Comparing Global Tourism Flows Measured by Official Census and Social Sensing

Lucas E. B. Skora\textsuperscript{a}, Helen C. M. Senefonte\textsuperscript{a,b}, Myriam Regattieri Delgado\textsuperscript{a}, Ricardo Lüders\textsuperscript{a} and Thiago H. Silva\textsuperscript{a,c}

\textsuperscript{a}Universidade Tecnológica Federal do Paraná (UTFPR), Av. Sete de Setembro, 3165, Curitiba, 80230-901, Brazil  
\textsuperscript{b}Universidade Estadual de Londrina (UEL), Rodovia Celso Garcia Cid, PR-445, Km 380, Londrina, 86057-970, Brazil  
\textsuperscript{c}University of Toronto, 55 St. George Street, Toronto, M5S 0C9, Canada

ARTICLE INFO

Keywords: Large-scale Assessment  
Mobility  
Tourists  
Social Media  
Networks

ABSTRACT

A better understanding of the behavior of tourists is strategic for improving services in the competitive and important economic segment of global tourism. Critical studies in the literature often explore the issue using traditional data, such as questionnaires or interviews. Traditional approaches provide precious information; however, they impose challenges to obtaining large-scale data, making it hard to study worldwide patterns. Location-based social networks (LBSNs) can potentially mitigate such issues due to the relatively low cost of acquiring large amounts of behavioral data. Nevertheless, before using such data for studying tourists’ behavior, it is necessary to verify whether the information adequately reveals the behavior measured with traditional data – considered the ground truth. Thus, the present work investigates in which countries the global tourism network measured with an LBSN agreeably reflects the behavior estimated by the World Tourism Organization using traditional methods. Although we could find exceptions, the results suggest that, for most countries, LBSN data can satisfactorily represent the behavior studied. We have an indication that, in countries with high correlations between results obtained from both datasets, LBSN data can be used in research regarding the mobility of the tourists in the studied context.

1. Introduction

Several studies available in the literature explore tourist behavior in different types of data sources. This type of study is strategic from an economic and social standpoint [1]. Many of these works use traditional survey data [2, 3], which are normally the ground truths for the various questions they address. For example, Lozano and Gutiérrez [4] study the global tourism network using data from the World Tourism Organization (WTO) obtained in a traditional way, leading to important insights for the development of new strategies concerning international tourism.

However, official data such as those from WTO are typically difficult to obtain because traditional approaches to data collection do not scale easily. Trying to solve this problem, recent studies are using data from location-based social networks (LBSNs) [5, 6]. LBSN data scale more easily and allow analysis at different granularities, going down to group or even individual level in specific areas. Traditional sources generally provide aggregated results for some particular locations, such as countries or specific cities [6], but usually do not reach the specificity of social networks. As such, data from these new sources have the potential to complement traditional ones in different ways.
2. Related Work

In this section, we present studies that explore tourist mobility with traditional and social networks data. For instance, Miguéns and Mendes [9] 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. Zheng et al. [10] analyzed 107 GPS logs of users over one year. The authors concluded that the mobility of residents and tourists is different; in addition, tourists’ behavior is influenced by their relationships and traveling experiences. Mobility is an under-explored behavior in part because it is hard to obtain an appropriate description for that [11, 12].

There are also several works using social media data to study the mobility of tourists. For example, Hawelka et al. [13] use georeferenced Twitter data to study international mobility. When analyzing the network topology between nations, the emerging clusters respect real geopolitical boundaries, even if physical distance is not used as a variable for the group. To validate the results, they were contrasted with statistics from the international economic forum, specifically on international tourist arrivals and tourism revenue, observing a strong linear correlation. Senefonte et al. [14] study regional influences on the behavior of tourists and residents in the context of mobility using Foursquare data. They show, among other results, that the ability to explore the cultural characteristics of each nationality in different destinations leads a promising way to improve recommendation systems for points of interest and other services to particular groups of tourists. Ferreira et al. [15] show that it is possible to go one step forward in the understanding of tourists’ mobility, identifying where and when places are more important to users in different cities, by using social media data.

Belyi et al. [16] 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 United Nations migration data. This "multi-layer" model shows patterns that are not visible with the three layers separately, better fitting expectations when compared with other international relationship networks, such as common language and trade relation graphs. Comparisons are made in terms of networks’ structure instead of directly comparing edges. The article also reports that country groups in the multi-layer network tend to be geographically connected, even though the physical distance between countries is not used in the algorithm. Another conclusion is that normalized weights of edges in the data sets extracted from Twitter and Flickr follow the same probability distribution with similar parameters. Still, the migration data tend to be more diverse.

Provenzano et al. [17] analyze tourism within Europe using both WTO data and georeferenced 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. [18] study 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.

The present work differs from similar efforts because its focus is on the extensive comparison between official data obtained using traditional ways, specifically those studied by Lozano and Gutiérrez [4], with mobility extracted using social media data. Although there are other comparisons between different data sources for the same context [16, 17], they focus on specific cases, not providing a clear indication of how extensively social media data reflects characteristics obtained by official data regarding the global tourism network. This work is also different because it studies tourist mobility on a global scale, while many others focus on small regions, such as the study of Provenzano et al. [17], Leung et al. [19], and Piazzi et al. [20].

