Comparing International Movements of Tourists

Official Census versus Social Media

Lucas Eduardo Bonancio Skora
lucasskora@alunos.utfpr.edu.br
Federal University of Technology - Paraná
Curitiba, Paraná

Thiago Henrique Silva (Advisor)
thiagoh@utfpr.edu.br
Federal University of Technology - Paraná
Curitiba, Paraná

ABSTRACT

Tourism is one of the most competitive, lucrative, and socially important economic segments in the world. Therefore, understanding the behavior of tourists is strategic for improving services and results. Many studies in the literature explore this issue using traditional data, such as surveys. These approaches provide reliable, precise information, but it is hard to obtain on a large scale, making studying worldwide patterns difficult. Location-Based Social Networks (LBSNs) could minimize these problems due to the ease of acquiring large amounts of detailed behavioral data. Nevertheless, before using such data, it is imperative to determine whether the information reveals behaviors comparable to traditional data - our ground truth. Thus, this work investigates whether the international flow of tourists measured with LBSN data is similar to the behavior estimated by the World Tourism Organization with traditional data sources. Our results suggest that LBSNs data represent the studied behavior well, indicating that they could be used in research regarding tourism mobility at different levels.

KEYWORDS

Large-scale Assessment, Mobility, Tourists, Social Media, Networks

1 INTRODUCTION

Several studies explore tourist behavior with different types of data sources. This type of study is socially and economically strategic [9], but the data sources normally explored, such as traditional surveys [7, 12], are typically challenging to create and organize on larger scales or lack the level of detail desired for some research paths. Despite these limitations, traditional survey data from official sources remains the ground truth for various tourism research questions. For example, Lozano and Gutiérrez [4] study the worldwide flow of tourists using data from the World Tourism Organization (WTO) obtained traditionally, leading to important insights for the development of new strategies concerning international tourism.

Seeking to address these limitations, recent studies are using data from location-based social networks (LBSNs) [8, 10]. LBSN data is usually easier to obtain and allows analysis at smaller scopes, such as specific people or venues. In contrast, traditional sources tend to aggregate results by cities or countries. Thus, data from these new sources, e.g., LBSNs, have the potential to complement traditional ones in different ways and alleviate some methodological problems of research based on only one of the two [1].

In: I Concurso de Trabalhos de Iniciação Científica (CTIC 2021), Minas Gerais, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2021.

ISSN 2596-1683

For this potential to be explored, it is necessary to validate LSBN data. Therefore, the question that guides this work is: Does LBSN data accurately represent international tourism flow compared to traditional data provided by the WTO? Seeking to answer this question, we compare differences and similarities between information obtained using official WTO data modeled and studied by Lozano and Gutierrez [4] and information obtained using LBSN data extracted in this work. Our results indicate that data from LBSNs, specifically Foursquare, are comparable to official data and satisfactorily represent the reality of international tourism flow.

2 RELATED WORKS

Other studies have already explored mobility with traditional data, social networks data, or both sources. This section describes some of them and their findings.

Miguens and Mendes [5] analyze the distributions of degree and strength in a directed international mobility graph where nodes are countries, and the weight of the edges represents the number of tourists traveling between those countries, generated based on data from a 2004 WTO survey. They discovered that degree distribution is random, but the weighted network is scale-free (follows a power-law distribution), demonstrating the importance of considering the weight of edges in a tourism graph. In addition, they investigated the correlation between degree and strength, finding that, for the analyzed network, in-degree has a strong correlation with in-strength. At the same time, out-strength grows almost quadratically concerning out-degree.

Hawelka et al. [3] use geolocalized Twitter data for a study of international mobility. The data used was collected during the entire year of 2012, with 944 million entries and 13 million different users. When analyzing the network topology between nations, the clusters that emerge respect real geopolitical boundaries, even if physical distance is not used as a variable for the grouping. To validate the results, they were contrasted with statistics from the international economic forum, specifically on international tourist arrivals and tourism revenue, reaching a strong linear correlation.

