Online data reliability for monitoring tourism activities in cities

Luis Encalada[1,2]
[1]Universidade de Lisboa
Institute of Geography and Spatial Planning
Branca Edmée Marques, 1600-276
Lisbon, Portugal
luisencalada@campus.ul.pt

[2]Universidad Espíritu Santo
Km. 2.5 vía La Puntilla,
Guayaquil, Ecuador

Jorge Rocha
Universidade de Lisboa
Institute of Geography and Spatial Planning
Branca Edmée Marques, 1600-276
Lisbon, Portugal
jorge.rocha@campus.ul.pt

Carlos C. Ferreira
University of Coimbra
Faculty of Arts and Humanities
Colégio de S. Jerónimo, 3004-530
Coimbra, Portugal
carlos.ferreira@uc.pt

Abstract

Studying the spatial structure of destinations has long been considered an important topic on tourism research. Researchers focus their attention on the analysis of tourism consumption and production. Recent studies found that non-traditional data, either from social media or peer-to-peer digital platforms, is useful for managing and monitoring tourism activities on dynamic destinations such as cities. However, more research is needed to address data reliability since it is important to understand the strengths and limitations of this novel data. This paper compares visitor counts to Lisbon city identified by their digital footprints on social media (i.e., visitors' geotagged photos from Flickr) with official hotel occupancy rates, as well as the accommodation supply from Airbnb listings with local accommodation official statistics. Findings demonstrated that the number of visitors identified by their digital footprints matches relatively well the visitation pattern reproduced by hotel occupancy rates. Datasets representing the accommodation supply uncovered the spatial concentration of local lodgings, not exclusively in the well-known tourist areas. Despite the strong correlation between datasets, there are some areas where the lodging counts deviate significantly, as we move from the periphery to tourism cores. Results showed the value of data from social media and peer-to-peer digital platforms as proxy for monitoring city tourism activities.

Keywords: Urban tourism; Social media data; Geotagged photos; Airbnb listings; Spatial analysis; Correlation analysis.

1 Introduction

Urban tourism has undergone a huge growth, becoming an important activity in most cities, where pressure from tourism is particularly intense, though not exclusively, in core areas known as tourism districts. Non-locals do intensive use of facilities and services, including transportation and accommodation, available in those areas. Studying the spatial structure of destinations has long been considered an important topic on tourism research, with implication on city management and planning (Pearce, 2001, 2013; Ashworth and Page, 2011). Researchers focus their attention on the analysis of tourism consumption and production. For instance, some studies monitor visitors' itineraries to analyse their spatial behaviour, while others explore the growth of tourism services (e.g., accommodation) to study the changes of functional use of some urban areas.

As denoted by some authors (Pearce, 2001; Ashworth and Page, 2011) the study of urban tourism is sometimes limited by the lack of suitable data. Traditional data collection techniques (e.g. surveys, interviews, counters) are often laborious, time-consuming and costly, also deprived of a longitudinal temporal facet. On the other hand, official data sources do not provide detailed information since there are some limitations about the spatial and temporal resolution of data (Batista e Silva *et al.*, 2018). However, after entering the digital age, the popularity of location-aware devices, the intensification of user-online activity (in social media), and the rise of peer-to-peer digital platforms (e.g., Airbnb, Tripadvisor, etc.) have made it possible to access information about visitor's behaviour and their context. Big data offers a potential for innovative statistics (Daas *et al.*, 2015), regarded as a self-sufficient source or by complementing official data or data collected through well-known classic methods. It is innovating the way various agencies, either state owned institutions or private companies, yield economic, social and demographic statistics. Big data is fostering the engagement between qualitative-quantitative approaches and its mutual benefits rather than advocate the convenience of one single approach (Sui and DeLyser, 2011).

