Bassolas et al. *EPJ Data Science* (2016) 5:12
DOI 10.1140/epjds/s13688-016-0073-5

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Touristic site attractiveness seen through Twitter

Aleix Bassolas[1], Maxime Lenormand[1*], Antònia Tugores[1], Bruno Gonçalves[2,3] and José J Ramasco[1]

*Correspondence:
maxime@ifisc.uib-csic.es
[1]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Palma de Mallorca, 07122, Spain
Full list of author information is available at the end of the article

Abstract

Tourism is becoming a significant contributor to medium and long range travels in an increasingly globalized world. Leisure traveling has an important impact on the local and global economy as well as on the environment. The study of touristic trips is thus raising a considerable interest. In this work, we apply a method to assess the attractiveness of 20 of the most popular touristic sites worldwide using geolocated tweets as a proxy for human mobility. We first rank the touristic sites based on the spatial distribution of the visitors' place of residence. The Taj Mahal, the Pisa Tower and the Eiffel Tower appear consistently in the top 5 in these rankings. We then pass to a coarser scale and classify the travelers by country of residence. Touristic site's visiting figures are then studied by country of residence showing that the Eiffel Tower, Times Square and the London Tower welcome the majority of the visitors of each country. Finally, we build a network linking sites whenever a user has been detected in more than one site. This allow us to unveil relations between touristic sites and find which ones are more tightly interconnected.

Keywords: tourism; human mobility; geolocated data; spatial network

1 Introduction

Traveling is getting more accessible in the present era of progressive globalization. It has never been easier to travel, resulting in a significant increase of the volume of leisure trips and tourists around the world (see, for instance, the statistics of the last UNWTO reports [1]). Over the last fifty years, this increasing importance of the economic, social and environmental impact of tourism on a region and its residents has led to a considerable number of studies in the so-called geography of tourism [2]. In particular, geographers and economists have attempted to understand the contribution of tourism to global and regional economy [3–7] and to assess the impact of tourism on local people [8–13].

These researches on tourism have traditionally relied on surveys and economic datasets, generally composed of small samples with a low spatio-temporal resolution. However, with the increasing availability of large databases generated by the use of geolocated information and communication technologies (ICT) devices such as mobile phones, credit or transport cards, the situation is now changing. Indeed, this flow of information has notably allowed researchers to study human mobility patterns at an unprecedented scale [14–19]. In addition, once these data are recorded, they can be aggregated in order to analyze the city's spatial structure and function [20–27] and they have also been successfully

tested against more traditional data sources [28–30]. In the field of tourism geography, these new data sources have offered the possibility to study tourism behavior at a very high spatio-temporal resolution [18, 31–37].

In this work, we propose a ranking of touristic sites worldwide based on their attractiveness measured with geolocated data as a proxy for human mobility. Many different rankings of most visited touristic sites exist but they are often based on the number of visitors, which does not really tell us much about their attractiveness at a global scale. Here we apply an alternative method proposed in [38] to measure the influence of cities. The purpose of this method is to analyze the influence and the attractiveness of a site based on the average radius traveled and the area covered by individuals visiting this site. More specifically, we select 20 out of the most popular touristic sites of the world and analyze their attractiveness using a dataset containing about 10 million geolocated tweets, which have already demonstrated their efficiency as useful source of data to study mobility at a world scale [18, 38]. In particular, we propose three rankings of the touristic sites' attractiveness based on the spatial distribution of the visitors' place of residence, we show that the Taj Mahal, the Pisa Tower and the Eiffel Tower appear always in the top 5. Then, we study the touristic site's visiting figures by country of residence, demonstrating that the Eiffel Tower, Times Square and the London Tower attract the majority of the visitors. To close the analysis, we focus on users detected in more than one site and explore the relationships between the 20 touristic sites by building a network of undirected trips between them.

2 Materials and methods

The purpose of this study is to measure the attractiveness of 20 touristic sites taking into account the spatial distribution of their visitors' places of residence. To do so, we analyze a database containing 9.6 million geolocated tweets worldwide posted in the period between September 10, 2010 and October 21, 2015. The dataset was built by selecting the geolocated tweets sent from the touristic places in the general streaming and requesting the time-lines of the users posting them. The touristic sites boundaries have been identified manually. Collective accounts and user exhibiting non-human behaviors have been removed from the data by identifying users tweeting too quickly from the same place, with more than 9 tweets during the same minute and from places separated in time and space by a distance larger than what is possible to be covered by a commercial flight (with an average speed of 750 km/h). Their spatial distributions and that of the touristic sites can be seen in Figure 1.