Belyi et al. [2] argue that human mobility is very complex because of its variety, including very different patterns such as tourist visits without recurrence or permanent migrations. Furthermore, each data source may emphasize a specific aspect of mobility, showing a distinct bias. Their study combines three data sources: Flickr (representing leisure activities and sightseeing), Twitter (any activity in an environment with internet access, whether business or leisure), and official UN migration data. The results of this "multi-layer" model show patterns that are not visible with the three layers separately, better fitting expectations when compared to other international relationship networks, such as common language and trade

relation graphs. Comparisons were made in terms of the structure of the networks instead of directly comparing edges. This article also reports that country groupings in the multi-layer network tend to be geographically connected, even though the physical distance between countries is not used in the algorithm. It also says that it is possible to use a single data set to study specific aspects of human mobility, but that is not the case when trying to model all mobility. Another conclusion was that the normalized weight of the edges in the data sets extracted from Twitter and Flickr follow the same probability distribution with similar parameters, but the migration data tend to be more diverse.

Provenzano et al. [6] analyzed tourism within Europe using both WTO data and geolocalized Twitter data, concluding that there is a large overlap between the two data sets, even though social network data is more complete. Zhou et al. [11] studied the international trade network by building a graph with edges only between a country and its largest economic partner and tries to justify the validity of this model using a triad census. As each country's trade is focused on a few other countries and several countries concentrate trade, it makes sense to focus on the most important relationship for each country.

Our work differs from similar efforts because the focus is on the extensive comparison between official data obtained in traditional ways, such as those studied by Provenzano et al. [6], with mobility extracted using social media data. Although there are other comparisons between different data sources for this context, they focus on specific cases, not providing a clear indication of the extent to which social media data reflects characteristics obtained with official data regarding the international flow of tourists.

3 METHODOLOGY

Two data sources are used in this work: i) one formed by check-ins grouped by user obtained from the geolocated social network Foursquare [1]; ii) another one that includes official data from WTO surveys ("*Tourism statistics- Arrivals of non-resident visitors at national borders, by nationality*" and "*Tourism statistics-Outbound tourism- trips abroad by resident visitors to countries of destination (basis: arrivals in destination countries)*"), through an analysis made in [4]. This second database and the previous analysis were used to compare the new social media data.

3.1 Graph generation

To represent the flow of tourists in this study, the data is analyzed considering **from-to** pairs, that is, **from** one country **to** another. The country where a Foursquare user made the most check-ins is considered their homeland, and they are considered tourists in all other countries. For each user from a given country, all different countries visited are counted. A directed graph $G = (V, E)$ is then constructed, where the set V represents the selected countries and a directed edge $e_{i,j} \in E$ with weight $w_{i,j} \in \mathbb{N}$ connects the countries $v_i, v_j \in V$ if the number $w_{i,j}$ of tourists who live in v_i and have visited v_j (made at least one check-in in v_j) is greater than zero.

Following the proposal of [4], the sub-graphs $G_{in,k}$ (incoming tourists) and $G_{out,k}$ (outgoing tourists) are created according to the following definitions: $G_{in,k} = (V, E_{in,k})$ with $E_{in,k} \subset E$ and $e_{i,j} \in E_{in,k}$ if $e_{i,j}$ is among the $k \in \{1, 2, 3\}$ edges with the highest weight entering v_j; $G_{out,k} = (V, E_{out,k})$ with $E_{out,k} \subset E$ and $e_{i,j} \in E_{out,k}$ if $e_{i,j}$ is among the $k \in \{1, 2, 3\}$ edges with the highest weight coming out of v_i. Therefore, this study uses 6 sub-graphs representing each of the two databases. A graphical representation of the top-2 outbound sub-graph for the Foursquare database is in Figure 1. In this figure, the size of a node represents its in-degree, and the size of an edge, its weight.

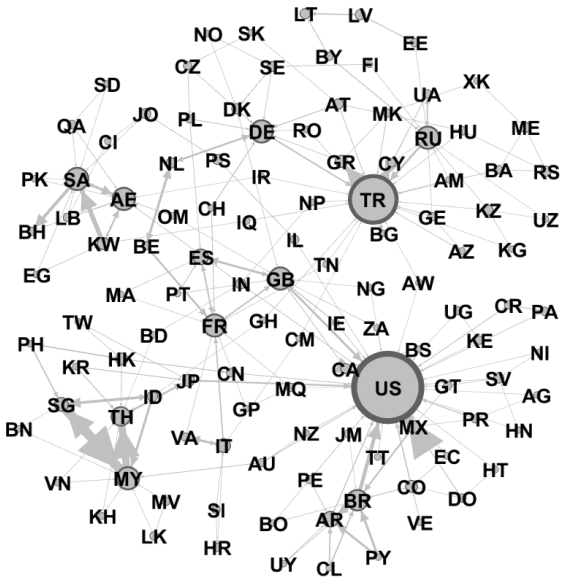

Figure 1: Graphical representation of the top-2 outbound sub-graph generated with Foursquare data.

