Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population

A case study for Lisbon Metro using *foursquare* check-in data

Christoph AUBRECHT[1*]; Joachim UNGAR[1]; Sergio FREIRE[2]
[1]AIT Austrian Institute of Technology GmbH
Donau-City-Str. 1, A-1220 Vienna, Austria
+43-50550-4522, christoph.aubrecht@ait.ac.at, joachim.ungar@gmail.com (corr. author*)
[2]e-GEO, FCSH, New University of Lisbon
Avenida de Berna 26-C, 1069-061 Lisbon, Portugal
sfreire@fcsh.unl.pt

Keywords: VGI, LBSN, foursquare, spatio-temporal, urban population distribution

Introduction

In recent years we have observed an incredible increase in location-specific information provided voluntarily by individuals and disseminated via the internet. The emergence of this Volunteered Geographic Information (VGI) as Goodchild first described it in 2007 [1] has attracted considerable interest within the GIScience research community. As a special type of user generated content, it offers great potential to produce up-to-date and near real-time information related to any place on Earth, even though overall accuracy remains an issue of debate. Location sharing services (LSS) such as ‘foursquare’, ‘Gowalla’, and ‘Facebook Places’ collect hundreds of millions of user-driven footprints or ‘check-ins’. Those footprints provide a unique opportunity to (1) study social and temporal characteristics of how people use these services and (2) model patterns of human mobility. However, the amount and frequency of VGI is not evenly distributed and recent research considers it directly related to socioeconomic characteristics of its contributors (i.e., geographic and economic constraints, individual social status) [2-3].
Particularly in the context of population dynamics studies, VGI may provide a data source that is more accessible and current as well as less expensive and time-consuming than traditional activity survey data. VGI generated on micro-blogging services and location-based social networks (LBSN) bear the greatest resemblance to the activity diary that time geographers are familiar with [4]. Noulas et al. [5] present a large-scale study of user behavior on the LBSN platform ‘foursquare’, analyzing user check-in dynamics and demonstrating how that reveals meaningful spatio-temporal patterns and offers the opportunity to study both user mobility and characteristics of urban spaces.
In this study we compare functionally categorized location-specific foursquare check-in information picturing one working week in the Lisbon Metropolitan Area to a daytime working population surface produced in previous work. The objective

is to analyze potential correlation patterns and explore options for modeling fine-scale spatio-temporal characteristics of urban land use based on VGI.

Study area and data

Lisbon Metro and its daytime working population

The study area comprises the main urban part of the Lisbon Metropolitan Area (LMA), located on the north side of the large Tagus River estuary. The LMA is the main metropolitan region in Portugal accounting for 36% of the national GDP and hosting 30% of the country's companies. The region is home to more than 2.5 million residents, i.e. 26% of the country's total population [6]. Although the average population density is 898 inhabitants per square kilometer, these densities vary widely in space and time. Beyond the more urbanized core the region still includes vast rural areas with scattered settlements whose uneven population density is not well captured and represented by census polygons, which can be heterogeneous even at the block level. Also, due to daily commuting for work and study, the daytime population of municipalities in the metro area of Lisbon can differ by more than 50% of the residential figures from the census [7].
In previous work [8] the population distribution of the LMA was modeled on very high spatial level of detail based on raster dasymetric mapping [9], i.e. spatially disaggregating census information from 2001 using residential streets as spatial reference units to reallocate population counts. The varying spatio-temporal characteristics were considered by referring to population mobility statistics in order to map daytime distribution. The final total daytime population distribution grid results from the sum of two raster surfaces at 50 m resolution: (1) the daytime working population, i.e. showing people in their places of work or study, and (2) the daytime residential population, i.e. the population that remains home during the day.

Location specific foursquare check-in data

'Foursquare' is a location-based social network (LBSN) relying on the growing usage of GPS-enabled mobile devices. Sharing their location with friends, users 'check in' at given venues to collect user points and virtual 'badges'. Earning badges stands for increasing the user's social status in the network. Users are also able to add new venues and thus extend the database, which in a further step can be verified by the respective venue owner to assure data quality. Foursquare currently counts over 8 million users worldwide, checking in around 2.5 million times per day [10].
Due to privacy policies it is not possible to access the raw data individually. However, there is an application programming interface (API) provided for companies and mobile application developers enabling retrieval of certain restricted data views. For this study the 'foursquare venues project (beta)' API endpoint was used (1) to extract the locations and types of venues and (2) to get the number of users currently checked in at these locations. With venues also containing functional information (e.g., office, restaurant, gym) it is possible to categorize the types of user activities. This can support in-depth analysis of time-dependent user behavior which would not be possible with other popular social network data like Twitter.

From 8-15 May 2011 (Sunday till Sunday) the entire area covered by approx. 1,400 request points (i.e., each identifying the respective 30 nearest venues) was observed on an hourly basis. In total, more than 250,000 requests were made, eventually resulting in a PostGIS database containing 10,185 venues and 22,664 check-ins.

