How Reliable Is Social Media Data? Validation of TripAdvisor Tourism Visitations Using Independent Data Sources

Shihan Ma(✉) and Andrei Kirilenko

Department of Tourism, Hospitality and Event Management, University of Florida, Gainesville, USA
andrei.kirilenko@ufl.edu

Abstract. Social media data has been rapidly applied as alternative data source for tourism statistics and measurement in recent years due to its availability, easy collection, good spatial coverage at multiple scales, and rich content. However, frequent criticism towards the social media is the bias towards the population of social media users leading to unknown representativeness of the entire population. The purpose of this study is to cross-validate the reliability and validity of visitation pattern of tourist destinations retrieved from the social media using alternative independent data sources. The primary social media data is TripAdvisor reviews of Florida attraction points, restaurants, and hotels. The inferred visitation pattern was validated against two independent datasets: cellphone tracking data and official visitor surveys. The validity was explored in tourist origins, destinations, and travel flows. Repetitively, travel patterns inferred from the social media were found strongly correlated to those from cellphone tracking and surveys. The visitation data obtained from social media was concluded to be reliable and representative.

Keywords: Social media · Validation · Tourism visitation · Cellphone data · User-generated content (UGC)

1 Introduction

The statistical measurement of tourism has been a vital task for all stakeholders in tourism fields since its emergence in modern economy [1,2]. Historically, major supranational organizations such as the United Nations Statistical Commission (UNSC) and World Tourism Organization (WTO), along with national and regional tourism entities have provided the official tourism data for public. However, this data largely rely on conventional surveys resulting in inconsistencies across countries, costly data collection, problems with respondents' mobility, and variability in sampled population [3–8]. The big data provided an alternative source of low-cost data tracing tourists' movements, preferences, points of interests, behaviors and even expenditures [9], together with novel data collection methodologies [10]. In the big data domain, social media is particularly promising due to its availability, seamless collection, good spatial coverage at multiple scales, and rich content [11], which has been convincingly demonstrated in multiple studies [12–15].

W. Wörndl et al. (Eds.): *Information and Communication Technologies in Tourism 2021*, pp. 286–293, 2021.
https://doi.org/10.1007/978-3-030-65785-7_26

Meanwhile, frequent criticism towards the social media is the suggested bias towards the population of social media users leading to unknown representativeness of the entire population [16,17]. Complicating the issue, population representativeness may vary time and across social media platforms [11]. The inherent bias of the social media data has long been debated [18], yet the attempts to measure its extent are extremely limited [19,20]. The purpose of this study is to cross-validate the reliability and validity of visitation pattern of tourist destinations retrieved from the social media with alternative independent data sources. The primary social media data is TripAdvisor reviews of Florida attraction points, restaurants, and hotels. The inferred visitation pattern was validated against two independent datasets: cellphone tracking data and official visitor surveys.

2 Data and Methods

2.1 Social Media Data

We collected all TripAdvisor reviews of Florida attractions, hotels, and restaurants (further – properties) published from January 2003 to October 2019. The collected variables included reviewers' self-reported place of living address, the total review numbers, property location, and review date. The data was cleaned in the following way: we (1) filtered out the abnormally active reviewers ranking in top 5%; (2) used Google location API to geotag the reviewers' place of living (at a city, county, state, of country level); and (3) classified the visitors into three groups based on their origins, that is, Floridians, USA domestic, and international. The home locations were kept with at least a city granularity for Floridians, state granularity for domestic visitors, and nation granularity for the international visitors.

Data cleaning resulted in a total of 2,162,249 reviews generated by 250,844 reviewers (visitors) to 51,525 Florida properties. Between the reviewers, 24.4% were Floridians, 57.4% domestic, and 18.2% were international tourists. These groups contributed 42.6%, 39.6%, 13.6% of reviews, respectively. Based on the visitors' origin (place of living) and destination (location of the visited property), the database was rearranged as a monthly visitation frequency for each visitor group in the origin-destination (OD) format (see Table 1).

