digital footprints allows us to uncover the presence of crowds, unexpected patterns of movement over time, as well as to compare behaviors and generate new hypotheses.

Spatial presence

A first step in the analysis consists of mapping the digital footprints. Data is stored in a matrix covering the study area. Each cell in the matrix includes data about the number of photos taken, the number of photographers present, and the number of phone calls made by foreigners over a given period of time. As shown in Figure 1, the matrix is made of 20x20 cells of 250x250 m2. This geovisualization reveals the main areas of tourist activity in a part of Rome over a 3-months period from September 2006 to November 2006. The presence of photographers is pictured in the orange color range, while areas of heavy mobile phone usage by foreigners are depicted in the green color range. The union between the two datasets shown in this image uncovers the city's major tourist-attractions such as the Colosseum and Piazza della Repubblica, next to the main train station. In addition, it appears that the Colosseum attracts the photographers, while foreign mobile phone users tend to be active around the train station. Based on these observations, it can be hypothesized that tourists take part in sightseeing activities in the orange areas, while being on the move or waiting in green zones – two rather different types of activities.

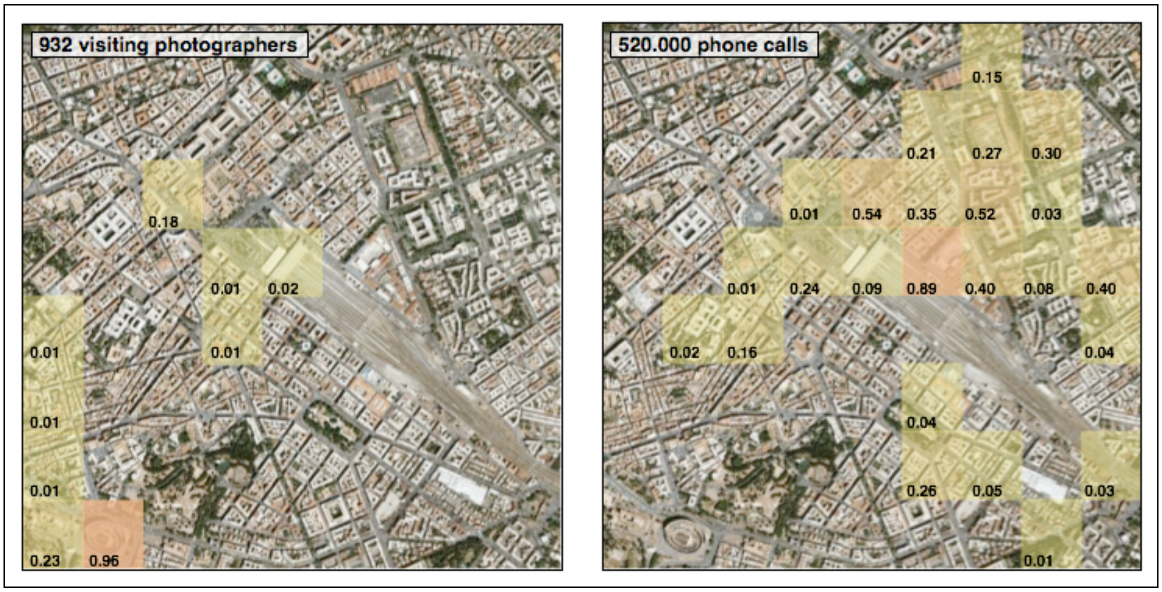

Figure 1. Geovisualizations of the presence of (a) 932 tourist photographers and (b) 520,000 phone calls from foreign mobile phones in the Coliseum and Piazza della Repubblica area from September to November 2006. Both types of data cover the train station area in the proximity of the Piazza della Republica. The values in each cell are normalized.

Temporal presence

The previous paragraph was discussing the difference in tourist activity based on area. This seems to be supported also by analyzing the temporal patterns obtained from the digital traces in the main areas of activity. We captured the number of photographers and the amount of phone calls for each day of the week of the 3-month study period. The graphs below offer a comparison of the average weekly distribution of phone calls made and presence of photographers by visitors between the areas around the Colosseum and Piazza della Repubblica (Figure 2). The histograms show the normalized variation between the average number of calls and the average number of photographs for each day of the week, and the average amount for the whole week. The plotted temporal signatures for the Colloseum area show similar trends, with a higher activity over the weekend than during weekdays. However, a corresponding temporal analysis of data related to Piazza della Repubblica and the train station area reveals two different patterns. Photographers, even if fewer by number than at the Colosseum, tend to be active on the weekend, whereas the foreign mobile phone users generated more voice and data traffic than at the Colosseum and mostly during the week days.