Analyzing figure 1 visually, we note that the United States (US) and Turkey (TR) have the greatest importance in the network. This is intuitive for the United States, as a high number of American tourists in other countries and foreign tourists in the United States are expected. On the other hand, Turkey's importance is due to the disproportionate popularity of the social network in that country during the period in which Foursquare data was collected (2014). In addition, it is possible to identify "local" tourism centers by analyzing the sub-graph: Russia (RU), Malaysia (MY), Brazil (BR), Saudi Arabia (SA), Mexico (MX), Germany (DE), France (FR) Italy (IT) and the United Kingdom (GB) have secondary prominence. These local tourism centers also tend to be local powers in their respective regions.

For the graphs created for the Foursquare data, 117 countries were considered. For the WTO data, there are 214 countries in the top-k out sub-graphs and 148 in the top-k in sub-graphs, as there were no inbound data for the other 66.

4 RESULTS

Several analyzes were conducted on the sub-graphs described in the previous section: i) calculations of statistics such as the number of edges, density, diameter, transitivity, average geodesic distance, diameter, degree centralization, average strength, the results of a dyad type census, and arc reciprocity; ii) Rankings of the countries

[1] https://foursquare.com.

in the graphs according to Pagerank centrality, betweenness, in-strength and out-strength, and in-degree or out-degree (depending on the sub-graph type); iii) Analysis of strongly connected components and hierarchical clusters for the top-3 in and outbound sub-graphs; iv) Intercontinental tourism flows; v) Triad analysis; vi) motif census; and vi) ERGM analysis. These are all analyses done by the study used for comparison [4].

All results obtained were compared with the reference using traditional data. Due to limitations, we only illustrate some of them in this paper, namely, those concerning intercontinental tourism flow in the top-3 inbound sub-graph, and in-degree, in-strength, out-strength, and betweenness rankings for the top-3 out sub-graph. General conclusions obtained in this study are presented in the conclusions section.

4.1 Intercontinental tourism flow

An analysis of intercontinental tourism flow was carried out for the top-3 in and top-3 out sub-graphs. The continents considered were North America, South America, Europe, Africa, Asia, and Oceania. In the same way as in our reference study, directed graphs were used, with notations corresponding to the continents: $G_{cont,in} = (V_{cont}, E_{cont})$ where V_{cont} is the set of six continents listed and the weight of $e_{i,j} \in E_{cont,in}$ is equal to the sum of the weights of the edges in the top-3 in sub-graph such that the country of origin is from the continent v_i and the destination country from the continent v_j. The same is true for the $G_{cont,out}$ graph, switching "in" for "out" in the definitions. Transitions within the same continent were maintained.

The results of this analysis for the top-3 inbound sub-graph are described in table 1: The lines represent the origin continent of a tourist, and the columns their destination continents. The first number in the first line in each cell is the number of different tourists traveling in that transition, and in parentheses is the number of edges (without weight) corresponding to this flow.

For comparison purposes, the data for both databases were normalized by dividing the weight of each edge by the sum of the weight of all edges, thus obtaining the percentage of tourism flow that occurs between each pair of continents for each database; and the percentile difference between the normalized flows (both weighted, in parentheses, and unweighted) in each transition was calculated. This is shown in the second line of table 1.

Table 1 makes it clear that intracontinental tourism flows are stronger than intercontinental ones. In addition, the smaller amount of data corresponding to tourists from Oceania and Africa is also confirmed in the WTO data, although in different proportions. The average of the percentile differences in the top-3 in sub-graphs is 1.59% for the number of edges and 1.52% for the number of tourists. These small differences indicate that intercontinental relationships are well represented by the Foursquare database.

4.2 Country rankings

Table 2 shows the rankings of countries in the top-3 out sub-graphs according to some quantities, namely, in-degree, in-strength, out-weight, and betweenness for both databases. Out-degree is left out due to the sub-graphs construction, making all nodes have an out-degree of 3. Table 2 shows the rankings for the WTO database, with the ranking of each country in the Foursquare database in parentheses. Cells marked with "NA" indicate countries not included in the Foursquare graphs.

One common characteristic of the database rankings is the importance of the United States in the rankings, even though it doesn't take the first position in the Foursquare out-weight ranking (where it is in fifth place). Also, once again, Turkey is disproportionately present in the Foursquare database.