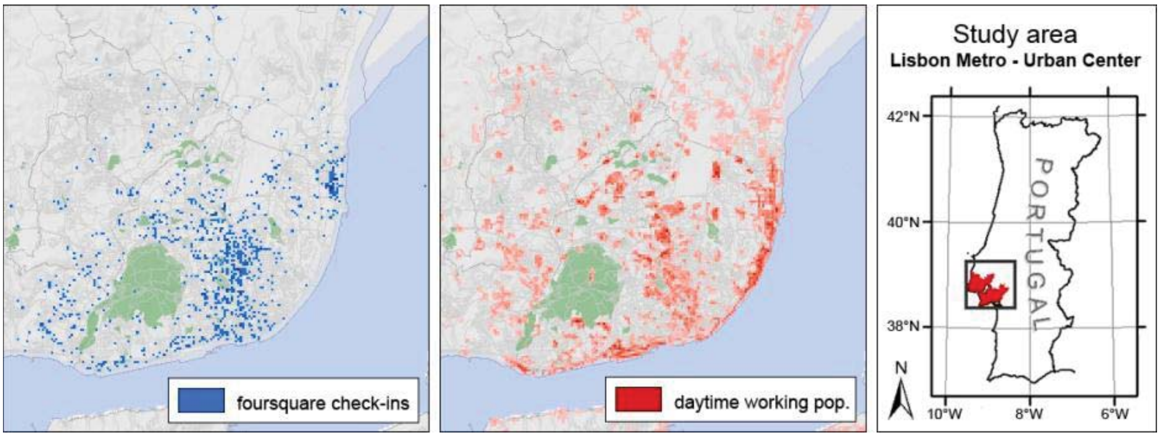

Figure 1: Comparison of 100 m foursquare check-in density raster (working week) and 100 m modeled daytime working population grid for the urban center of Lisbon Metro

Methods

To make the recorded foursquare data comparable to the disaggregated working population raster (see figure 1) some requirements had to be defined: (1) Covering a typical workday, only user activity between 9 a.m. and 5 p.m. is considered. (2) Only user activity between Monday and Friday is taken into account (working week). (3) Only those foursquare venue categories relevant for daytime working population are used for the analysis (e.g., office, food).

Rasterizing the foursquare data to a 50 m cell size (i.e., resolution of the reference dataset) showed little useful spatial patterns and a rather small area covered. In order to cope with this problem the spatial resolution was decreased and two raster surfaces of 100 m and 200 m respectively were produced. In an additional step the previously modeled daytime working population grid is aggregated accordingly and overlaid with the foursquare grids in order to apply some spatial statistics and correlation measures. Not the absolute numbers are checked in that context, but rather relative density patterns. Bearing in mind that the two data sets do not refer to the same point in time (2001 and 2011 respectively), we assume checking spatially explicit but nonetheless relative patterns a worthwhile approach. In fact, the subjects of comparison in that context are not two population distribution models, but rather location-specific online activity vs. human mobility patterns.

Discussion and outlook

Particularly in urban areas an increasing number of persons are equipped with ‘location sensors’ in the form of GPS-enabled mobile devices. The willingness to share situational experiences with others is generally increasing rapidly and is boosted by rising new technologies supporting the spatial component of social networks. These new developments result in collection of a vast amount of data about people's locations and enable analyses of spatio-temporal movements. With further research certainly needed, this kind of data has a high potential to be very useful in the future for urban planning purposes, disaster and crisis management, and for other fields requiring population data on high spatio-temporal resolution.

Undoubtedly the data in its current form still show gaps and are biased to some extent by various factors: (1) To date there has not been any quantitative information available on the socioeconomic structure of foursquare users. This is most welcome for privacy reasons but complicates the assessment of how representative the data is with respect to the total population. (2) Foursquare users usually use the service on a regular basis which leads to redundant records when aggregating data from multiple days or weeks. (3) The motivation of users to check in varies depending on a multitude of factors including their general social behavior, activities (e.g., special events), and rewards like badges or free items (e.g., coffee, ice cream) they receive from foursquare or venue owners. Therefore, the captured information often covers just a small sample of the total user mobility and activity in real life.

References

[1] **Goodchild, M.F.** (2007), Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), pp. 211-221

[2] **Cheng, Z. et al.** (2011), Exploring Millions of Footprints in Location Sharing Services. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11). Barcelona, Spain

[3] **Li, L. and Goodchild, M.F.** (2011), The spatio-temporal patterns of VGI. Symposium on Space-Time Integration in Geography and GIScience. AAG Annual Meeting 2011. Seattle, WA, USA

[4] **Rush, J. and Kwan, M.-P.** (2011), Using Volunteered Geographic Information to Study Population Dynamics. Symposium on Space-Time Integration in Geography and GIScience. AAG Annual Meeting 2011. Seattle, WA, USA

[5] **Noulas, A. et al.** (2011), An Empirical Study of Geographic User Activity Patterns in Foursquare. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11). Barcelona, Spain

[6] **INE (Instituto Nacional de Estatística)** (2001), Recenseamento Geral da População e da Habitação, Lisbon

[7] **INE (Instituto Nacional de Estatística)** (2003), Movimentos Pendulares e Organização do Território Metropolitano: Área Metropolitana de Lisboa e Área Metropolitana do Porto 1991-2001, Lisbon

[8] **Freire, S., and Aubrecht, C.** (2011), Assessing Spatio-Temporal Population Exposure to Tsunami Hazard in the Lisbon Metropolitan Area. 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011). Lisbon, Portugal

[9] **Mennis, J. and Hultgren, T.** (2006), Intelligent dasymetric mapping and its application to areal interpolation. Cartography and Geographic Information Science, 33(3), pp. 179-194

[10] **Foursquare** website https://foursquare.com/about (last visited: 23 May 2011)