# Digital Footprinting: Uncovering Tourists with User-Generated Content

**Article** *in* IEEE Pervasive Computing · January 2009

DOI: 10.1109/MPRV.2008.71 · Source: IEEE Xplore



**5 authors**, including:

Francesco Calabrese
IBM

Filippo Dal Fiore
University of Bologna

Josep Blat
University Pompeu Fabra

# Digital Footprinting: Uncovering Tourists with User-Generated Content


**Fabien Girardin**
**Josep Blat**
**Francesco Calabrese**
**Filippo Dal Fiore**
**Carlo Ratti**

# Digital Footprinting: Uncovering Tourists with User-Generated Content

**Keywords** digital footprinting, geovisualization, user-generated data, user-generated content

# Digital Footprinting:
## Uncovering Tourists with User-Generated Content

*Novel methods and tools are being developed to explore the significance of the new types of user-related spatiotemporal data. This approach helps uncover the presence and movements of tourists from cell phone network data and the georeferenced photos they generate.*

**Fabien Girardin and Josep Blat**
*Universitat Pompeu Fabra*

**Francesco Calabrese, Filippo Dal Fiore, and Carlo Ratti**
*Massachusetts Institute of Technology*

Today, it's possible to gather every click of every move of every user who interacts with any software in a database and submit it to a second-degree data-mining operation. Along with the growing ubiquity of mobile technologies, the logs produced have helped researchers create and define new methods of observing, recording, and analyzing a city and its human dynamics.[1] In effect, these personal devices create a vast, geographically aware sensor web[2] that accumulates tracks to reveal both individual and social behaviors with unprecedented detail.[3] The low cost and high availability of these *digital footprints* will challenge the social sciences, which have never before had access to the volumes of data used in the natural sciences,[4] but the benefits to fields that require an in-depth understanding of large group behavior could be equally great.

Accordingly, this article illustrates the potential of user-generated electronic trails to remotely reveal the presence and movement of a city's visitors. (This research was led at the Massachusetts of Technology, in collaboration with Universitat Pompeu Fabra.) We anticipate that validating these trails with respect to existing surveys might lead to an improved understanding of several aspects of urban mobility and travel. We therefore present several novel data-collection techniques, analytical methods, and visualization tools that we've been developing to uncover urban dynamics. Although the nature of digital footprints renders the information derived both more credible and reliable, we must further consider how to validate this pervasively user-generated content.

In a previous work, we showed that explicitly disclosed spatiotemporal data from open platforms provide novel insights on visitor dynamics in an urban space.[5] Understanding population dynamics by type, neighborhood, or region would enable customized services (and advertising) as well as the accurate timing of urban service provisions, such as scheduling monument opening times based on the daily, weekly, or monthly tourist demand. In general, more synchronous management of service infrastructures clearly could play an important role in tourism management.

## Working with digital footprints

A city's visitors have many ways of leaving voluntary or involuntary electronic trails: prior to their visits, tourists generate server log entries when they consult digital maps[6] or travel Web sites;[7] during their visit, they leave traces on wireless networks whenever they use their mo-


Published by the IEEE CS ■ 1536-1268/08/$25.00 © 2008 IEEE

bile phones;[8] and after their visit, they might add online reviews and photos. Broadly speaking then, there are two types of footprint: active and passive. *Passive tracks* are left through interaction with an infrastructure, such as a mobile phone network, that produces entries in locational logs; *active prints* come from the users themselves when they expose locational data in photos, messages, and sensor measurements.

In this article, we consider two types of digital traces from Rome, Italy: georeferenced photos made publicly available on the photo-sharing Web site Flickr and aggregate records of wireless network events generated by mobile phone users making calls and sending text messages on the Telecom Italia Mobile (TIM) system.

### Explicit footprints

People using the Flickr service to share and organize photos also have the option to add geographical attributes. Each time a photo is anchored to a physical location, Flickr assigns longitude and latitude values together with an accuracy attribute derived from the zoom level of the map in use to position the photos. So photos positioned on a map when the user zooms in at the street level receive a higher accuracy estimate than ones positioned when the user pulls back in the online map view. The system also adds metadata embedded by the camera into the image using exchangeable image file (EXIF) format information, completing the spatiotemporal data.