In order to measure the site attractiveness, we need to identify the place of residence of every user who have been at least once in one of the touristic sites. First, we discretize the space by dividing the world into squares of equal area (100×100 km^2) using a cylindrical equal-area projection. Then, we identify the place most frequented by a user as the cell from which he or she has spent most of his/her time. To ensure that this most frequented location is the actual user's place of residence the constraint that at least one third of the tweets has been posted from this location is imposed. The resulting dataset contains about 59,000 users' places of residence. The number of valid users is shown in Table 1 for each touristic site. In the same way, we identify the country of residence of every user who have posted a tweets from one of the touristic sites during the time period.

Two metrics have been considered to measure the attractiveness of a touristic site based on the spatial distribution of the places of residence of users who have visited this site:

Figure 1 Density of geolocated tweets and positions of the touristic sites.

Table 1 Number of valid users by touristic sites

Site	Users	Site	Users
Alhambra (Granada, Spain)	1,208	Angkor Wat (Cambodia)	947
Corcovado (Rio, Brazil)	1,708	Eiffel Tower (Paris, France)	11,613
Forbidden City (Beijing, China)	457	Giza (Egypt)	205
Golden Pavilion (Kyoto, Japan)	1,114	Grand Canyon (US)	1,451
Hagia Sophia (Istanbul, Turkey)	2,701	Iguazu Falls (Argentina-Brazil)	583
Kukulcan (Chichen Itzá, Mexico)	209	London Tower (London, UK)	3,361
Machu Pichu (Peru)	987	Mount Fuji (Japan)	2,241
Niagara Falls (Canada-US)	920	Pisa Tower (Pisa, Italy)	1,270
Saint Basil's (Moscow, Russia)	262	Taj Mahal (Agra, India)	378
Times Square (NY, US)	13,356	Zocalo (Mexico City, Mexico)	16,193

- *Radius*: The average distance between the places of residence and the touristic site. The distances are computed using the Haversine formula between the latitude and longitude coordinates of the centroids of the cells of residence and the centroid of the touristic site. In order not to penalize isolated touristic sites, the distances have been normalized by the average distance of all the Twitter users' places of residence to the site. It has been checked that the results are consistent if the median is used instead of the average radius for the rankings.
- *Coverage*: The area covered by the users' places of residence computed as the number of distinct cells (or countries) of residence.

To fairly compare the different touristic sites which may have different number of visitors, the two metrics are computed with 200 users' place of residence selected at random and averaged over 100 independent extractions. Note that unlike the coverage the radius does not depend on the sample size but, to be consistent, we decided to use the same sampling procedure for both indicators.

3 Results

3.1 Touristic sites' attractiveness

We start by analyzing the spatial distribution of the users' place of residence to assess the attractiveness of the 20 touristic sites. In Figure 2(a) and Figure 2(b), the touristic sites are ranked according to the radius of attraction based on the distance traveled by the users

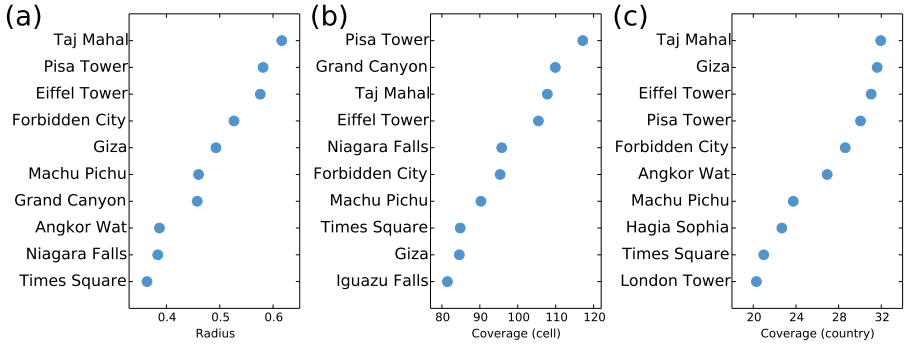

Figure 2 Ranking of the touristic sites according to the radius and the coverage. (a) Radius. **(b)** Coverage (cell). **(c)** Coverage (country).

from their cell of residence to the touristic site and the area covered by the users' cells of residence. In both cases, the results are averaged over 100 random selection of 200 users. The robustness of the results have been assessed with different sample sizes (50, 100 and 150 users), we obtained globally the same rankings for the two metrics. Both measures are very correlated and for most of the site the absolute difference between the two rankings is lower or equal than 2 positions. However, since the metrics are sensitive to slightly different information both rankings also display some dissimilarities. For example, the Grand Canyon and the Niagara Falls exhibit a high coverage due to a large number of visitors from many distinct places in the US but a low radius of attraction at the global scale.