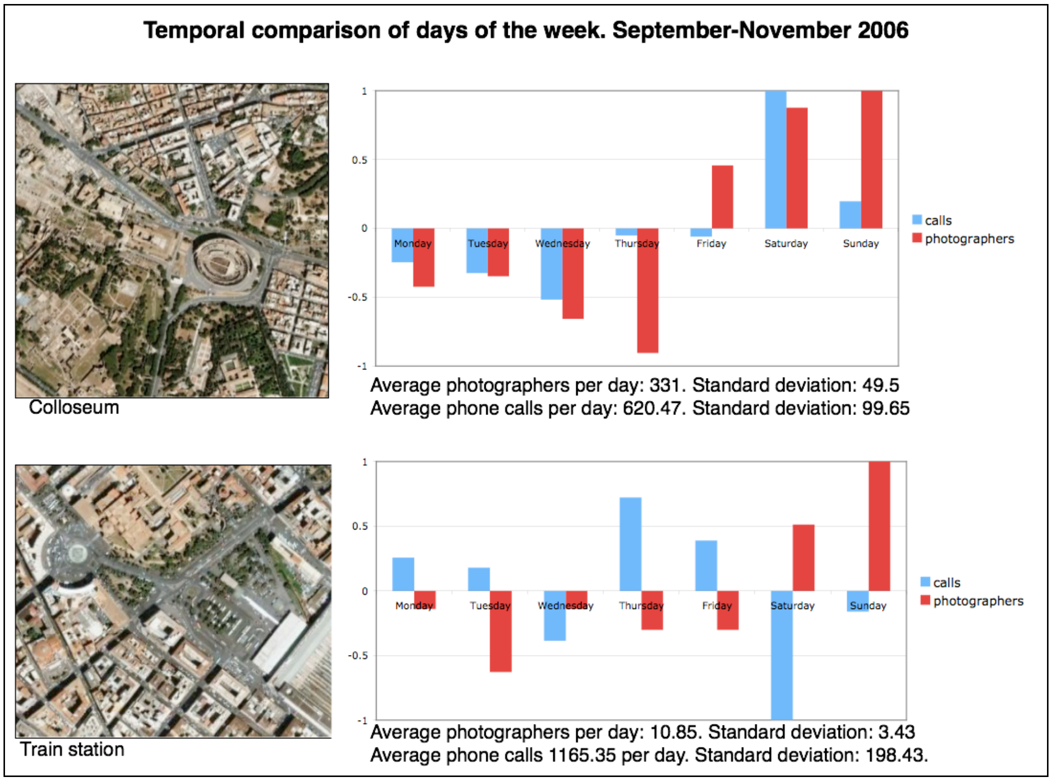

Figure 2. Comparison of the temporal signature of foreigners phone activity and number of tourist photographers. It reveals similar patterns of low below average activity during the week days and a rise of presence over the weekend at the Colloseum. In opposition, the temporal signature of the train station shows a higher presence of foreigners calling from their mobile phones during the week, while photographers indicate a reverse pattern and increased presence over the weekend.

These temporal signatures provide further evidences to the different types of presence that occur at the tourist points of interest. It can be further hypothesized that the Colloseum attracts sightseeing (i.e. photographers) activities over the weekend and the neighborhood of the train station provides facilities for visitors on the move (e.g. people on business trips) during the week days. In other words, While difficult to interpret, is that there are two types of visitors, one present for business that dominates the pattern the weekly pattern of photographers.

Revealing desire lines from digital traces

Another novel aspect of the study of digital footprints that are referenced geographically resides in the ability to uncover digital desire lines. A desire line is a path developed by erosion caused by human footfall. Based on the time stamp and the disclosed location of the photos, our software records the movement of people to form digital traces. We construct a trace from the set of geographically-referenced chronologically-ordered photos taken by one person over the course of one day. Next, we aggregate these personal footprints both for anonymizing and in order to generate the digital desire lines by accumulation. Formatting them in KML[3] should reveal the lines and hints at traveling behaviors of specific types of visitors (Figure 3).