Flickr also provides a public API that enables anyone to query its public data store for photos. We elected to analyze three years of data, from November 2004 to November 2007, for Rome because it's a popular and highly photographed tourist destination. For that three-year period, we were able to extract 144,501 georeferenced photos that 6,019 different users had uploaded. For each of these publicly available photos, we retrieved the geographical coordinates, time stamp, accuracy level, and the owner's obfuscated identifier.

Because we were particularly interested in the behavior of tourists in Rome, we separated the photographers into two groups based on their presence in the city over time. Discriminating between locals and visitors required dividing the study period into 30-day blocks. If the photographers took all their photos within a period of 30 days, the algorithm labeled them as a visitor, but if they uploaded photographs at intervals of more than 30 days, then it categorized them as a resident. From our population of 6,019 photographers, we classed 4,719 as one-time visitors.

To find out more about the nature of our photographers, we took advantage of a social function in Flickr that invites users to voluntarily provide additional information about themselves such as their city and country of residence. In some cases, because of spelling errors or user idiosyncrasies (such as using "The Big Apple" to mean New York City), we were forced to manually process the city or country information. However, after cleaning, we found that 59 percent of the users had disclosed meaningful origin information, broken into several main populations: 991 Italians, 1,171 other Europeans, 807 North Americans, 104 South Americans, 71 Asians, and 70 Australians and New Zealanders.

Mapping and visualization are critical first steps to interpreting and explaining digital footprints.

### Implicit footprints

Previous research has shown that the diffusion of mobile phones and the widespread coverage of mobile phone wireless networks in urban areas make these technologies interesting means to identify and track both groups and individuals.[9,10] Our collaboration with TIM took advantage of new a system called Lochness (Localizing and Handling Network Event Systems), which is a software platform that localizes and stores user-generated events as they occur on the mobile network. Calls in progress, short message service (SMS) transmissions, and call handovers are all captured through external probes that localize and collate incoming messages before transmitting the results to Lochness. The messages are then aggregated to produce raster-format maps of the users' distribution. (A detailed introduction to the platform is available elsewhere.[11])

TIM installed the Lochness platform and related probes on a set of base station controllers (BSCs) located in the city's northeast quadrant, covering an area of approximately 100 km$^2$. The system can reliably localize users to within an area of 250 by 250 m$^2$ before assigning them to the corresponding grid reference. Lochness divided the users into two groups (Italians and foreigners) based on the country code information embedded in their international mobile subscriber identity (IMSI) number. Over a period of three months, timed to coincide with the Venice Biennale from September to November 2006, the system calculated these attributes every five minutes and transmitted the results to servers at the Massachusetts Institute of Technology (MIT).

### Processing and visualization

Jo Woo and his colleagues suggest that the large volumes of data coming from these types of sources can only be interpreted through geovisualization,[12] which is to say that after collection, mapping and visualization are critical to interpreting and explaining user

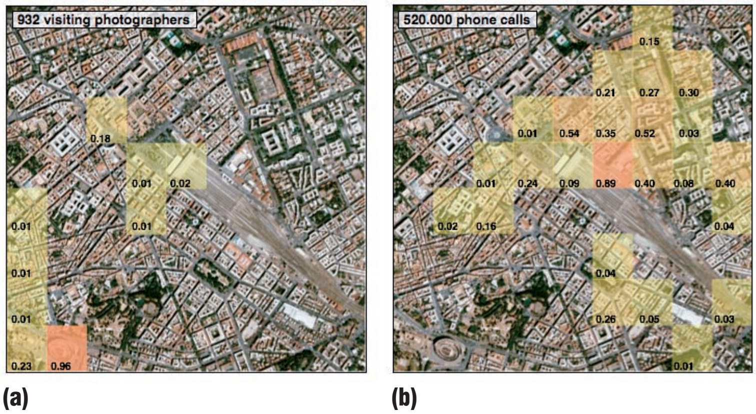



Figure 1. Geovisualizations of the presence of (a) 932 tourist photographers and (b) 520,000 phone calls from foreign mobile phones in the Coliseum and Piazza della Repubblica area from September to November 2006. Both types of data cover the train station area in the proximity of the Piazza della Republica.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

behaviors. We elected to use Google Earth to support visual synthesis and our preliminary investigation of digital traces. Accordingly, we stored data collected by the Lochness platform and the Flickr service on a MySQL server, enabling us to flexibly query and aggregate the data further as required. Using software developed in house, we then exported the aggregate results in a format compatible with Google Earth for interactive visual exploration. Precise digital satellite imagery from Telespazio, which is a company providing satellite services, was added as image overlay. Applying these techniques and tools to process digital footprints lets us uncover the presence of crowds and the patterns of movement over time as well as compare user behaviors to generate new hypotheses.