3. Data and Methods

This section presents the data used and the key methodological steps adopted in the present work. We virtually replicate the main results obtained by Lozano and Gutiérrez [4]. However, instead of using WTO survey, the work considers Foursquare data, whose details are presented in the next section. Additionally, and as one of the contributions of the present work, we perform at the end of the paper a multidimensional analysis joining all the metrics to show where - in which countries - there are high correlations between datasets and where there are not.

3.1. Data Description

Two datasets are used in this work. The first one is composed of check-ins grouped by users obtained from the location-based social network Foursquare. The same dataset was also used in [14, 15]. It represents globally shared data spanning four months, April–July of 2014. The total number of check-ins in the dataset is 19,716,435. The total number of countries in the dataset is 230, but only countries with more than 1000 check-ins are considered in this study, resulting in 117 countries. Appendix A shows all 117 countries considered in this study grouped by continent. Africa and Oceania are under-represented continents in the

1https://foursquare.com.
Comparing Global Tourism Flows Measured by Official Census and Social Sensing

dataset due to: i) small population for Oceania; and ii) less access to the internet and weaker tourism industry for Africa.

Due to the origin of the dataset, it is important to consider the fact that it would be nearly impossible to acquire more recent data from the same source, as Foursquare is not as used as it was in 2014. However, it would be possible to acquire similar data from other social networks, such as Yelp. The fact the dataset spans three months does mean it also has a seasonal bias, as it would be autumn in the southern hemisphere and spring in the northern hemisphere. This may also be a cause of some of the disparities observed between the two datasets.

The other dataset contains 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").). These data represent the number of tourists arriving in/leaving a country from/to every other country under study. The WTO dataset has been analyzed by Lozano and Gutiérrez [4]; thus, the data and results are used in the present paper as a comparison basis with the Foursquare data. The WTO dataset under evaluation was collected originally in 2016.

3.2. Graph Generation

The global tourism network is modeled by graphs in this study. In the graphs obtained from Foursquare data, 117 different countries are considered. In the WTO data, there are 214 countries in the outbound tourism (Out) subgraphs and 148 countries in the inbound tourism (In). For 66 countries, there is no inbound data.

Similar to the assumptions made in [14], [13], and [15], the country where a user of Foursquare made the most check-ins is considered the homeland of that user, and for all the remaining countries, such user is considered a tourist. For each user from a given country, all different countries visited are counted. A directed graph \( G = (V, E) \) is then built, where the set \( V \) of nodes represents the selected countries and a directed edge \( e_{i,j} \in E \) with weight \( w_{i,j} \in \mathbb{N} \) connects countries \( v_i \) to \( v_j \in V \) if the number \( w_{i,j} \) of tourists who live in \( v_i \) and have visited \( v_j \) is greater than zero – i.e., at least one check-in was made in \( v_j \).

Following the proposal presented in the baseline work [4], subgraphs \( G_{\text{in,k}} \) (incoming tourists) and \( G_{\text{out,k}} \) (outgoing tourists) are created according to the following definitions: \( G_{\text{in,k}} = (V, E_{\text{in,k}}) \) with \( E_{\text{in,k}} \subset E \) and \( e_{i,j} \in E_{\text{in,k}} \) if \( e_{i,j} \) is among the \( k \in \{1, 2, 3\} \) edges with the highest weight entering \( v_j \); \( G_{\text{out,k}} = (V, E_{\text{out,k}}) \) with \( E_{\text{out,k}} \subset E \) and \( e_{i,j} \in E_{\text{out,k}} \) if \( e_{i,j} \) is among the \( k \in \{1, 2, 3\} \) edges with the highest weight coming out of \( v_j \). Therefore, this study uses 6 subgraphs representing each of the two datasets.

Different analyses are performed on the subgraphs, from basic statistics – obtained with the Python 3 packages NetworkX [21] and Pandas [22] – to more complex structural properties. First, we replicated all analyses performed by Lozano and Gutiérrez [4], which are used as baselines for comparing with the results obtained from social media data. Then we provide a final multidimensional analysis to show where there is correlation and where there is not.

3.3. Country Rankings Based on Centrality Measures

We use different centrality measures to rank the most important countries based on specific perspectives. First, we apply the PageRank centrality [23] in each subgraph. Next, we compute In-strength, Out-strength, and Betweenness centrality [24] for the Top-3 In and Top-3 Out subgraphs. We also compute In-degree for the Top-3 Out subgraph and Out-degree for the Top-3 In subgraph. These centralities are calculated using the NetworkX and Pandas libraries of the programming language Python. Further details on these centralities measures can be found in [24].

3.4. Hierarchical Clustering and Strongly Connected Components

The strongly connected components of the Top-3 In and Top-3 Out subgraphs are determined for the Foursquare dataset by a NetworkX library function. By following the procedures applied in [4], components containing only one country are ignored.

A hierarchical clustering algorithm with average linkage criteria [25] is used as in [4]. Here it is accomplished in Python with libraries Numpy [26], NetworkX, and Sklearn [27]. In this analysis, it is necessary to represent graphs as distance matrices. This is done by normalizing the adjacency matrices of the graphs. Normalization occurs by row for the Top-3 Out subgraph and by column for the Top-3 In subgraph. This is done so that the sum of the normalized elements \( n(w_{i,j}) \) of the rows or columns of the matrix results in 1. Afterward, each element of the distance matrix is updated with the value \( 1 - n(w_{i,j}) \), transforming normalized affinity into the normalized distance. To make the comparison simpler, the dendrogram cut-off point is defined to generate the same number of clusters as defined by Lozano and Gutiérrez [4], excluding those with only one country.