2.2 Cellphone Data

The primary independent dataset used for cross-validation was the trilaterated mobile phone signal tower data provided by AirSage (www.airsage.com). The anonymized data (over 8 billion records) covered Florida and adjacent areas from October 2018 to September 2019 and was organized in a form of OD trip counts for visitors from different home zones with a census tract granularity. The raw was preprocessed to filter out non-tourism travels and aggregated at a monthly time scale. Then, data was separated into two market segments: Floridians and domestic visitors. The origins of the domestic were aggregated at the state level. International visitors' information was largely unavailable in cellphone database and was excluded from research (Table 1).

2.3 VISIT FLORIDA Survey Data

The secondary cross-validation dataset was the Florida Visitor Study survey from Visit Florida (visitflorida.org). The annual survey is the premier reference guide on visitors to Florida. These data largely rely on conventional survey tools such as questionnaires and interviews. The data used in this study cover 2015–2018 and include quarterly statistics on domestic and international visitors: the origins at a state and nation scales and the total number of Florida visitors. The data on destinations visited in Florida is not provided; the local Florida tourists is also not included. Data summary is provided in Table 1.

Table 1. The data used in this research.

	Origin	Destination	Geo resolution	Timeframe	Time frequency
Social media data					
Floridian	Yes	Yes	County - County	2003–2019	Monthly
Domestic	Yes	Yes	State - County	2003–2019	Monthly
Int'l	Yes	Yes	Nation - County	2003–2019	Monthly
Cellphone data					
Floridian	Yes	Yes	Tract - Tract	2018.10–2019.9	Monthly
Domestic	Yes	Partial	State - Tract	2018.10–2019.9	Monthly
Int'l	No	No	Not applicable	2018.10–2019.9	Monthly
Survey data					
Floridian	No	No	Not applicable	Not applicable	Not applicable
Domestic	Yes	Partial	State - Region	2015–2018	Seasonal
Int'l	Yes	Partial	Nation - Region	2015–2018	Seasonal

2.4 Methods

Based on data availability and spatial resolution, the validation methodology was as follows:

- to validate the origins of Floridians inferred from the social media, their spatial distributions were compared with the cellphone data. Pearson's r correlation between the log-transformed paired data on the number of visits from each origin was used to estimate the match between different data sources.
- in a similar way, to validate the origins of domestic visitors, the destination of Floridians, and the travel flows of Floridians, their respective representations in different databases were used.

3 Results

3.1 Validation of Trip Origins

The validation of the origins of Floridian travel was based on the social media and cellphone data at a county resolution. The data on the top travel origins from both datasets are shown in Table 2. The inferred numbers of trips (log-transformed) from

same origins estimated from social media and cellphone data are highly correlated ($r = 0.93$, $p < 0.001$). The preliminary estimation implies that one TripAdvisor trip approximately corresponds to 100 trip counts from the cellphone data (Fig. 1).

Table 2. Top origin counties of Floridians

Origin	N Trips Cellphone	N Trips Social media
Palm Beach	1,531,156	15,309
Hillsborough	1,435,614	13,325
Miami-Dade	1,205,709	12,986
Duval	1,147,412	8,711
Orange	1,128,544	14,447
Broward	866,250	15,048
Lee	830,158	9,218
Pinellas	816,220	10,156
Polk	757,789	4,618
Brevard	674,611	6,898

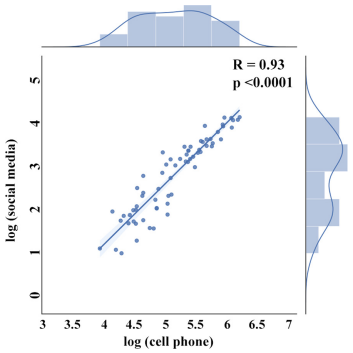

Fig. 1. Correlation of log (social media) * log(cellphone) trip origin counts

Validation of the origins of domestic US visitors was based on the comparison between social media, cellphone, and survey data, at a state level resolution. The data on the top 15 origin states provided in the Survey was compared with data from the other two datasets (Table 3) and demonstrated high cross-correlation (Fig. 2). The data