## Analyzing digital footprints

We used user-originated digital footprints to uncover some new aspects of the presence and movement of tourists during their visit to Rome. Specifically, we used spatial and temporal presence data to visualize the user-generated information.

### Spatial presence

To map users' spatial distribution, we store data in a matrix covering the entire 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. The geovisualization in Figure 1 reveals the main areas of tourist activity in part of central Rome over the three-month period of September to November 2006.

Figure 1a shows the presence of photographers, and Figure 1b depicts the areas of heavy mobile phone usage by foreigners. The union between visiting photographers and foreign mobile phone customers quickly uncovers the area's major visitor attractions such as the Coliseum and the main train station next to Piazza della Repubblica. It appears that the Coliseum attracts sightseeing photographers whereas foreign mobile phone users, typically on the move, tend to be active around the train station.

### Temporal presence

Turning to the temporal patterns obtained from the digital traces, we compared the number of photographers and the volume of phone activity for each day over the three-month study period. Figure 2 shows the difference between the average weekly distribution of phone calls made by visitors and the presence of visiting photographers in the areas around the Coliseum and Piazza della Repubblica. The histograms show the normalized variation between the average number of calls and photographs for each day and the average amount for the whole week.

The resulting temporal signatures for the Coliseum area show related trends for both data sets, with higher activity over the weekend than on weekdays. However, the Piazza della Repubblica area reveals a markedly different pattern: photographers, though fewer in number than at the Coliseum, also tend to be active on the weekend, whereas the foreign mobile phone users are much more active during the weekdays.

These temporal signatures provide further evidence to the different types of presence that occur at the tourist points of interest. We can further hypothesize that the Coliseum attracts sightseeing activities (photographers) over the weekend and the train station neighborhood provides facilities for visitors on the move (such as people on business trips) during the weekdays.

### Desire lines from digital traces

The study of digital footprints also lets us uncover the digital *desire lines*, which embody people's paths through the city. Based on the time stamp and location of photos, our software organizes the images chronologically to reconstruct the photographers' movement. More precisely, we start by revealing the most active areas obtained by spatial data clustering. Next, we aggregate these individual paths to generate desire lines that capture the sequential preferences of visitors. We check the location of each user activity (photo) to see if it's contained in a cluster and, in the case of a match, add the point to the trace generated by the