3.5. Cross-regional Tourism Flows

In addition to the flow between countries, an analysis of cross-regional tourism flow is also carried out. The regions considered are the continents of North and South America, Europe, Africa, Asia, and Oceania. In the same way, as in the country analysis, directed graphs are used, now with notations corresponding to the continents: \( G_{\text{cont,in}} = (V_{\text{cont}}, E_{\text{cont}}) \) where \( V_{\text{cont}} \) are the continents listed and the weight of \( e_{i,j} \in E_{\text{cont,in}} \) is equal to the sum of the edge weights in the Top-3 In subgraph such that the country of origin is from the continent \( v_i \) and the destination country from the continent \( v_j \).

Transitions within the same continent, meaning tourists’ travels to a different country in their home continent, were maintained, as in the reference study. For fairness in comparison with WTO data, the LBSN data are 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 dataset.
3.6. Motifs

This section describes the techniques used to reveal the local patterns of the subgraphs: motif analysis, which is performed with the mfinder program [28]. The triad census analysis, which has overlapping results with the motif ones, is shown in the Appendix D.

Motif analysis consists in counting and classifying the patterns of edges between each set of nodes in the graph. A comparison with random networks of the input graph size is made to determine the most relevant patterns.

The algorithm of mfinder uses the Z-value of the frequency of a particular motif in relation to the frequency of that motif in a set of random networks of the same size to choose the most relevant motifs in the input graph. Taking advantage of this, the results of the motif census using three nodes for Foursquare and WTO data have been compared through the percentage difference of the Z-values for each motif. This metric is used because it diminishes the influence of the subgraph size difference for the two datasets.

4. Results

This section presents the results obtained using the Foursquare dataset; they are further compared with those obtained by [4] using the WTO dataset. Full details on the results for the WTO dataset may be consulted in [4].

4.1. Overview of Subgraphs

First, subgraphs created with Foursquare data are evaluated. Figures 1(a), 1(b), and 1(c) show the subgraphs for the Top-1, Top-2, and Top-3 In mobility graphs, respectively, whereas, Figures 2(a), 2(b), and 2(c) show the Top-1, Top-2, and Top-3 Out mobility graphs, respectively. Each country in the network is represented by a node identified by the country’s ISO-3166 alpha-2 country code. The thickness of edges is proportional to their weight, and the size of the nodes is proportional to their Out-degree for Top-k In subgraphs and their In-degree for Top-k Out subgraphs.

Note that the United States (US) and Turkey (TR) have the greatest importance in the network. This phenomenon is intuitive for the US, as a great number of American tourists is expected in other countries as well as a great number of foreign tourists is expected in the US. On the other hand, the importance of Turkey is due to the great popularity of the social network in that country during the studied period. In addition, it is possible to identify regional tourism centers by analyzing the subgraphs. In almost all of them, Russia (RU), Malaysia (MY), Brazil (BR), Saudi Arabia (SA), Mexico (MX), Germany (DE), France (FR), Italy (IT), and the United Kingdom (GB) have a secondary position.

From these observations, it is possible to conclude that the United States is the largest recorded source and destination of tourists in the network. The fact that countries such as the United Kingdom, France, and Germany stand out more in the Top-k Out graphs are indications that those countries are more important globally as destinations than as tourist sources.

Comparing these graphs with those generated from WTO data, the most striking difference is the outstanding representation of Turkey in the dataset. This finding is important to keep in mind, especially for other analyses that are presented below. Different values of k provide particular information. For k = 1, the Top-1 In/Out subgraphs are extremely focused, useful for identifying the most critical relationship of each country. The Top-2 and Top-3 In/Out subgraphs, i.e., k = 2 and k = 3, are useful to identify small communities and relationships between more than two countries. This type of relationship can be helpful, for instance, to recommend countries to visit.

4.2. General Statistics

Tables 1 and 2 summarize general statistics about the created subgraphs. It is interesting to note that some information in the tables is determined by how the subgraphs are constructed: the number of edges for the Top-k In and Out subgraphs is determined by multiplying the number of countries by k. The average degree is equal to k, and the graph density is equal to k divided by the number of countries minus one, which means the subgraphs tend to have very low densities, as the number of countries is far greater than k. It is also notable that the transitivity of Top-1 graphs is always 0, as it is impossible to form triangles in these subgraphs. In the following statistics, only average strength takes into account the weight of the edges.