To complete the previous results, we also consider the number of countries of origin averaged over 100 random selection of 200 users. This gives us new insights on the origin of the visitors. For example, as it can be observed in Figure 3, the visitors of the Grand Canyon are mainly coming from the US, whereas in the case of the Taj Mahal the visitors' country of residence are more uniformly distributed. Also, it is interesting to note that in most of the cases the nationals are the main source of visitors except for Angkor Wat (Table 2). Some touristic sites have a national attractiveness, such as the Mont Fuji or Zocalo hosting about 84% and 93% of locals, whereas others have a more global attractiveness, this is the case of the Pisa Tower and the Machu Pichu welcoming only 21% of local visitors.

More generally, we plot in Figure 2(c) the ranking of touristic sites based on the country coverage. The results obtained are very different than the ones based on the cell coverage (Figure 2(b)). Indeed, some touristic sites can have a low cell coverage but with residence cells located in many different countries, this is the case of the Pyramids of Giza, which went up 7 places and appears now in second position. On the contrary, other touristic sites have a high cell coverage but with many cells in the same country, as in the previously mentioned cases of the Grand Canyon and the Niagara Falls. Finally, the ranks of the Taj Mahal, the Pisa Tower and Eiffel Tower are consistent with the two previous rankings, these three sites are always in the top 5. Finally, we compare quantitatively the rankings with the Kendall's τ correlation coefficient which is a measure of association between two measured quantities based on the rank. In agreement with the qualitative observations, we obtain significant correlation coefficients comprised between 0.66 and 0.77 confirming the consistency between rankings obtained with the different metrics.