---

[3] http://www.opengeospatial.org/standards/kml/

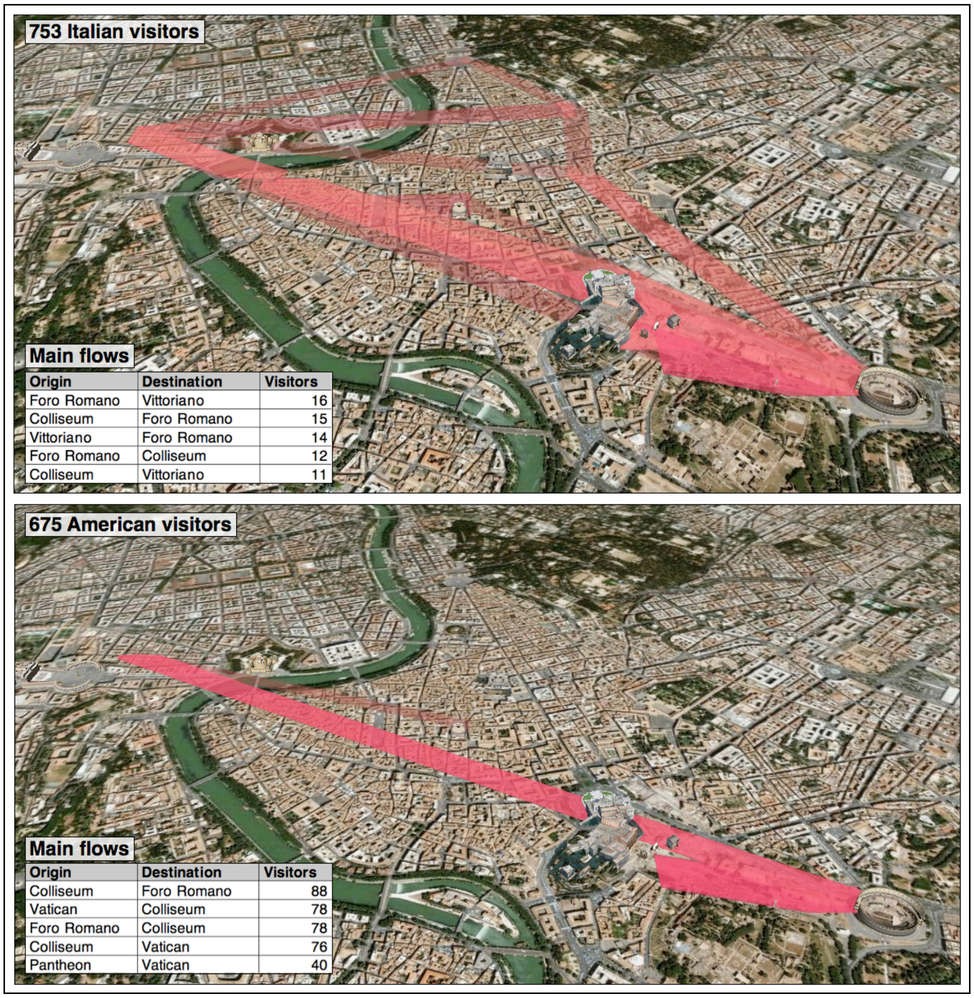

Figure 3. Geovisualiation of the main paths tourists took to visit the points of interests of the city. 753 Italian visitors (top) are activity off the beaten tracks with a large amount of points of interets visited but little flow. 675 American visitors stay on a narrow desire line between Vatican, Foro Romano and Colloseum. Note that different scales imply for each geovisualization.

More precisely, we start by revealing the most active areas obtained by spatial clustering of the data[4]. It results in a set of the main points of interests in the city. The location of each point of the trace (i.e. photo) is then checked if it is contained in a cluster. In case of a match, the point of interest gets added to the trace generated by the owner of the photo. This process produces multiple directed graphs that support better quantitative analysis. We can obtain, the number of sites visited on average, by season, the most visited and photographed points of interests or where do photographers start and end their journeys, etc.

From another perspective, we can build (asymmetric) matrices by aggregating individual digital traces whose elements are the number of photographers who moved from point of interest X to point of interest Y. Depending on specialized queries to collect the digital traces that form the matrix, different patterns of movements can be revealed. For instance, queries can be based on the nationality of the user, number of days of activity in the city, number of photos taken, areas covered during a visit.

---

[4] Direct aggregation of the traces was not providing results easy to visualise and analyze.