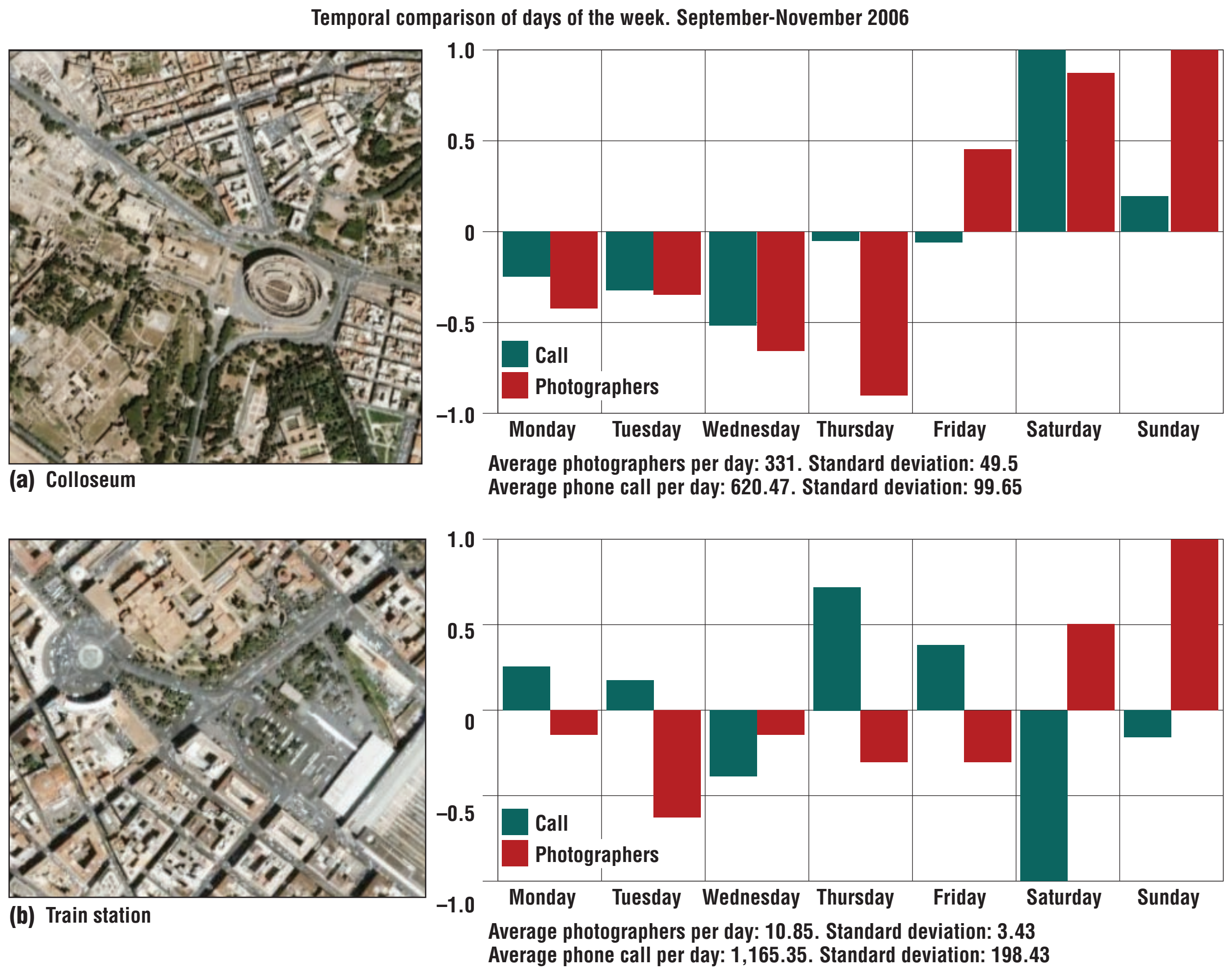



**Figure 2. Comparison of the temporal signature of foreigners' phone activity and number of tourist photographers. It reveals patterns of below-average activity on weekdays and a rise of presence over the weekend at the Coliseum. In contrast, the train station's temporal signature shows a higher presence of foreigners calling from their mobile phones during the week, whereas photographers indicate a reverse pattern and increased presence over the weekend.**

photo's owner. This process produces multiple directed graphs that support better quantitative analysis, which gives us the number of sites visited by season, the most visited and photographed points of interests, and data on where photographers start and end their journeys.

Formatting this data according to the open Keyhole Markup Language standard lets us import it into Google Earth to explore the traveling behaviors of specific types of visitors. The resulting visualization in Figure 3 suggests the main points of interest in the city as a whole. Building asymmetric matrices of the number of photographers who moved from point of interest $x$ to point of interest $y$ reveals the predominant sequence of site visits. We can also base queries on the users' nationality, the number of days of activity in the city, the number of photos taken, and areas visited during a trip.

### Semantic description

Previous work has demonstrated that we can use spatially and temporally annotated material available on the Web to extract place- and event-related semantic information.[13] In a similar vein, we analyzed the tags associated with the user-originating photos to reveal clues of people's perception of their environment and the semantics of their perspective of urban space. For instance, the word "ruins" is one of the most-used tags to describe photos in Rome. Mapping the distribution of this tag for 2,866 photos uncovers the most ancient and "decayed" part of the city: the Coliseum and the Forum (Figure 4).