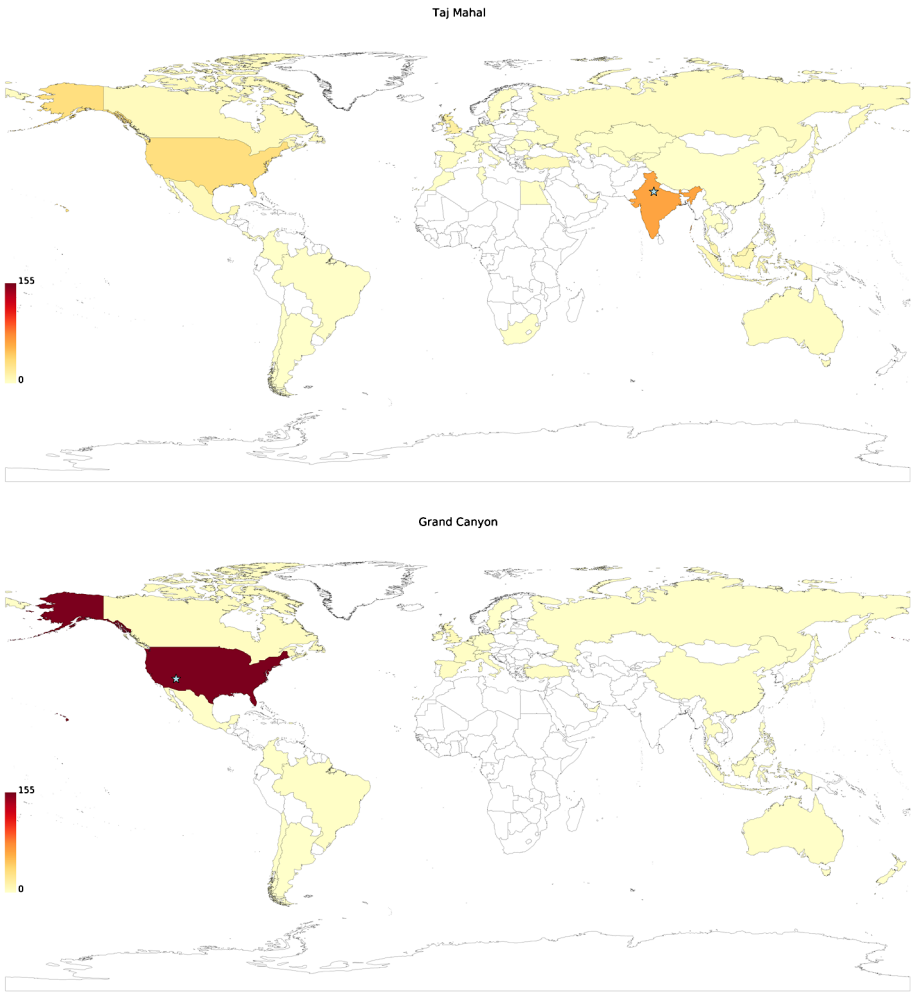

Figure 3 Heat map of the spatial distribution of the visitors' country of origin for the Taj Mahal and the Grand Canyon. The results have been averaged over 100 extractions of 200 randomly selected users.

3.2 Touristic site's visiting figures by country of residence

We can also do the opposite by studying the touristic preferences of the residents of each country. We extract the distribution of the number of visitors from each country to the touristic sites and normalize by the total number of visitors in order to obtain a probability distribution to visit a touristic site according to the country of origin. This distribution can be averaged over the 70 countries with the higher number of residents in our database (gray bars in Figure 4). The Eiffel Tower, Times Square and the London Tower welcome in average 50% of the visitors of each country. It is important to note that these most visited touristic sites are not necessarily the ones with the higher attractiveness presented in the previous section. That is the advantage of the method proposed in [38], which allows us to measure the influence and the power of attraction of regions of the world with different number of local and non-local visitors.

We continue our analysis by performing a hierarchical cluster analysis to group together countries exhibiting similar distribution of the number of visitors according to the touristic sites. Countries are clustered together using the ascending hierarchical clustering

Table 2 The three countries hosting most of the visitors for each touristic site

Site	Top 1	Top 2	Top 3
Alhambra (Spain)	Spain 71.14%	US 6.06%	UK 2.61%
Angkor Wat (Cambodia)	Malaysia 19.64%	Philippines 17.4%	US 9.59%
Corcovado (Brazil)	Brazil 81.13%	US 4.92%	Chile 3.08%
Eiffel Tower (France)	France 26.75%	US 16.62%	UK 8.92%
Forbidden City (China)	China 26.48%	US 14.46%	Malaysia 10.95%
Giza (Egypt)	Egypt 30.65%	US 9.8%	Kuwait 5.85%
Golden Pavilion (Japan)	Japan 60.74%	Thailand 11.72%	US 4.84%
Grand Canyon (US)	US 75.79%	UK 2.87%	Spain 2.16%
Hagia Sophia (Turkey)	Turkey 71.26%	US 5.48%	Malaysia 1.67%
Iguazu Falls (Arg-Brazil-Para)	Argentina 48.26%	Brazil 26.61%	Paraguay 8.63%
Kukulcan (Mexico)	Mexico 73.78%	US 10.07%	Spain 2.83%
London Tower (UK)	UK 65.61%	US 10.24%	Spain 2.77%
Machu Pichu	Peru 20.43%	US 19.95%	Chile 10.43%
Mount Fuji (Japan)	Japan 84.01%	Thailand 5.83%	Malaysia 2.66%
Niagara Falls (Canada-US)	US 60.5%	Canada 16.31%	Turkey 3.25%
Pisa Tower (Italy)	Italy 20.85%	US 13.56%	Turkey 10.95%
Sant Basil (Russia)	Russia 66.71%	US 5.06%	Turkey 3.77%
Taj Mahal (India)	India 27.97%	US 15.59%	UK 7.61%
Times Square (US)	US 74.32%	Brazil 3.26%	UK 2.31%
Zocalo (Mexico)	Mexico 92.22%	US 3.1%	Colombia 0.77%

The countries are ranked by percentage of visitors.

method with the average linkage clustering as agglomeration method and the Euclidean distance as similarity metric, respectively. To choose the number of clusters, we used the average silhouette index [39]. The results of the clustering analysis are shown in Figure 4. Two natural clusters emerge from the data, these clusters are without surprise composed of countries which tend to visit in a more significant way touristic sites located in countries belonging to their cluster. The first cluster gather countries of America and Asia whereas the second one is composed of countries from Europe and Oceania.