Table 1
Statistics of the Top-k Out subgraphs.

| Quantity                      | Top-1 Out | Top-2 Out | Top-3 Out |
|-------------------------------|-----------|-----------|-----------|
| Number of edges               | 117       | 234       | 351       |
| Density                       | 0.009     | 0.017     | 0.026     |
| Average geodesic distance     | 1.877     | 3.81      | 3.888     |
| Diameter                      | 5         | 13        | 12        |
| Average degree                | 1         | 2         | 3         |
| In-degree centralization      | 0.210     | 0.289     | 0.342     |
| average strength              | 208,085   | 316,632   | 388,513   |
| number of                     | 11/95/    | 29/176/   | 51/249/   |
| mutual/assymmetric/null dyads | 6680      | 6581      | 6486      |
| arc reciprocity               | 0.188     | 0.248     | 0.291     |
| transitivity                  | 0.000     | 0.410     | 0.413     |

Table 2
Statistics of the Top-k In subgraphs.

| Quantity                      | Top-1 In  | Top-2 In  | Top-3 In  |
|-------------------------------|-----------|-----------|-----------|
| Number of edges               | 117       | 234       | 351       |
| Density                       | 0.009     | 0.017     | 0.026     |
| Average geodesic distance     | 1.769     | 2.729     | 3.129     |
| Diameter                      | 4         | 9         | 9         |
| Average degree                | 1         | 2         | 3         |
| Out-degree centralization     | 0.210     | 0.342     | 0.430     |
| Average strength              | 265,932   | 367,726   | 437,726   |
| number of                     | 7/103/    | 18/198/   | 35/281/   |
| mutual/assymmetric/null dyads | 6676      | 6570      | 6470      |
| arc reciprocity               | 0.120     | 0.154     | 0.199     |
| transitivity                  | 0.000     | 0.029     | 0.049     |
Comparing Global Tourism Flows Measured by Official Census and Social Sensing

Figure 1: Top-k In graphs generated for the Foursquare dataset. Nodes represent countries identified by their ISO-3166 alpha-2 country code. The size of nodes is proportional to their Out-degree, and the size of edges to their weight.

Figure 2: Top-k Out graphs generated for the Foursquare dataset. Nodes represent countries identified by their ISO-3166 alpha-2 country code. The size of nodes is proportional to their In-degree, and the size of edges to their weight.

The densities of the diad types are not comparable between WTO and Foursquare data, as the density of the subgraphs directly influences the densities of the diads. Analyzing the average strength, the number of tourists represented by WTO data is around ten times greater than the Foursquare data for the inbound (In) subgraphs and from three to five times greater for the outbound (Out) subgraphs.

Average geodesic densities between Foursquare and WTO data are mostly similar. In-degree centralization is smaller in Foursquare data than in WTO data, but Out-degree centralization is greater in Foursquare than in WTO data. Arc reciprocity is slightly greater in the Top-k Out Foursquare subgraphs than in the WTO ones, but between Top-k In subgraphs, it stays basically the same.

4.3. Centrality Analyses

We first study the results from the PageRank centrality. Table 3 shows the 20 countries with the highest PageRank centralities. Countries highlighted in bold are among the top 20 countries of all three Top-k In subgraphs or all three Top-k Out subgraphs. Countries in bold with an asterisk following their names appear in all six rankings.

The countries’ names written in bold in the Top-k Out subgraphs of the Foursquare base are the US, Mexico, Turkey, Cyprus, Netherlands, Saudi Arabia, Bahrain, Malaysia, Thailand, Germany, and the UK (11 out of 20 countries), and in the WTO dataset (obtained from [4]), the US, Mexico, South Africa, Thailand, Botswana, Malaysia, France, Spain, Ukraine, Israel, Mauritius, Hong Kong, Benin, Greece, Ethiopia and the Philippines (16 out of 20 countries). Thus, it is clear that the classification of the centrality in the Top-k Out subgraphs is more homogeneous in WTO data than in Foursquare data. Furthermore, in the three subgraphs of the two datasets – Top-1/2/3 Out –, the United States has the highest centrality. Turkey’s disproportionate popularity in the Foursquare dataset has influenced its presence at the top of the three indexes as well as Cyprus, an important tourist destination for Turks. Finally, no African country appears in these indices, unlike the WTO data.
Table 3
Countries with the Top-20 highest Pagerank centralities for each Foursquare subgraph. Countries in bold appear in all 3 Top-k In or all Top-k out rankings. Countries in bold with an asterisk following their name appear in all six rankings.

| Subgraph     | Top-1 Out | Top-2 Out | Top-3 Out |
|--------------|-----------|-----------|-----------|
| Top-1 Out    |           |           |           |
| US*          | US*       | US*       |           |
| Mexico       | Turkey    |           |           |
| Turkey       | Mexico    | Canada    |           |
| Cyprus       | Canada    | Turkey    |           |
| Netherlands* | Cyprus    | Spain     |           |
| Saudi Arabia | Saudi Arabia | Mexico   |           |
| Belgium      | Greece    | Canada    |           |
| Bahrain      | UAE       | Germany   |           |
| Malaysia*    | United Kingdom | Saudi Arabia |           |
| Thailand     | Malaysia* | Italy     |           |
| Italy        | Spain     | UAE       |           |
| Vatican      | Germany   | Belgium   |           |
| Latvia       | Bahrain   | Netherlands* |           |
| Palestine    | France    | Malaysia* |           |
| Israel       | Thailand  | cyprus    |           |
| Lithuania    | Brazil    |           |           |
| Germany      | Netherlands* | Thailand |           |
| United Kingdom | Singapore | Vatican  |           |
| Sri Lanka    | Kuwait    |           |           |
| Maldives     | Japan     | Bahrain   |           |
| Top-1 In     | Top-2 In  | Top-3 In  |           |
| Maldives     | Guadeloupe | Martinique |           |
| Netherlands* | Martinique | Guadeloupe |           |
| Belgium      | Sri Lanka | Macedonia |           |
| Sri Lanka    | Maldives  | Kosovo    |           |
| Peru         | Macedonia | Sri Lanka |           |
| Chile        | Kosovo    | Maldives  |           |
| Portugal     | Latvia    | Kazakhstan |           |
| Malaysia*    | Lithuania | Quirguistan |           |
| Brazil       | Netherlands* | Uganda |           |
| US*          | Belgium   |           |           |
| Turkey       | US*       | Netherlands* |           |
| Oman         | Serbia    |           |           |
| Saudi Arabia | Malaysia* | US*       |           |
| Austria      | Montenegro | Serbia   |           |
| Argentina    | Germany   | Costa Rica |           |
| Slovakia     | Palestine | Latvia    |           |
| New Zealand  | Singapore | Belgium   |           |
| Vatican      | Peru      | El Salvador |           |
| Hong Kong    | Estonia   | Panama    |           |
| Palestine    | Saudi Arabia | Slovakia |           |