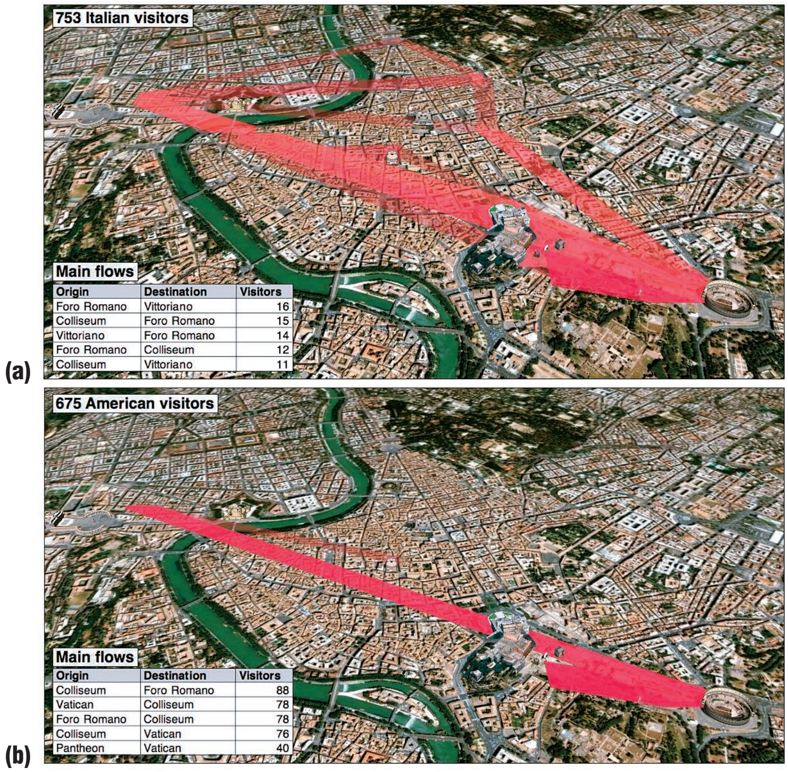



**Figure 3. Geovisualiation of the main paths taken by photographers between points of interests in Rome. Significantly, (a) the 753 visiting Italian photographers are active across many areas of the city, whereas (b) the 675 American visitors stay on a narrow path between the Vatican, Forum, and Coliseum. (Different scales apply for each geovisualization.)**

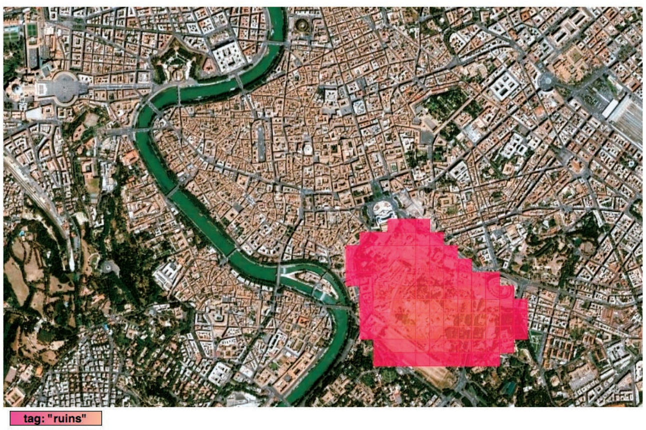



**Figure 4. Geovisualization of the areas defined by the position of the 2,886 photos with the tag "ruins" as uploaded by 260 photographers. It reveals the Coliseum and Forum areas known for their multitude ancient ruins.**

## Significance of user-generated data

These aggregate spatiotemporal records seem to lead to an improved understanding of different aspects of mobility and travel. Although the results are still fairly coarse, we've shown the potential for geographically referenced digital footprints to reveal patterns of mobility and preference among different visitor groups. However, in the context of our study, traditional methods would help us better define the usefulness of pervasive user-generated content. For example, hotel occupancy and museum surveys would let us observe and quantify visitors' presence and movements. Along this vein, the Rome tourism office supplied us with monthly ticket receipts for the Coliseum in 2006.

Figure 5 compares sales figures with mobile usage and photographic activity. Ticket receipts show that there are slightly more Coliseum visitors in October than September, with a major drop in attendance in November. This pattern matches the activity of foreign-registered mobile phones in the area, but it doesn't coincide with photographer activity. These discrepancies likely exist because the data sets are capturing the activity of different sets of visitors. For example, correlation with ticket sales from the Coliseum fails to account for the fact that users can easily photograph the arena or make a call from the vicinity of the monument without bothering to pay the entry fee. Due to the large difference in the nature of the activity producing the data, it might be that correlating it with user-generated content doesn't reinforce existing tourism and travel knowledge, but does reveal new dimensions of user behavior.