3.3 Network of touristic sites

In the final part of this work, we investigate the relationships between touristic sites based on the number of Twitter users who visited more than one site during a time window between September 2010 and October 2015. More specifically, we built an undirected spatial network for which every link between two touristic sites represents at least one user who has visited both sites. As a co-occurrence network, the weight of a link between two sites is equal to the total number of users visiting the connected sites. The network is represented in Figure 5 where the width and the brightness of a link is proportional to its weight and the size of a node is proportional to its weighted degree (strenght). The Eiffel Tower, Times Square, Zocalo and the London Tower appear to be the most central sites playing a key role in the global connectivity of the network (Table 3). The Eiffel Tower alone accounted for a 50% of the total weighted degree. The three links exhibiting the highest weights connect the Eiffel Tower with Time Square, the London Tower and the Pisa Tower representing 30% the total sum of weights. Zocalo is also well connected with the Eiffel Tower and Time Square representing 11% of the total sum of weights.

4 Discussion

We study the global attractiveness of 20 touristic sites worldwide taking into account the spatial distribution of the place of residence of the visitors as detected from Twitter. In-

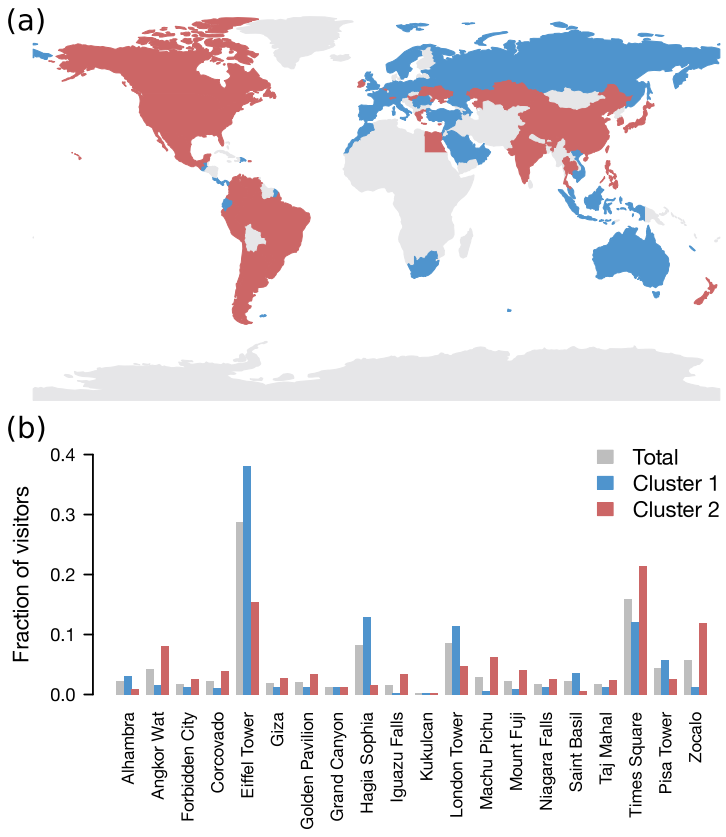

Figure 4 Clustering analysis. (a) Map of the spatial distribution of the country of residence according to the cluster. **(b)** Fraction of visitors according to the touristic site.

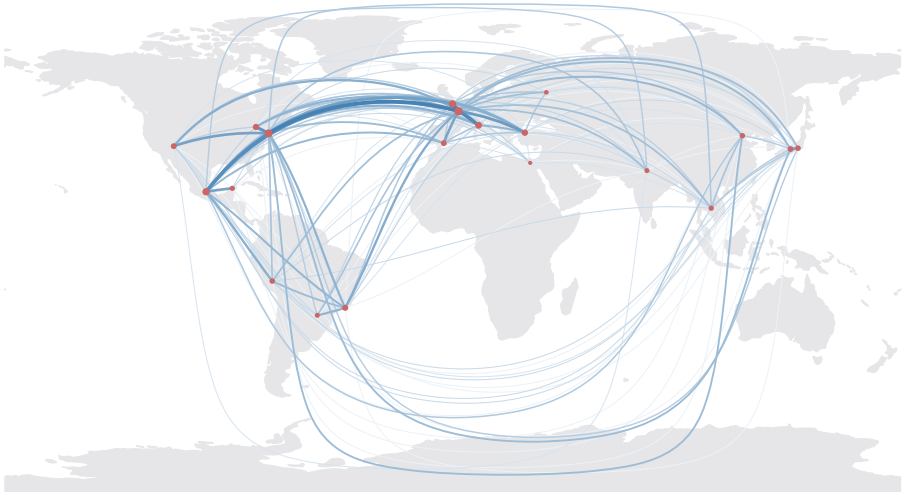

Figure 5 Network of undirected trips between touristic sites. The width and the brightness of a link is proportional to its weight. The size of a node is proportional to its weighted degree.