As shown in the second part of Table 3, for the Top-k In subgraphs of the Foursquare dataset, the countries in bold are Maldives, Netherlands, Belgium, Sri Lanka, Malaysia, and the US (6 of 20). The ones in the WTO dataset are the US, Canada, China, Hong Kong, Germany, Argentina, France, South Africa, the Russian Federation, Ukraine, and Uzbekistan (11 of 20). Interestingly, the countries with the highest centrality for the Foursquare Top-k In subgraphs are Maldives, Guadeloupe, and Martinique, with the US close to tenth place. At the same time, the US continues to dominate the Top-K In in the WTO dataset. This may indicate that, when considering individual countries, Foursquare data represents those that act primarily as tourist sources more accurately than those that serve as destinations; this last case is probably influenced by the LBSN bias.

Table 4
In-degree, In-strength, Out-strength, and Betweenness centralities for the Top-3 Out Foursquare subgraphs. Countries in bold appear in the first 20 positions for at least three out of four centrality rankings shown in the table, and countries with asterisks appear in the first 20 positions for all the four centrality rankings.

| Country       | In-degree | In-strength | Out-strength | Betweenness |
|---------------|-----------|-------------|--------------|-------------|
| 1 US*         | US*       | Malaysia    | US*          |             |
| 2 Turkey*     | Malaysia  | Singapore   | United Kingdom* |           |
| 3 United      | Thailand* | Turkey*     | France       |             |
| 4 France      | Turkey*   | Mexico*     | Turkey*      |             |
| 5 Germany*    | Singapore | US*         | Germany*     |             |
| 6 UAE         | Germany*  | Brazil*     | Mexico*      |             |
| 7 Russian Federation | France | Kuwait  | Belgium |             |
| 8 Thailand*   | United Kingdom | Saudi Arabia |         |
| 9 Italy       | Saudi Arabia | United Kingdom | Federation |
| 10 Malaysia   | Brazil*   | Indonesia   | Spain*       |             |
| 11 Saudi Arabia | Argentina | Thailand* | Italy         |
| 12 Mexico*    | Mexico*   | Canada      | UAE          |             |
| 13 Spain*     | Spain*    | Cyprus      | Oman         |             |
| 14 Brazil*    | UAE       | Paraguay    | Thailand*    |             |
| 15 Kuwait     | Greece    | Belgium     | Greece       |             |
| 16 Argentina  | Cyprus     | Netherlands | Cyprus       |             |
| 17 Serbia     | Indonesia | Italy       | Ukraine      |             |
| 18 Singapore  | Canada     | Russian     | Argentina    |             |
| Federation    | Germany*   | Japan       |             |             |

Complementing the analysis of the studied countries, Tables 4 and 5 show three centralities (In-strength, Out-strength, and Betweenness) common to the Top-3 In and Out subgraphs, whereas In-degree is shown only for Top-3 Out and Out-degree only for Top-3 In.

For simplicity, only the first 20 positions for each centrality measure are shown. Countries appearing in at least three columns are shown in bold for each table, whereas an asterisk identifies countries appearing in all four columns.

One common characteristic of the centrality rankings is the importance of the United States in the Top-3 Out rankings (Table 4), even though it does not take the first position in the Foursquare Out-strength ranking. The fact that the United States is less dominant in the output rankings (Table 5), appearing at the second place in the WTO centrality rankings and the first place only in one of the Foursquare rankings (In-strength), indicates that the United States plays a more predominant role as a “source” than as a destination in both datasets. Also, Turkey is, once again, disproportionately present in the Foursquare dataset, being first in both Out-degree and Out-strength rankings for the Top-3 In Foursquare subgraphs, even though it does not appear in the Top-20 of the equivalent WTO rankings. To provide fairness in the comparison, Tables 6 and 7 show the rankings of countries in the WTO datasets and, in parentheses, the position of that country in the corresponding Foursquare ranking. Cells marked with "NA" indicate countries with less than 1000
Table 5
Out-degree, Out-strength, In-strength, and Betweenness centralities for the Top-3 In Foursquare subgraphs. Countries in bold appear in the first 20 positions for at least three out of four centrality rankings shown in the table, and countries with asterisks appear in the first 20 positions for all the four centrality rankings.