## Challenges of user-generated data sets

Our data-processing techniques have tried to account for the fluctuating

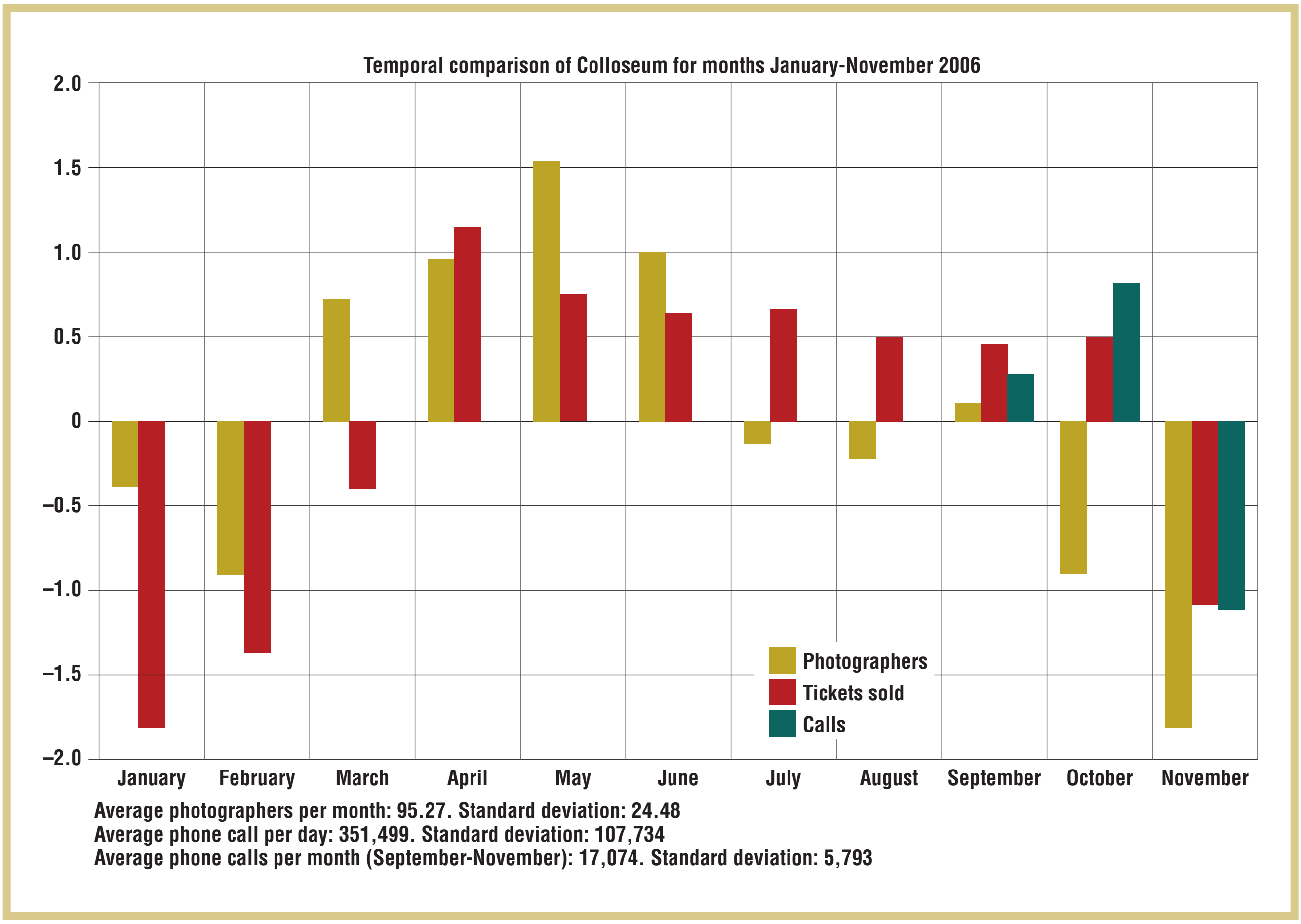



**Figure 5. Comparison of visitors to the Coliseum area between January and November 2006 using the number of tickets sold, number of calls made, and number of photographers active in the zone. The values represent the variations from the monthly average, scaled by the standard deviation.**

quality of user-generated data, which can substantially impede our ability to generate accurate information. For instance, the time stamps extracted from the camera-generated EXIF metadata don't necessarily match the real time at which a photo was taken; the user must take the time to set the camera's clock and then adjust it to the local time. User-generated data points can also be idiosyncratic and, for instance, indicate not the point where photo was taken but the location of the photographed object.

Including mobile phone data introduces challenging scale issues because camera phone and photo data set resolutions vary substantially. Challenges also arise from the fact that only phone activity handled by a subset of BSCs in Rome were monitored, leading to the risk of *border effects*, where calls near the border of a monitored area might be handled by other BSCs and thus not counted by the Lochness platform. This last consideration applies in particular to areas to the southwest of the Coliseum (see Figure 1).