Note that some countries are always in similar positions regardless of the database, such as the United States and Thailand, and others in very different positions, such as Israel in the in-degree ranking and Ireland in the in-strength ranking. This is probably caused by the smaller number of check-ins in the database made by inhabitants of these smaller countries, leading to a less accurate model in these cases.

5 CONCLUSIONS

The tourism industry has become essential in the global economy, generating, according to the WTO, more than US$1 billion in 2019 alone and sustaining millions of jobs and companies worldwide. Therefore, studying tourists and their behavior is essential to facilitate the growth and improvement of this industry. The specific field of tourist mobility is still poorly studied, especially on a larger scale. One of the reasons for this is the difficulty in building the necessary data sets, as traditional data sources, such as surveys, do not scale easily and sometimes lack the necessary detail. In this context, there is an evident need to investigate alternative sources to study these phenomena. In this work, we identified that location-based social networks, specifically Foursquare, are comparable to official data and satisfactorily represent the reality of the international tourism flow. This happens even with some known limitations of LBSNs, such as the predominance of some demographic groups - mainly young people with regular internet access [8].

We noticed some differences between the survey and Foursquare results - which was to be expected. Most of these differences occurred when analyzing individual countries for which Foursquare data is limited due to weaker penetration or smaller presence in international tourism. Analyzes that took into account the structure of the sub-graphs as a whole, such as the intercontinental tourism flow analysis, tended to have closer results. It is thus possible to conclude that, although it cannot reliably represent specific smaller countries in every case, the general characteristics of the international tourism network are replicated in the LSBN data.

Besides, LBSN data's strength also goes beyond considering countries or regions as units. With LBSN, it is possible to work at the level of groups of individuals and/or specific addresses, allowing for a much more detail. In addition to providing unprecedented access level of details in scale, LBSNs are also useful for macroscopic analyses, as long as the researcher considers the bias towards more central countries and/or countries with greater penetration by the social network in question.

Our results open up a range of new opportunities for expanding and complementing studies on tourist movement using traditional sources, such as data from the WTO. New applications and services, for example, can benefit from this resource to develop innovative solutions to meet the demands of competitive global tourism.

Table 1: Intercontinental tourism flow for the top-3 in Foursquare sub-graph. Cell values represent number of tourists in a transition, in parentheses is the number of edges in that transition, and below is the percentile difference for the normalized tourism flow/edge count between the Foursquare and WTO databases.

	North America	South America	Europe	Africa	Asia	Oceania
North America	6862(42) 0.19%(-1.96%)	1172(9) 2.04%(0.99%)	1917(8) 3.15%(1.60%)	50(4) 0.03%(-0.43%)	1059(9) 1.32%-0.13%)	116(2) 0.20%(-1.00%)
South America	1215(9) 2.29%(1.66%)	3057(21) 4.16%(0.36%)	113(1) 0.22%(0.28%)	85(1) 0.16%(0.28%)	3(1) 0.00%(0.28%)	0(0) 0.00%(0.00%)
Europe	41(6) -0.87%(-3.00%)	0(0) -0.05%(-0.67%)	8266(77) -11.34%(10.47%)	269(12) -0.06%(-3.54%)	1687(10) 0.86%(0.6%)	0(0) -0.11%(-0.67%)
Africa	0(0) 0.00%(0.00%)	0(0) 0.00%(0.00%)	0(0) 0.00%(0.00%)	19(3) -3.02%(-8.58%)	0(0) -0.37%(-0.44%)	0(0) 0.00%(0.00%)
Asia	0(0) -0.07%(-2.69%)	0(0) 0.00%(0.00%)	7093(28) 11.75%(6.4%)	458(13) 0.50%(2.13%)	17389(91) -11.07%(3.45%)	295(3) 0.46%(-0.49%)
Oceania	0(0) 0.00%(0.00%)	0(0) 0.00%(0.00%)	0(0) 0.00%(0.00%)	0(0) 0.00%(-0.22%)	0(0) -0.16%(-0.44%)	48(1) -0.21%(-4.20%)

Table 2: Country rankings in the top-3 out sub-graph considering traditional data. Numbers in parenthesis refers to the position in the ranking using Foursquare data.