| Ranking | Out-degree | Out-strength | In-strength | Betweenness |
|---------|------------|--------------|-------------|-------------|
| 1       | Turkey     | *Turkey*     | US*         | Turkey*     |
| 2       | US*        | US*          | Malaysia*   | US*         |
| 3       | Russian    | Malaysia*    | Singapore   | Russian     |
| 4       | Malaysia*  | Singapore    | Thailand    | Brazil*     |
| 5       | United     | Mexico*      | Turkey*     | Belarus     |
| 6       | Mexico*    | Brazil*      | Germany*    | United*     |
| 7       | Brazil*    | Russian      | Singapore   | Paraguay    |
| 8       | UAE*       | Singapore    | Mexico*     | Poland      |
| 9       | Singapore  | Kuwait        | Argentina   | Germany*    |
| 10      | Kuwait     | United       | Cyprus      | Ukraine     |
| 11      | France*    | Thailand     | United      | Mexico*     |
| 12      | Germany*   | Indonesia    | Greece      | Cyprus      |
| 13      | Singapore  | Cyprus       | France*     | Latvia      |
| 14      | Thailand   | Germany*     | Brazil*     | UAE*        |
| 15      | Chile      | Netherlands  | Canada       | Netherlands |
| 16      | Japan      | Japan        | Netherlands | Spain       |
| 17      | Serbia     | Italy        | Indonesia   | Malaysia*   |
| 18      | Colombia   | Belgium      | UAE*        | France*     |
| 19      | Philippines| France*      | Spain       | Kuwait      |
| 20      | Spain      | UAE*         | Italy       | Japan       |

Table 6
Comparison of the country rankings in the Top-3 Out subgraph. Countries are arranged according to their position in the WTO rankings, and the number in parentheses is their position in the Foursquare ranking. If a country does not appear in the Foursquare ranking, the position is replaced by "NA."

| WTO Ranking | In-degree | In-strength | Out-strength | Betweenness |
|-------------|-----------|-------------|--------------|-------------|
| 1           | US(1)     | US(1)       | US(5)        | US(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       | Spain(10)   |
| 6           | Thailand(8) | Malaysia(2) | Singapore    | Malaysia(26) |
| 7           | Greece(22)  | Hong Kong(29) | Mexico(4)   | Ukraine(17) |
| 8           | Spain(13)   | Mexico(12)  | Russia       | 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)    | South Korea(29) | Guadeloupe(100) |              |
| 18          | Barbados(NA) | Cambodia(87) | Indonesia(10) | Dominic(NA) |
| 19          | Botswana(NA) | Philippines(42) | Belarus(30) | Hong Kong(30) |
| 20          | Mauritius(NA) | Botswana(NA) | Portugal(46) | Antigua and Barbuda(87) |

Foursquare check-ins registered in their territory, ignored in the graph.

Note that some countries are always in similar positions regardless of the dataset, such as the United States and Thailand in the Top-3 Out rankings. A few others appear in very different positions, such as Israel (for In-degree ranking it appears at 11th in the WTO Top-3 Out and 66th in Foursquare) and Ireland (for In-strength ranking it appears at 12th in the WTO Top-3 Out and 96th in Foursquare). This is probably caused by the smaller number of check-ins in the dataset made by inhabitants of these smaller countries, leading to a less accurate model in these cases.

4.4. Strongly Connected Components
An analysis of the strongly connected components in the Top-3 In/Out subgraphs is also performed for both datasets. Full lists of the countries in each component together with their geographical displacement are shown in Appendix B.

Figures 3 and 4 help us to understand the results. They illustrate the correspondence between Foursquare and WTO regarding this metric, where on the left side are the connected components for the Foursquare data, and on the right are those of the WTO data. For the Top-3 Out subgraph, there is a large component in the WTO data that is made up of many members from two of the Foursquare components. In both datasets’ Top-3 In subgraphs, we have very similar small components strongly tied to geographic regions: South American, South Asian, Central American, Eastern European, and Arabic components. Note that in both Top-3 In and Out cases, we observe the tendency of smaller components linked to the same bigger one in the other dataset. This indicates considerable similarities among the results despite particular characteristics of the datasets that could have induced a component bigger or smaller.

4.5. Hierarchical Clustering
According to the procedure described in Section 3, clusters of countries can be obtained for both Top-k In and Top-k Out subgraphs. For comparison, Figures 5(a) and 5(b) show the results for Top-3 In subgraphs of Foursquare and WTO, respectively. Figures 6(a) and 6(b) show the results for Top-3 Out subgraphs of Foursquare and WTO, respectively.

Not surprisingly, most clusters are aligned with geographical regions. The complete list of countries in each cluster for Foursquare data is shown in Appendix C, and the list of countries in the WTO clusters can be found in [4].