The big data deluge has been associated with a production of papers, while promoting its key role in tourism research (Li *et al.*, 2018). Recent attempts (Straumann, Çöltekin and Andrienko, 2014; García-Palomares, Gutiérrez and Mínguez, 2015; Vu *et al.*, 2015, 2017; Li, Zhou and Wang, 2018) explored the attractiveness of places, tourist' areas of concentration and intra-destination mobility, by leveraging geotagged data from social networks. Accommodation supply data from Airbnb led the analysis of critical issues associated to tourism impacts in urban destinations (Guttentag, 2015; Gutiérrez *et al.*, 2017; Eugenio-Martin, Cazorla-Artiles and González-Martel, 2019). Still, the analysis of information from TripAdvisor can also give an indication on how tourists perceive, experience and use the destination (Batista e Silva *et al.*, 2018; van der Zee, Bertocchi and Vanneste, 2018).

All studies found that non-traditional data either from social media and peer-to-peer digital platforms is useful for managing and monitoring tourism activities on dynamic destinations such as cities. However, more research is needed to address (big)data reliability (Li *et al.*, 2018), since there can exist spatial and temporal biases. It is important to understand the strengths and limitations of data and only few studies have used more controlled data sources (e.g., surveys or official statistics) to validate social media data.

This paper compares visitor counts to Lisbon city identified by their digital footprints on social media (i.e., visitors' geotagged photos from Flickr) with hotel occupancy rates from the city Tourism Bureau (Observatório de Turismo de Lisboa –OTL), as well as the accommodation supply from Airbnb listings with local accommodation statistics from the National Local Lodging Registry (Registo Nacional de Alojamento Local – RNAL). The aim of this paper is to identify spatial and temporal (dis)similarities between authoritative and online data regarding tourist visitation and services. We focus our analysis on several years, using visitors' information between 2012 and 2017, and lodgings information from 2015 to 2018.

2 Data

Geotagged photos available online on Flickr within the city boundary were retrieved through its Application Programming Interface. As in previous works (Girardin *et al.*, 2008; García-Palomares, Gutiérrez and Mínguez, 2015), metadata related to photos' timestamps (the date when the picture was created) yielded the selection of pictures taken by visitors. The number of days between the first and last uploaded pictures was computed to identify users that do not overpass the average length of stay (3 days) within the destination. These pictures were considered to belong to visitors, otherwise they were considered as belonging to locals. The dataset comprises >69,400 photos belonging to more than 6,700 users considered as city tourists (Table 1).

Hotel occupancy rates were retrieved from OTL monthly reports (https://www.visitlisboa.com/about-turismo-de-lisboa/observatório).

Table 1: Yearly counts of Lisbon visitors from Flickr.

Year	Visitors' photos	Visitors (Flickr)	Visitors - Monthly Mean/Std. Dev.
2012	9,543	1,232	103/22
2013	11,434	1,250	104/26
2014	12,507	1,262	105/25
2015	14,096	1,178	98/16
2016	10,745	1,073	89/20
2017	11,083	793	66/23

Accommodation data refers to some information about Airbnb lodgings obtained from Inside Airbnb (http://insideairbnb.com/get-the-data.html) and Tom Slee websites (http://tomslee.net/category/airbnb-data). Both sites let the access to data packages containing public information compiled from Airbnb website, including lodging geolocation, room and host IDs, room type, number of accommodates (acc.), bedrooms, reviews, etc. This data was retrieved by above-mentioned contributors, from 2015 to 2018. We use the available data packages. In 2015, data from a single day in March. In 2016 from 4 days in March, June, September and December. In 2017, we use data from 8 days between January and July. And, in the last year, from 6 days between April and October, excluding June. Data compilation from contributors correspond to available lodgings in those days. Since the total number of rooms may vary according the daytime of data collection, we use all datasets about lodging availability in several days to have a more precise number resembling the existing accommodation supply. Data from different days was merged according to each year. Repeated hosting rooms IDs were removed so we count distinct rooms offered during the year. In order to compile each room information, we compare each room data from different days and select only the max number of accommodates and bedrooms available in the corresponding listings. Coordinates from repeated rooms were compared to verify whether they match, no matching rooms were also removed.