Semantic description

Previous work demonstrated that, when subtly aggregated and processed, mobile phone data could reveal events. Other scholars have examined geographically-annotated material available on the Web: Rattenbury et al. [10] used the location and time metadata associated with photos and their tags to extract "place" and "event" semantic information. Following a similar line, we have analyzed the tags linked to the photos in our collection and revealed some clues related to people's perception of their environment and the semantics of urban space from their perspective. For instance, the word "ruins" is one of the most used tag to describe georeferenced photos in Rome. Mapping the distribution of this tag over the city uncovers the most antique and physically destructed part of the city: The Colosseum and the Foro Romano (Figure 4).

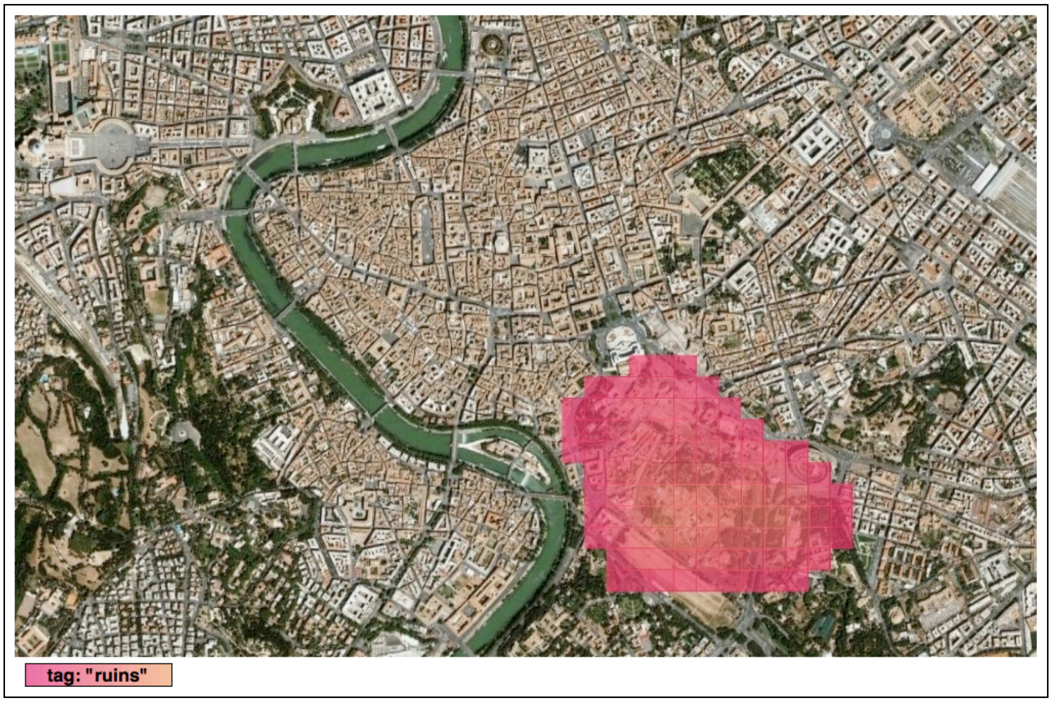

Figure 4. Geovisualization of the areas defined by the position of the 2886 photos with the tag "ruins" uploaded by 260 photographers. It reveals the Coliseum and Foro areas known for their multitude ancient ruins.

The significance of these user-generated data

The accumulated and aggregated records of where and when people were seem to lead to an improved understanding of different aspects of mobility and travel. Arguably, the emerging patterns supplied by the data mining of large sets of user-generated content should provide more reliable insights. We have shown some usage potential for geographically-referenced digital footprints during this paper. However, there is a strong need to examine their significance in comparison with existing mobility and activity data available in cities.

In the context of our study, there are traditional ways to observe and quantify the presence and movements of visitors in a city like Rome that could help us better define the significance of using pervasive user-generated content are hotel occupancy and museum surveys. We could access monthly statistics regarding the number of people who bought a ticket to visit the Colloseum in 2006, provided by the Rome Tourism office. In Figure 4 we compare these figures with the presence data from mobile users (phone calls) and photographers (photos) that was previously collected over a period of 3 months (September to November 2006).

Numbers show that there are slightly more Colloseum visitors in October than September with a major drop of attendance in November. This pattern matches with the activity of foreign mobile phone users in the area. However, it does not coincide with the detected number of photographers. Yet both georeferenced-photos and tickets sold show that the months of April and May are those of the highest activity during the year. We hypothesize that the main reason for these discrepancies is that the datasets relate to different types of visitors. One set of data is generated by visitors paying to visit an attraction, others by mobile service customers wealthy enough to pay roaming charges, and finally, a tech savvy community of people who are familiar with digital photography and social software who upload content on Flickr.