As other studies have already observed [13, 16, 29, 30], many clusters are formed in geographically close regions, even if physical distance is not one of the variables used to calculate the clustering. Furthermore, the clusters found with the Foursquare data are similar to those obtained based on the WTO data. There are some small divergences between the clusters found in this work and in [4]. The observed exceptions are supposed to occur as we deal with two different data sources. This can be attributed to the lack of data for countries less central in international tourism and smaller Foursquare penetration.
Table 7

Comparison of the country rankings in the Top-3 In subgraph. Countries are arranged according to their position in the WTO rankings, and the number in parentheses is their position in the Foursquare ranking. If a country does not appear in the Foursquare ranking, the position is replaced by "NA."

| WTO Ranking | Out-degree | Out-strength | In-strength | betweenness |
|-------------|------------|--------------|-------------|-------------|
| 1           | US(2)      | Hong(103)    | China(32)   | US(2)       |
| 2           | United     | Kong(76)     | Mongolia    | China(69)   |
| 3           | France(11) | Germany(16)  | Italy(20)   | Russia(93)  |
| 4           | China(76)  | United       | Spain(19)   | China(69)   |
| 5           | Germany(12)| France(21)   | Poland(52)  | Ukraine(10) |
| 6           | Russia(3)  | Canada(14)   | France(13)  | Macao(NA)   |
| 7           | Japan(16)  | Federation(7)| Malaysia(8) | Taiwan(96)  |
| 8           | Australia(59)| Macao(NA) | Singapore(4)| Indonesia(35)|
| 9           | New Zealand(116)| | Ukraine(35)| | |
| 10          | Russia(109)| South Africa(29)| Federation(29)| | |
| 11          | Argentina(22)| Czech Republic(43)| Canada(15)| France(18)| |
| 12          | Brazil(7)  | Switzerland(72)| Sweden(58)| Egypt(70) |
| 13          | Italy(21)  | Taiwan(99)   | Turkey(5)   | Greece(31) |
| 14          | South Korea(63)| United| Kingdom(11)| Tanzania(NA)| |
| 15          | Netherlands(39)| Malaysia(3)| Thailand(4)| Saudi Arabia(22)| |
| 16          | Saudi Arabia(9)| Slovakia(90)| South Africa(88)| Turkey(1)| |
| 17          | Taiwan(99) | Kazakhstan(61)| South Australia(55)| | |
| 18          | Taiwan(99) | Kazakhstan(61)| Korea(25)| Australia(55)| |
| 19          | Malaysia(4) | Moldavia(NA) | Singapore(3)| Malaysia(17)| |

Even with these divergences, the clusters are geopolitically consistent with what is expected, including geographically close countries and/or countries with affinities such as similar languages or cultural habits. Some examples from the top-3 Out subgraph are the cluster of south-west Asian/north African and majority Muslim countries (United Arab Emirates, Egypt, Kuwait, Qatar, Oman, Sudan, Bahrain, Pakistan, and Saudi Arabia) and the cluster of South-American countries (Peru, Argentina, Uruguay, Paraguay, Brazil, and Chile). Examples from the top-3 In subgraph are a cluster comprised of France, two of its overseas departments and two of its former colonies (France, Martinique, Guadeloupe, Tunisia, and Morocco), and a cluster of Baltic/southern Slavic countries (Bosnia and Herzegovina, Montenegro, and Serbia).

4.6. Cross-regional Tourism Flow

To provide a macroscopic view of Top-3 In and Top-3 Out subgraphs, Tables 8 and 9 describe the flow of tourism between continents instead of countries. The value in each cell is the number of different tourists traveling from any country in a specific continent (row) to any country in a particular continent (column). For example, the cell North America to South America in Table 8 represents 332 trips of unique users from a country in North America to a country in South America.

As observed in [4], Tables 8 and 9 make it clear that intracontinental tourism flows (i.e., tourism flows between countries of the same continent, represented by the values in the main diagonal) are greater than intercontinental ones. In addition, the small amount of data corresponding to tourists...
Comparing Global Tourism Flows Measured by Official Census and Social Sensing

Figure 5: Clusters found for the Top-3 In graph generated for the Foursquare and WTO datasets. Colors in each figure represent a specific cluster. Countries in white have not been included in the model or have been ignored because they are part of a single-country component/cluster.

Figure 6: Clusters found for the Top-3 Out graph generated for the Foursquare and WTO datasets. Colors in each figure represent a specific cluster. Countries in white have not been included in the model or have been ignored because they are part of a single-country component/cluster.

Table 8
Cross-regional tourism flows for the Top-3 Out Foursquare subgraph. Cell values represent number of tourist’s trips in a transition.

|          | North America | South America | Europe | Africa | Asia | Oceania |
|----------|---------------|---------------|--------|--------|------|---------|
| North America | 6406          | 332           | 930    | 0      | 1    | 0       |
| South America | 3104          | 0             | 0      | 0      | 0    | 0       |
| Europe     | 566           | 0             | 9627   | 0      | 2461 | 0       |
| Africa     | 46            | 0             | 99     | 23     | 204  | 0       |
| Asia       | 560           | 0             | 3124   | 5      | 15989| 0       |
| Oceania    | 113           | 0             | 0      | 0      | 97   | 79      |

from Oceania and Africa is also confirmed in the WTO data, although in different proportions.

A comparison between the data from both datasets is made by calculating the percentage difference between the flows of each possible transition. The results are shown in Figures 7 and 8.

For Top-3 In subgraphs, the average percentage difference in the number of tourists is 1.52%. For Top-3 Out, the average is 1.74%. Due to many empty transitions (no tourist trips registered) for both datasets, there are several null cells.