WTO Ranking	in-degree	in-strength	out-weight	Betweenness
1	USA(1)	USA(1)	USA(5)	USA(1)
2	Malaysia(11)	Spain(13)	Germany(19)	France(3)
3	South Africa(77)	France(7)	Canada(12)	Greece(15)
4	Canada (27)	Ukraine(32)	China(38)	Cyprus(16)
5	Ukraine(76)	Thailand(3)	United Kingdom (9)	Spain (10)
6	Thailand (8)	Malaysia (2)	Singapore (2)	Malaysia (26)
7	Greece(22)	Hong Kong (29)	Mexico (4)	Ukraine(17)
8	Spain(13)	Mexico(12)	Russian Federation(18)	Andorra (NA)
9	Benin(NA)	Canada(18)	France(21)	Philippines (64)
10	France(4)	Greece(15)	Italy(17)	Thailand(14)
11	Israel(66)	South Africa(79)	Netherlands (16)	South Africa (54)
12	Brazil (14)	Ireland(96)	Switzerland(31)	Canada (49)
13	Colombia(24)	Indonesia(17)	Spain(20)	Sri Lanka (110)
14	Angola(NA)	Brazil (10)	Japan (22)	Mexico(6)
15	Ethiopia(NA)	Uzbekistan (82)	Moldavia(NA)	Brazil(8)
16	Mali (NA)	Andorra(NA)	Malaysia (1)	Mauritius(NA)
17	Peru(37)	Peru(51)	South Korea (29)	Guadalupe(100)
18	Barbados(NA)	Cambodia (87)	Indonesia(10)	Dominica (NA)
19	Botswana(NA)	Philippines (42)	Belarus(30)	Hong Kong (30)
20	Mauritius(NA)	Botswana(NA)	Portugal(46)	Antigua and Barbuda(87)

ACKNOWLEDGMENTS

The student and advisor would like to thank professors Helen C. M. Senefonte, Myriam R. B. S. Delgado, and Ricardo Lüders for their help during the development of this work. They would like to thank also the São Paulo Research Foundation (FAPESP), Project GoodWeb process 2018/23011-1, for the funding provided.

REFERENCES

[1] Tarek Al Baghal, Alexander Wenz, Luke Sloan, and Curtis Jessop. 2021. Linking Twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Science* 10, 32 (2021), 1–20. https://doi.org/10.1140/epjds/s13688-021-00286-7

[2] Alexander Belyi, Iva Bojic, Stanislav Sobolevsky, Izabela Sitko, Bartosz Hawelka, Lada Rudikova, Alexander Kurbatski, and Carlo Ratti. 2017. Global multi-layer network of human mobility. *International Journal of Geographical Information Science* 31, 7 (2017), 1381–1402.

[3] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41, 3 (2014), 260–271.

[4] Sebastián Lozano and Ester Gutiérrez. 2018. A complex network analysis of global tourism flows. *International Journal of Tourism Research* 20, 5 (2018), 588–604.

[5] JIL Miguéns and JFF Mendes. 2008. Travel and tourism: Into a complex network. *Physica A: Statistical Mechanics and its Applications* 387, 12 (2008), 2963–2971.

[6] Davide Provenzano, Bartosz Hawelka, and Rodolfo Baggio. 2018. The mobility network of European tourists: a longitudinal study and a comparison with geo-located Twitter data. *Tourism Review* (2018).

[7] Raffaele Scuderi and Chiara Dalle Nogare. 2018. Mapping tourist consumption behaviour from destination card data: What do sequences of activities reveal? *International Journal of Tourism Research* 20, 5 (2018), 554–565.

[8] Thiago H. Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio Loureiro, and Daniele Quercia. 2019. Urban Computing Leveraging Location-Based Social Network Data: A Survey. *ACM Comput. Surv.* 52, 1, Article 17 (Feb. 2019), 39 pages. https://doi.org/10.1145/3301284

[9] WTO. 2021. *International Tourism Highlights, 2020 Edition.* World Tourism Organization, Madrid, Spain. https://doi.org/10.18111/9789284422456 arXiv:https://www.e-unwto.org/doi/pdf/10.18111/9789284422456

[10] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol.* 5 (2014), 38:1–38:55.

[11] Min Zhou, Gang Wu, and Helian Xu. 2016. Structure and formation of top networks in international trade, 2001–2010. *Social Networks* 44 (2016), 9–21.

[12] Marta Zieba. 2017. Cultural participation of tourists–Evidence from travel habits of Austrian residents. *Tourism Economics* 23, 2 (2017), 295–315.