We can also expect the use of photograph and mobile data to raise privacy and ethical concerns related to collecting data without the individual's consent. However, our approach addresses these concerns on two levels. First, our photography data set includes only information that users explicitly disclose on an open platform. Second, we aggregate the cell phone data in a way that removes all traces of the individual. On the Flickr service, users have direct control over who can access their locational data, but we supplemented this by applying an obfuscation algorithm to erase the relationship with individuals' Web identity and their digital trails. Thus, we could only analyze anonymous records of information they already publicly disclosed.

Collecting and analyzing aggregate network usage data fully complies with the 2002 Directive of the European Parliament and Council on privacy. Data was only reported to us in aggregate, so we received no data about an individual's identity or trajectory. In effect, we could only count the total number of people—either Italian or foreign—that used a mobile phone at a given point in the city and at a given moment. We couldn't identify individual users in any way based on the data that we collected and analyzed, and consequently we avoided the significant privacy issues that have other methodologies have raised.[10]

From a methodological perspective, the data we analyzed in this article has a clear advantage over more traditional location data obtained through controlled studies in which subjects carried sensors and were thus aware of being tracked. Although we couldn't determine the sample used, our mobile phone data covers the usage habits of more than 1 million people and thus represents a step-change in the scale of localizable data-collection efforts.

These collection methods also contain several important potential advantages over other pervasive tracking systems. Solutions that require people to carry a separate GPS-enabled device not only remind users that their movements are being followed—which might encourage them to pursue highbrow activities during their visit—but also generate fatigue effects and don't always function well in urban areas because of signal multipath and urban canyon obstructions. The alternative of a distributed, but fixed web of sensors entails onerous maintenance and data-transmission costs. These issues strongly suggest that the research community should investigate and evaluate the use of these new data types as well as consider approaches that don't rely on the deployment of ad hoc and costly infrastructures.

Our results demonstrate that further development is required to validate our observations and gain new insights into factors such as a space's temporal usage-signature, its attractiveness to different groups of people, and the degree of similarity to usage of other spaces.

The explicit character of photo geotagging and manual disclosure provides additional dimensions of interest; positioning a photo on a map isn't simply adding information about its location, it's an act of communication that embodies locations, times, and experiences that individuals consider to be relevant to themselves and others. There is a real richness to the intentional weight that people attach to disclosing their photos, and the results clearly show that Flickr users have a tendency to point out the highlights of their visit while skipping over their trip's lowlights.

Still, our analysis and visualization are meant to complement, not replace, traditional surveys and other means of data collection. In the predigital age, tourism officials could know how many visitors spent a night in a hotel, but now we can also use feedback mechanisms on public Web sites to estimate how much they enjoyed their stay. Similarly, we could know how many tourists visited a given attraction, but now we can also infer their experiences by their uploaded pictures and semantic descriptions. Direct observation let us know the number of tourists in an area, but through the mobile phone network, we can know their nationalities.

The shortcomings of single-site ticket sales as a correlating data set requires us to pursue alternate strategies for relating our mobile and photographic data to real-world activity with traditional surveys. An additional research avenue is understanding the circumstances under which users tag their content with at a street level or when they're tagged to a larger region. An initial analysis of our Flickr data set suggests that the 123 German users tended to provide more accurate locational information than their 175 Spanish counterparts.

The results of further analysis might reveal distinct profiles of georeferencing and geotagging photos. These profiles might be based on culture or nationality, the type of tourist in terms of their length of stay or familiarity with the city, their level of technical expertise or spatial orientation ability, and the type of task or type of environment visited. Other questions we need to consider relate to the types of situations during which users are more or less likely to use their mobile devices for data generation. Answers to these types of questions should help us better define the meaning of the data and further explore their potential usage in social sciences and urban studies.

Our analysis and visualization are meant to complement, not replace, traditional surveys and data collection.

## ACKNOWLEDGMENTS

We thank Barcelona Media and Telecom Italia for their support and Telespazio for satellite digital imagery. Also, we're 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, Jon Reades, and Andrea Vaccari for letting us pick their brains. Of course, any shortcomings are our sole responsibility.

## REFERENCES

1. E. O'Neill et al., "Instrumenting the City: Developing Methods for Observing and Understanding the Digital Cityscape," *Proc. 8th Int'l Conf. Ubiquitous Computing* (Ubicomp 2006), LNCS 4206, Springer, 2006, pp. 315–332.
2. M.F. Goodchild, "Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0," *Int'l J. Spatial Data Infrastructures Research*, vol. 2, 2007, pp. 24–32.