Table 3. Top origin states for the US domestic visitors

Origin State	N Trips Cellphone	N Trips Social media	N Trips Survey 2018
Georgia	834,620	30,639	11,935,176
New York	661,042	34,242	10,021,044
California	368,015	11,633	4,503,840
Texas	364,626	16,726	4,841,628
North Carolina	326,272	15,850	5,292,012
New Jersey	277,725	16,661	4,729,032
Ohio	271,976	18,393	4,841,628
Pennsylvania	270,474	19,783	5,742,396
Alabama	266,281	7,955	5,404,608
Virginia	266,275	12,605	3,265,284
Illinois	261,418	18,124	5,517,204
Massachusetts	203,044	14,162	3,152,688
Michigan	200,719	13,589	4,278,648
Tennessee	175,951	12,853	4,616,436
Indiana	150,297	9,280	3,603,072
Maryland	141,362	8,943	2,927,496

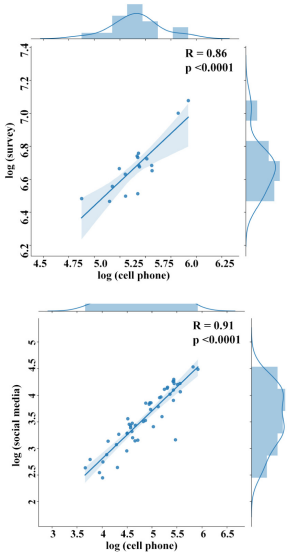

Fig. 2. Correlations of origin trip counts estimated from three datasets.

implies that one TripAdvisor trip count is equivalent to 100 trips inferred from the cellphone data and 2000 trips inferred from Visit Florida survey, hence providing the base to ranslate the social media and cellphone record data to real visitation data.

4 Validation of Destinations

Validation of the destination choices of Floridian travelers was based on data from social media and cellphone, on a county level resolution. The comparative data for the top destinations from both datasets are found in Table 4. The comparative numbers of trips are highly correlated ($r = 0.89$, $p < 0.0001$) (Fig. 3). The preliminary estimation implies that each trip count from the social media approximates 100 trip count from cellphone data.

Table 4. Top destination counties for Floridian tourists.

County	N Trips Cellphone	N Trips Social media
Orange	4,222,721	32,014
Miami-Dade	2,573,859	9,787
Hillsborough	1,878,658	8,779
Broward	911,854	8,149
Palm Beach	690,836	7,145
Polk	677,824	2,978
Duval	622,774	5,804
Pinellas	549,217	11,055
Osceola	534,319	5,746
Seminole	534,261	1,893

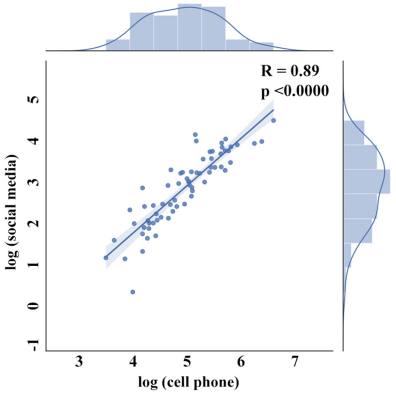

Fig. 3. Correlation of log (social media) * log (cellphone). Floridian travelers only.

5 Validation of Travel Flows

The validation on the origin-destination travel flows of Floridians was based on data from the social media and cellphones at a county level resolution. The number of trips for the top network links are shown in Table 5. The number of OD trips are strongly correlated ($r = 0.72$, $p < 0.01$) (Fig. 4). One travel estimated from the social media approximates 180 travels estimated from the cellphone data.

Table 5. Top OD flows for Floridian tourists

Origin - destination	N Trips Cellphones	N Trips Social media
Palm Beach - Miami-Dade	806,358	1,619
Hillsborough - Orange	699,457	3,606
Duval - Orange	339,657	2,187
Pinellas - Orange	326,103	2,286
Lee - Miami-Dade	325,180	622
Duval - Hillsborough	295,452	496
Orange - Hillsborough	284,670	1,314
Miami-Dade - Palm Beach	234,129	1,109
Miami-Dade - Orange	224,680	3,748
Marion - Orange	222,944	668

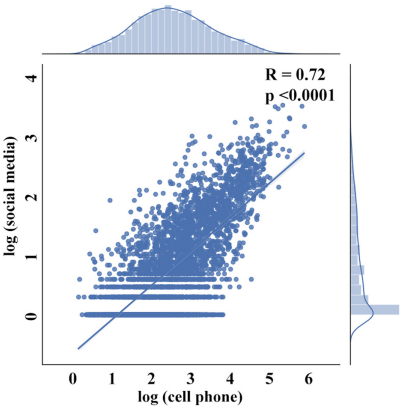

Fig. 4. Cross-plot of log (Social media) * log (cellphone)