Data from RNAL refers to local lodgings registered and legally operating that provide temporary accommodation services. We segment lodgings by year with reference to their day of registration. All lodging registered from 2010 up to 2015 were considered for the first year of analysis, and so on for the remaining years (Table 2).

Table 2: Yearly counts of lodgings from Airbnb and RNAL in Lisbon.

Year	Airbnb lodgings	Sum (Max. acc.)	RNAL lodgings	Sum (# acc.)
2015	5,653	13,903	3,091	20,305
2016	15,165	57,185	6,159	36,151
2017	16,925	64,595	10,211	57,886
2018	19,260	73,609	16,696	93,005

The number of lodgings were spatially aggregated with reference to Lisbon city blocks.

3 Methods

3.1 Correlation analysis

For each year, we compare the similarity of monthly visitor counts from social media (Flickr) with hotel occupancy rates from OTL by using the Pearson correlation coefficient, which is a commonly used statistical method to measure the linear relationship between two datasets. We also present results from *F*-test between time-series data.

3.2 Geographically weighted regression (GWR)

We use the GWR (Fotheringham, Brunsdon and Charlton, 2002) to assess the degree of association between the aggregated number of lodgings from RNAL and Airbnb. GWR enables local variations (over space) in the estimation of coefficients of determination. We focus on the analysis of the standardised residuals to get an overview of the differences concerning the spatial distribution of local accommodation supply between both data sources. The analysis was performed for each of the 4 years.

4 Results and Discussion

4.1 Inbound visitors counts

There is a similar trend between the seasonal pattern of Flickr' users considered as visitors and the rates of hotel occupancy. Monthly counts from Flickr show the same peaks in the two semesters, following the tourism seasonality evidenced on hotel occupancy rates from OTL reports. In the last six years, there seems to be a significant positive correlation between the two time-series (Table 3). Although, this correlation was lower in 2012 and 2015 (when compared to other years), not showing a powerful significance at the level of $p < 0.01$, the tests for remaining years correlate strongly and highly significant.

Additionally, results from *F*-test and its low statistical significance values (Table 3), suggest that there is not enough evidence to reject the null hypothesis (H_0 - ratio of variances is equal to 1). Therefore, data distributions on both time-series, regarding monthly visitors' presence within the destination, show similar variances.

Table 3: Pearson correlation (*r*), *F*-test and significance level between monthly visitor counts from Flickr and monthly hotel occupancy rates, from 2012 to 2017.

Year	*r*	*F*	*F* significance (Level=0.05)
2012	0.69*	2.43	0.16
2013	0.81**	3.11	0.07
2014	0.86***	2.40	0.16
2015	0.64*	1.16	0.80
2016	0.89***	1.96	0.27
2017	0.71**	3.61	0.04

p-values: ≤ 0.05(*); ≤ 0.01(**); ≤ 0.001(***)

4.2 Comparison between local accomodation datasets

The coefficient of determination reveals the common part of variation between datasets (Table 4). Resulting adjusted R^2 denote a medium-high positive correlation between lodging counts from both sources. Still, the multi-year analysis reveals that the correlation is higher in the recent years.

Table 4: Adjusted R^2 from GWR.

Year	Adjusted R^2
2015	0.49
2016	0.51
2017	0.63
2018	0.73

By looking at GWR standardised residuals (Figure 1), it is possible to identify areas where there are significant differences between observed and expected values, represented by higher or lower residuals. In the 4 years, the accommodation supply located in the 'inner-city' is consistently represented in both datasets. There are no strong differences in most of the peripheral city blocks. Still, the amount of city blocks matching lodging counts in areas with intensive tourist activity is not numerous but still relevant.

Similarities are less marked near tourism cores. There is some evidence suggesting that there are more Airbnb lodgings than expected (on RNAL), most of them located in the proximity of tourism cores (city blocks with negative residuals – blue colored). Moreover, there are some city blocks within well-known tourist areas where lodging counts deviate significantly, like the ones located in Graça and Alfama neighborhoods.