Table 3 Ranking of nodes based on the total weight

Site	Node Total Weight
Eiffel Tower (France)	0.51
Times Square (US)	0.35
Zocalo (Mexico)	0.21
London Tower (UK)	0.2
Pisa Tower (Italy)	0.12
Hagia Sophia (Turkey)	0.08
Niagara Falls (Canada-US)	0.07
Corcovado (Brazil)	0.05
Alhambra (Spain)	0.05
Grand Canyon (US)	0.05

stead of studying the most visited places, the focus of the analysis is set on the sites attracting visitors from most diverse parts of the world. A first ranking of the sites is obtained based on cells of residence of the users at a geographical scale of 100 by 100 kilometers. Both the radius of attraction and the coverage of the visitors' origins consistently point toward the Taj Mahal, the Eiffel tower and the Pisa tower as top rankers. When the users' place of residence is scaled up to country level, these sites still appear on the top and we are also able to discover particular cases such as the Grand Canyon and the Niagara Falls that are most visited by users residing in their hosting countries. At country level, the top rankers are the Taj Mahal and the Pyramids of Giza exhibiting a low cell coverage but with residence cells distributed in many different countries.

Our method to use social media as a proxy to measure human mobility lays the foundation for even more involved analysis. For example, when we cluster the sites by the country of the origin of their visitors, two main clusters emerge: one including the Americas and the Far East and the other with Europe, Oceania and South Africa. The relations between sites have been also investigated by considering users who visited more than one place. An undirected network was built connecting sites visited by the same users. The Eiffel Tower, Times Square, Zocalo and the London Tower are the most central sites of the network.

In summary, this manuscript serves to illustrate the power of geolocated data to provide world wide information regarding leisure related mobility. The data and the method are completely general and can be applied to a large range of geographical locations, travel purposes and scales. We hope thus that this work contribute toward a more agile and cost-efficient characterization of human mobility.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
AB and ML designed the study, analysed the data and wrote the manuscript. BG, AT and JJR designed the study and wrote the manuscript. All authors read, commented and approved the final version of the manuscript.

Author details
[1]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Palma de Mallorca, 07122, Spain. [2]Center for Data Science, New York University, 726 Broadway, 7th Floor, New York, 10003, USA. [3]Aix Marseille Université, Université de Toulon, CNRS, CPT, UMR 7332, Marseille, 13288, France.

Acknowledgements
Partial financial support has been received from the Spanish Ministry of Economy (MINECO) and FEDER (EU) under project INTENSE@COSYP (FIS2012-30634), and from the EU Commission through project INSIGHT. The work of ML has been funded under the PD/004/2013 project, from the Conselleria de Educación, Cultura y Universidades of the Government of the Balearic Islands and from the European Social Fund through the Balearic Islands ESF operational program for 2013-2017. JJR acknowledges funding from the Ramón y Cajal program of MINECO. BG was partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018).

Received: 17 November 2015 Accepted: 16 March 2016 Published online: 25 March 2016