6 Conclusions

We found that the social media is a reliable source of data on tourism visitations representative not only of the social media users, but also of the general population. The travel patterns extracted from social media are strongly correlated to those retrieved from the cellphone tracking data and official tourist surveys. The reliability of social media data is evidenced not only in the counts of tourists arriving from various origins or going to various destinations, but also in the travel origin-to-destination travel flows. A longitudinal comparison based on visitation temporal patterns in a future study is suggested to improve the robustness of our results.

This strong correlation in addition implies the potential of social media to represent the real visitation data by fusing the high-resolution social media with the overall tourism measurements from the state or national tourism organizations. In our data, one trip count from the social media approximately represents 2000 visitations from the survey data.

The two high-resolution data sources used in this study, social media and cell phone tracking, can both be used in visitation measurements. Notably, social media data has lower granularity, especially in determining visitor origins. We however found that the seemingly high resolution of the cell phone data can result in significant errors in urban areas. In addition, very high costs of the cellphone data determine its primary area of use in social media validation in key areas.

References

1. Burkart AJ, Medlik S (1981) Tourism, Past. Present and Future, London
2. Lickorish LJ (1997) Travel statistics—the slow move forward. Tour Manag 18(8):491–497
3. Hannigan K (1994) A regional-analysis of tourism growth in Ireland. Reg Stud 28(2):208–213
4. Guizzardi A, Bernini C (2012) Measuring underreporting in accommodation statistics: evidence from Italy. Curr Issues Tour 15(6):597–602
5. Frechtling DC, Hara T (2016) State of the world's tourism statistics and what to do about it. Tour Econ 22(5):995–1013
6. Volo S, Giambalvo O (2008) Tourism statistics: methodological imperatives and difficulties: the case of residential tourism in island communities 1,3. Curr Issues Tour 11(4):369–380
7. Latham J, Edwards C (2003) The statistical measurement of tourism. Prog Tour Recreat Hosp Manag 1:55–76
8. Aroca P, Brida JG, Volo S (2017) Tourism statistics: correcting data inadequacy. Tour Econ 23(1):99–112
9. Volo S (2018) Tourism data sources: from official statistics to big data. In: The SAGE Handbook of Tourism Management: Theories, Concepts and Disciplinary Approaches to Tourism, 2018, pp 193–201
10. Li J, Xu L, Tang L, Wang S, Li L (2018) Big data in tourism research: a literature review. Tour Manag 68:301–323
11. S. (David) Ma, A. P. Kirilenko, and S. Stepchenkova, "Special interest tourism is not so special after all: Big data evidence from the, (2017) Great American Solar Eclipse". Tour Manag 77:2020
12. Leung D, Law R, van Hoof H, Buhalis D (2013) Social media in tourism and hospitality: a literature review. J Travel Tour Mark 30(1–2):3–22
13. Donaire JA (2011) Barcelona Tourism image within the flickr community. Cuad Tur 27:1061–1062
14. Zheng Y-T, Zha Z-J, Chua T-S (2012) Mining travel patterns from geotagged photos. ACM Trans Intell Syst Technol 3(3):1–8
15. Hernández JM, Kirilenko AP, Stepchenkova S (2018) Network approach to tourist segmentation via user generated content. Ann Tour Res 73:35–47
16. Diaz F, Gamon M, Hofman JM, Kiciman E, Rothschild D (2016) Online and social media data as an imperfect continuous panel survey. PLoS ONE 11(1):e0145406
17. Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2(13):13
18. Crampton JW et al (2013) Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartogr Geogr Inf Sci 40(2):130–139

19. Steiger E, Westerholt R, Resch B, Zipf A (2015) Twitter as an indicator for whereabouts of people? correlating Twitter with UK census data. Comput Environ Urban Syst 54:255–265
20. Heikinheimo V, Di Minin E, Tenkanen H, Hausmann A, Erkkonen J, Toivonen T (2017) User-Generated Geographic Information for Visitor Monitoring in a National Park: A Comparison of Social Media Data and Visitor Survey. ISPRS Int J Geo-Inf 6(3):85