3. N. Eagle and A.S. Pentland, “Reality Mining: Sensing Complex Social Systems,” *Personal Ubiquitous Computing*, vol. 10, no. 4, 2006, pp. 255–268.

4. B. Latour, “Beware, Your Imagination Leaves Digital Traces,” *Times Higher Education Supplement*, 6 Apr. 2007; www.bruno-latour.fr/poparticles/poparticle/P-129-THES-GB.doc.

5. F. Girardin et al., “Leveraging Explicitly Disclosed Location Information to Understand Tourist Dynamics: A Case Study,” *J. Location-Based Services*, vol. 2, Mar. 2008, pp. 41–54.

6. D. Fisher, “Hotmap: Looking at Geographic Attention,” *IEEE Trans. Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1184–1191.

7. K. Wöber, “Similarities in Information Search of City Break Travelers — A Web Usage Mining Exercise,” *Information and Communication Technologies in Tourism 2007*, Springer Vienna, 2007, pp. 77–86.

8. R. Ahas et al., “Mobile Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia,” *Information and Communication Technologies in Tourism 2007*, Springer Vienna, 2007.

9. C. Ratti et al., “Mobile Landscapes: Using Location Data from Cell-Phones for Urban Analysis,” *Environment and Planning B: Planning and Design*, vol. 33, no. 5, 2006, pp. 727–748.

10. M. González, C. Hidalgo, and A Barabási, “Understanding Individual Human Mobility Patterns,” *Nature*, vol. 453, 2008, pp.779–782

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

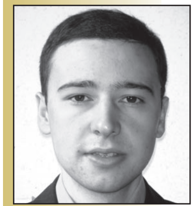
12. J. Woo et al., “Interactive Visual Exploration of a Large Spatio-Temporal Dataset: Reflections on a Geovisualization Mashup,” *IEEE Trans. Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1176–1183.

13. T. Rattenbury, N. Good, and M. Naaman, “Towards Automatic Extraction of Event and Place Semantics from Flickr Tags,” *Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 07), ACM Press, 2007, pp. 103–110.

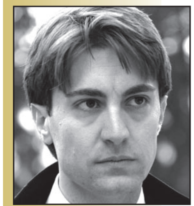



## the AUTHORS


**Fabien Girardin** is a PhD student in computer sciences and digital communication at the Universitat Pompeu Fabra, Spain, and a visiting PhD student at the Massachusetts Institute of Technology’s SENSEable City Laboratory. He is also a cofounding partner of Simpliquity, a user experience consultancy in the domain of emerging technologies. His research interests include integrating ubiquitous technologies in the everyday urban environments. Girardin received his MPhil in computer science and digitial communication from the Universitat Pompeu Fabra. Contact him at Fabien.Girardin@upf.edu



**Francesco Calabrese** is a post-doctoral associate at the Massachusetts Institute of Technology’s SENSEable City Laboratory. His research interests include ubiquitous computing; analysis of urban dynamics through sensor networks; analysis and design of distributed, hybrid, embedded control systems; and CNC machines. He received his PhD in computer engineering from the University of Naples Federico II, Italy. He is a member of the IEEE and the IEEE Control Systems Society. Contact him at fcalabre@mit.edu.



**Filippo Dal Fiore** is a researcher affiliated with the Massachusetts Institute of Technology’s SENSEable City Laboratory and the Austrian Academy of Science, while a PhD candidate in applied economics at the Vrije Universiteit Amsterdam. His research interests include the behavioral implications of information and mobile technologies, encompassing topics in different disciplines, from mobility and transportation to education and innovation. Contact him at dalfiore@mit.edu.

**Carlo Ratti** is an associate professor of the practice of urban technologies at the Massachusetts Institute of Technology, where he directs the SENSEable City Laboratory. He’s also founding partner and director of carlorattiassociati, an architectural firm. He received his PhD in architecture from the University of Cambridge. He’s a member of the Ordine degli Ingegneri di Torino and the Association des Anciens Elèves de l’École Nationale des Ponts et Chaussées and is a UK Registered Architect. Contact him at the ratti@media.mit.edu.

**Josep Blat** is a professor of computer science and director of the Department of Information and Communication Technologies at Universitat Pompeu Fabra, Spain. He received a PhD in Mathematics at the Heriot-Watt University (Edinburgh). His research interests include cooperative environments, educational telematics, multimedia and GIS. Contact him at Josep.Blat@upf.edu