References

1. United Nations World Tourism Organization UNWTO (2015) World Tourism Barometer
2. Christaller W (1964) Some considerations of tourism location in Europe: the peripheral regions-underdeveloped countries-recreation areas. Pap Reg Sci 12:95-105
3. Williams A, Shaw G (1991) Tourism and economic development: western European experiences, 2nd edn. Belhaven Press, London
4. Hazari R, Sgro M (1995) Tourism and growth in a dynamic model of trade. J Int Trade Econ Dev 4:253-256
5. Durbarry R (2004) Tourism and economic growth: the case of Mauritius. Tour Econ 10:389-401
6. Proença S, Soukiazis E (2008) Tourism as an economic growth factor: a case study for southern European countries. Tour Econ 14:791-806
7. Matias A, Nijkamp P, Sarmento M (2009) Advances in tourism economics: new developments. Physica-Verlag, Heidelberg
8. Long PT, Perdue R, Allen L (1990) Rural resident tourism perceptions and attitudes by community level of tourism. J Travel Res 28:3-9
9. Madrigal R (1993) A tale of tourism in two cities. Ann Tour Res 20:336-353
10. Jurowski C, Uysal M, Williams DR (1997) A theoretical analysis of host community resident reactions to tourism. J Travel Res 36:9-11
11. Lindberg K, Johnson RL (1997) Modeling resident attitudes towards tourism. Ann Tour Res 24:402-424
12. Andereck KL, Vogt C (2000) The relationship between residents' attitudes toward tourism and tourism development options. J Travel Res 39:27-36
13. Gursoy D, Jurowski C, Uysal M (2002) Resident attitudes: a structural modeling approach. Ann Tour Res 29:79-105
14. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439:462-465
15. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779-782
16. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327:1018-1021
17. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7:e37027
18. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as a proxy for global mobility patterns. Cartogr Geogr Inf Sci 41:260-271
19. Lenormand M, Louail T, Garcia Cantú O, Picornell M, Herranz R, Barthelemy M, San Miguel M, Ramasco JJ (2015) Influence of sociodemographic characteristics on human mobility. Sci Rep 5:10075
20. Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. IEEE Pervasive Comput 6:30-38
21. Soto V, Frías-Martínez E (2011) Automated land use identification using cell-phone records. In: Proceedings of the 3rd ACM international workshop on MobiArch. HotPlanet '11. ACM, New York, pp 17-22. doi:10.1145/2000172.2000179
22. Frías-Martínez V, Soto V, Hohwald H, Frías-Martínez E (2012) Characterizing urban landscapes using geolocated tweets. In: SocialCom/PASSAT. IEEE, Amsterdam, pp 239-248
23. Pei T, Sobolevsky S, Ratti C, Shaw SL, Zhou C (2014) A new insight into land use classification based on aggregated mobile phone data. Int J Geogr Inf Sci 28:1988-2007
24. Louail T, Lenormand M, Garcia Cantú O, Picornell M, Herranz R, Frías-Martínez E, Ramasco JJ, Barthelemy M (2014) From mobile phone data to the spatial structure of cities. Sci Rep 4:5276
25. Grauwin S, Sobolevsky S, Moritz S, Gódor I, Ratti C (2014) Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong. In: Helbich M, Jokar Arsanjani J, Leitner M (eds) Computational approaches for urban environments
26. Lenormand M, Picornell M, Garcia Cantú O, Tugores A, Louail T, Herranz R, Barthelemy M, Frías-Martínez E, Ramasco JJ (2015) Comparing and modeling land use organization in cities. R Soc Open Sci 2:150449
27. Louail T, Lenormand M, Picornell M, Garcia Cantú O, Herranz R, Frías-Martínez E, Ramasco JJ, Barthelemy M (2015) Uncovering the spatial structure of mobility networks. Nat Commun 6:6007
28. Lenormand M, Picornell M, Garcia Cantú O, Tugores A, Louail T, Herranz R, Barthélemy M, Frías-Martínez E, Ramasco JJ (2014) Cross-checking different source of mobility information. PLoS ONE 9:e105184
29. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111:15888-15893
30. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, González MC, Colizza V (2014) On the use of human mobility proxy for the modeling of epidemics. PLoS Comput Biol 10:e1003716
31. Asakura Y, Iryo T (2007) Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transp Res, Part A, Policy Pract 41:684-690
32. Shoval N, Isaacson M (2007) Tracking tourists in the digital age. Ann Tour Res 34:141-159
33. Girardin F, Calabrese F, Fiore FD, Ratti C, Blat J (2008) Digital footprinting: uncovering tourists with user-generated content. IEEE Pervasive Comput 7:36-43
34. Freytag T (2010) Déjá-vu: tourist practices of repeat visitors in the city of Paris. Soc Geogr 5:49-58
35. Poletto C, Tizzoni M, Colizza V (2012) Heterogeneous length of stay of hosts' movements and spatial epidemic spread. Sci Rep 2:476
36. Poletto C, Tizzoni M, Colizza V (2013) Human mobility and time spent at destination: impact on spatial epidemic spreading. J Theor Biol 338:41-58. http://www.sciencedirect.com/science/article/pii/S0022519313004062
37. Bajardi P, Delfino M, Panisson A, Petri G, Tizzoni M (2015) Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Sci 4:3. http://www.epjdatascience.com/content/4/1/3
38. Lenormand M, Gonçalves B, Tugores A, Ramasco JJ (2015) Human diffusion and city influence. J R Soc Interface 12:20150473
39. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53-65