5 Conclusions

This study examines the reliability of data from social media and peer-to-peer digital platforms regarding its temporal and spatial representativeness in comparison to official data and, therefore, its value as proxy for monitoring tourism activities in cities.

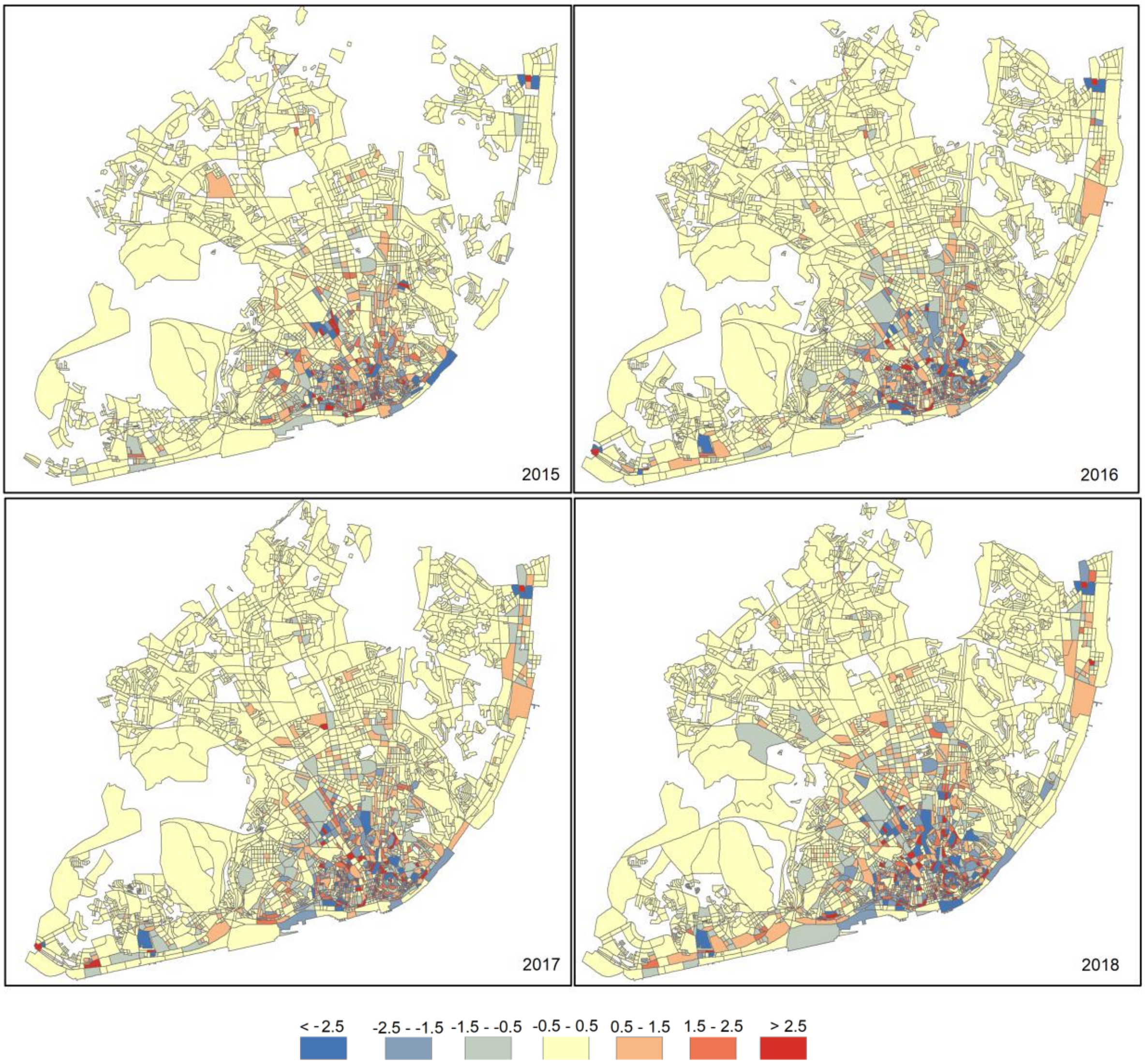
Our analysis revealed that the number of visitors identified by their digital footprints matches relatively well the visitation pattern (e.g. tourism seasonality) seen on hotel occupancy reports.