More generally, it seems too early to say whether comparing the different datasets supports better validation of our observations. Due to the large difference in the nature of the activity producing the data we compare, it might be that correlating different datasets does not only reinforce observations, but also reveals additional dimensions of user behavior that we might not yet have accounted for. A current challenge is to understand more precisely the user-activity that is reflected in each of these types of datasets.

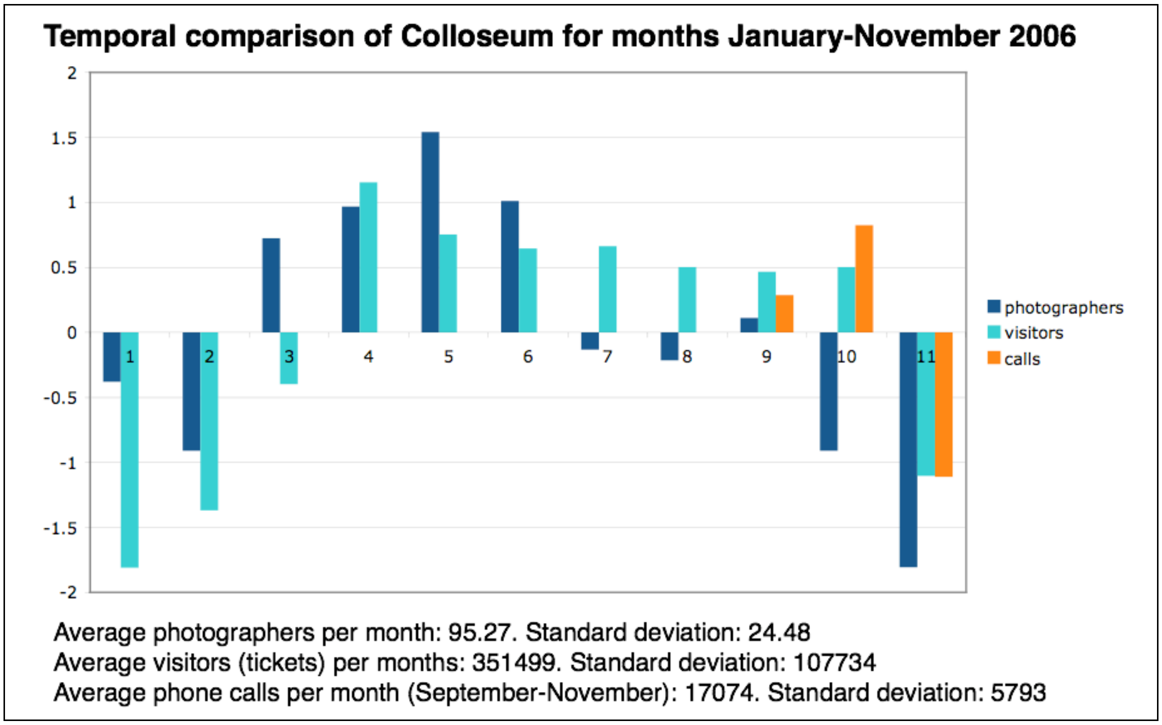

Figure 5. Comparison of the presence of visitors in the Colloseum area from January to November 2006, from the number of tickets sold, number of phone calls and number of photographers active in the zone. The values represent the variations from the monthly average, scaled by the standard deviation.

Data quality

In our analysis we have tried account for the fluctuating quality in the data, which can impede on, or restrict the ability to generate accurate information. For instance, the timestamps extracted from the camera-generated EXIF metadata does not necessarily match the correct time at which a photo was taken. A limitation of the mobile phone dataset arises from the fact that only phone calls handled by cell-phone towers connected to a subset of all the BSCs in Rome were monitored. This study, then, does not consider all areas in the center of Rome, which leads to the presence of “border effects.” – places that are at the borders of a monitored area are also partially covered by cell-phone towers that are connected to BSCs that are not monitored. This implies that part of the calls taking place in such areas were not counted by the LocHNESs platform. The last consideration applies in particular to areas in the city located south west of the Colosseum (see also Figure 1).

Discussion

The explosion in the use of capture and communication devices (e.g. mobile phones, digital cameras) and the introduction of content sharing platforms has led to the emergence of a wealth of georeferenced-data generated by people that can provide new opportunities to urban studies and social sciences. This paper illustrates the value of explicitly-disclosed geographically-referenced photos and implicitly-generated records of mobile phone network usage. We showcased the analysis of these accumulated user-generated digital footprints to uncover novel viewpoints on the presence and movements of tourists during their visit of a city. We introduced novel tools and techniques for this analysis, although the results demonstrate that they must be developed further in order to validate our observations as well as to allow making further ones, for example: the temporal usage-signature of a space, its attractiveness to different groups of people, and its degrees of similarity in terms of different types of usage to other spaces.

From another methodological perspective, the data we have analyzed in this paper have a clear advantage over previous location data which were obtained rather through lab experiments – with subjects carrying sensors and thus being aware of being tracked. Over other commonly used methods, they do not come from samples – although, as we have indicated they might be biased as coming from population segments.

Another promising aspect user generated data lies in its explicit character. Unlike the automatic capturing of traces, the manual disclosure of location in the act of geotagging of photo provides additional qualities: positioning a photo on a map is not simply adding information about its location; it is also an act of communication which contains what people consider as relevant for themselves and others. A specific richness lies in the intentional weight people put on disclosing their photos. We have shown that Flickr users have a tendency to point out the highlights of their discovery of the city and discarding the downtimes of their trip[5].

Our data analysis and visualization are meant to complement traditional surveys and other traditional means to collect information about people's behaviors. For instance, in the pre-digital age we knew how many tourists spent a night in a given hotel; now we can also estimate how much they liked it. We knew how many tourists visited a given attraction site; now we can also know how many uploaded a photo of it and the semantics of their description of it. At the same time, by knowing the roaming cell phone activity we can quantify anecdotal evidence: in the past, we roughly knew the kind of tourists that populated a given square by observing them physically; now we also know the nationalities of tourists who are using a mobile phone.

Future works will aim at further systematic assessment of the quality of these user-generated data and determining their validity limits, as well as the relationships between results that come from different data sources. Another promising approach is to correlate them with the results of traditional surveys.

An additional research avenue lies in understanding the practice of generating and georeferencing data. For instance, under which circumstances do users tag their content with a street address and

[5] And, related to the previous discussion in this section, more implicitly generated footprints help to uncover additional information for this reason, as the example of the presence of visitors in the Piazza de la Repubblica shows.

when is it tagged to a bigger region? The results of such analysis should reveal profiles of behavior of georeferencing and geotagging photos. These profiles might be based on culture (nationality[6], residents vs. visitors), type of tourist (length of stay, number of photos taken), technical expertise, spatial orientation ability, type of tasks (social sharing, individual indexing), or types of urban environments visited (monuments, events, activities). Other questions that should be answered are related to the types of situations during which users are more likely to use their mobile devices. Answers to these types of questions should allow us to define better the meaning of the data and explore further their potential usage in social sciences and urban studies.

Acknowledgments

We would like to thank Barcelona Media,Telecom Italia and and Telespazio for satellite digital imagery for their support. Also, we are indebted to many people at the Massachusetts Institute of Technology and the Universitat Pompeu Fabra for providing extremely stimulating research environments and for their generous feedback. In particular, thanks to Assaf Biderman, Liang Liu, Nicolas Nova, Toni Navarrete and Andrea Vaccari for letting us pick their brains. Of course, any shortcomings are our sole responsibility.

References

1. Eagle, N. and Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268.

2. Latour, B. (2007). Beware, your imagination leaves digital traces, Column for Times Higher Education Supplement, 6th of April 2007. http:// www.bruno-latour.fr/poparticles/poparticle/P-129-THES-GB.doc

3. Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2:24–32.

4. Girardin, F., Dal Fiore, F., Blat, J., and Ratti, C. (2007). Understanding of tourist dynamics from explicitly disclosed location information. In The 4th International Symposium on LBS & TeleCartography.

5. Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. 2006. Mobile landscapes: Using location data from cell-phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727 – 748.

6. Calabrese, F., Ratti, C., Real Time Rome. *Networks and Communication Studies*, vol. 20, nos. 3 & 4, 2006, pp. 247–258.

7, Dykes, J., Slingsby, A., and Clarke, K. 2007. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1176–1183.

8. Rattenbury, T., Good, N., and Naaman, M. 2007. Towards automatic extraction of event and place semantics from flickr tags. In SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 103–110, New York, NY, USA. ACM Press.

---

[6] For instance, at a very general and caricatural level, the Flickr dataset in Rome suggests that 123 German users tend to provide more accurate location information to their photos (average accuracy level of 14.13) than 175 Spaniards (average accuracy level of 13).