Datasets representing the accommodation supply uncovered the spatial concentration of local lodgings, not exclusively in the well-known tourist areas. Despite the strong correlation observed on the spatial distribution of lodgings, there are some dissimilarities between datasets, in terms of volume, more pronounced as we move from the periphery to tourism cores. Moreover, there is some evidence indicating that the current increase of lodgings may not be reflected consistently in official statistics, neither in consolidated nor in newer tourist areas.

Using online data available on social media and peer-to-peer digital platforms can be an effective way for monitoring tourist visitation and services. It provides new perspectives of understanding tourist behavior, as well as the processes related to city tourism production and consumption.

Our findings support previous works indicating that online data complement existing authoritative data. However, it should be taken with caution. As denoted in this paper, there is the need to evaluate non-traditional sources against other data sources to better know their strengths and limitations.

Figure 1: GWR residuals between local accommodation datasets of Lisbon city, from 2015 to 2018.



References

Ashworth, G. and Page, S. J. (2011) Urban tourism research: Recent progress and current paradoxes, *Tourism Management*, 32(1), 1–15.

Batista e Silva, F. *et al.* (2018) Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources, *Tourism Management*, 68, 101–115.

Daas, P. J. H. *et al.* (2015) Big Data as a Source for Official Statistics, *Journal of Official Statistics*, 249.

Eugenio-Martin, J. L., Cazorla-Artiles, J. M. and González-Martel, C. (2019) On the determinants of Airbnb location and its spatial distribution, *Tourism Economics*. SAGE Publications Ltd.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. E. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, England: Wiley.

García-Palomares, J. C., Gutiérrez, J. and Mínguez, C. (2015) Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS, *Applied Geography*, 63, 408–417.

Girardin, F. *et al.* (2008) Leveraging explicitly disclosed location information to understand tourist dynamics: a case study, *Journal of Location Based Services*. Taylor & Francis, 2(1), 41–56.

Gutiérrez, J. *et al.* (2017) The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona, *Tourism Management*, 62, 278-291

Guttentag, D. (2015) Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector, *Current Issues in Tourism*. Routledge, 18(12), 1192–1217.

Li, D., Zhou, X. and Wang, M. (2018) Analyzing and visualizing the spatial interactions between tourists and locals:

A Flickr study in ten US cities, *Cities*, 74, 249–258.

Li, J. *et al.* (2018) Big data in tourism research: A literature review, *Tourism Management*, 68, 301–323.

Pearce, D. G. (2001) An integrative framework for urban tourism research, *Annals of Tourism Research*, 28, 926–946.

Pearce, D. G. (2013) Toward an Integrative Conceptual Framework of Destinations, *Journal of Travel Research*. SAGE Publications Inc, 53(2), 141–153.

Straumann, R. K., Çöltekin, A. and Andrienko, G. (2014) Towards (Re)Constructing Narratives from Georeferenced Photographs through Visual Analytics, *The Cartographic Journal*. Taylor & Francis, 51(2), 152–165.

Sui, D. and DeLyser, D. (2011) Crossing the qualitative-quantitative chasm I: Hybrid geographies, the spatial turn, and volunteered geographic information (VGI), *Progress in Human Geography*. SAGE Publications Ltd, 36(1), 111–124.

Vu, H. Q. *et al.* (2015) Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos, *Tourism Management*, 46, 222–232.

Vu, H. Q. *et al.* (2017) Tourist Activity Analysis by Leveraging Mobile Social Media Data, *Journal of Travel Research*. SAGE Publications Inc, 57(7), 883–898.

van der Zee, E., Bertocchi, D. and Vanneste, D. (2018) Distribution of tourists within urban heritage destinations: a hot spot/cold spot analysis of TripAdvisor data as support for destination management, *Current Issues in Tourism*. Routledge, 1–22.