Table 9
Cross-regional tourism flows for the Top-3 In subgraph Foursquare. Cell values represent number of tourist’s trips in a transition.

|          | North America | South America | Europe | Africa | Asia | Oceania |
|----------|---------------|---------------|--------|--------|------|---------|
| North Amer. | 6862          | 1172          | 1917   | 50     | 1059 | 116     |
| South Amer. | 1215          | 3057          | 113    | 85     | 3    | 0       |
| Europe    | 41            | 0             | 8266   | 269    | 1687 | 0       |
| Africa    | 0             | 0             | 19     | 0      | 0    | 0       |
| Asia      | 0             | 0             | 7093   | 458    | 17389| 295     |
| Oceania   | 0             | 0             | 0      | 0      | 48   | 48      |

These small differences between the two datasets indicate that intercontinental relationships are well represented in the Foursquare dataset.

4.7. Motif Analysis
Motif analysis identifies the most common edge patterns in the network, comparing their densities with the density of the same patterns in random nets of equal size. Table 10 shows important motifs identified by mfinder using both datasets. The cells contain the Z-score of a pattern frequency.
Comparing Global Tourism Flows Measured by Official Census and Social Sensing

Figure 7: Percentage difference between Foursquare and WTO dataset ([4]) of the intercontinental tourism flows for the Top-3 Out subgraphs. Negative values indicate lower values for Foursquare than WTO.

Figure 8: Percentage difference between Foursquare and WTO dataset ([4]) of the intercontinental tourism flows for the Top-3 In subgraphs. Negative values indicate lower values for Foursquare than WTO.

Relative to its frequency in a set of random graphs of the same size: To the left, data regarding the Foursquare dataset, and to the right, WTO data. Cells with NA represent patterns not identified as relevant for that subgraph and dataset.

There is a high correspondence between the most relevant motifs of Foursquare and WTO networks. The only pattern that has been identified as relevant for a Foursquare subgraph and not relevant for any WTO subgraphs is the complete triad (last row of Table 10). It may indicate a particular point that deserves further investigation in the future because it could be the case that LBSN data express certain relationships more precisely. Probably, it is complementary to other investigated sources, such as the WTO surveys investigated in the present work.

Figure 9 shows the percentage differences between the Z-values in the Top-3 In (red) and Out (blue) subgraphs of the two datasets for all three-node motifs (not only ones identified as relevant by mfinder in Table 10). Two difference values (-818.66 and -792.67) are not properly shown in Figure 9 due to scale limits. The average of the absolute values of the percentage differences is 32.86% for the Top-3 In subgraph and 216.77% for the Top-3 Out. These high differences may be due to bias in the LSNB data, favoring countries with a central role in international tourism whose residents/visitors have better access to the internet.

4.8. Multidimensional Comparison

So far, we have observed, exploring individual metrics, considerable similarities between the two datasets under analysis, yet some dissimilarities have also been observed. In this section, we improve the previous analyses by considering several metrics together. For this multidimensional experiment, we aggregate all metrics that could be directly interpreted as features. Five metrics are common to Top-k In and Out subgraphs, k={1,2,3}: In-strength, Out-strength, betweenness, pagerank, and ID of the strongly-connected component the country belongs to. Two are complementary for both types of subgraphs: In-degree for Top-k Out and

Table 10

| Motif            | Top-2 Out | Top-3 Out | Top-2 In | Top-3 In |
|------------------|-----------|-----------|----------|----------|
| A←B→C, A→C      | 7.4/4.21  | 20.9/3.41 | 12.5/4.73| 3.2/5.66 |
| A←B→C, A→C      | 0.8/NA    | 3.6/NA    | 4.2/7.06 | 12.1/10.38|
| A←B→C, A→C      | NA/NA     | 2.0/NA    | NA/5.31 | 12.1/NA  |
| A←B→C, A→C      | 2.0/7.19  | 9.5/4.15  | NA/NA   | 3.9/5.8  |
| A←B→C, A→C      | NA/NA     | 2.2/7.2   | 0.3/NA  | 2.0/15.09|
| A←B→C, A→C      | NA/NA     | 0.2/NA    | NA/NA   | NA/NA    |
| A←B→C, A→C      | NA/NA     | 0.2/NA    | NA/NA   | NA/NA    |

Figure 9: Percentage differences between the Z-values for each motif in each Top-3 subgraph for each dataset.
Out-degree for Top-k In. Aiming to aggregate numerical and nominal metrics, in the preprocessing step we perform the one-hot encoding process on the nominal variable ‘component’ and standardize all features by removing the mean and scaling to unit variance.

Considering that we have Top-(1,2,3) In and Out subgraphs, each one providing different metric values, every country is associated with six distinct multidimensional vectors - one for each subgraph. Therefore, for each country, an n-dimensional vector of features is built, associated to each subgraph. The dimension n includes five elements for In-strength, Out-strength, betweenness, pagerank, and In-degree (or Out-degree) features plus as many elements as needed by the one-hot encoding. This last number of elements varies between 63 (in the case of Top-3 Out) to 107 (in the case of Top-1 In) in WTO data. By considering m countries, a square m-dimensional symmetric matrix of distances is built with the euclidean distance $d_{ij}$ between the n-dimensional feature vectors of countries $i$ and $j$. Therefore, six matrices are obtained for all Top-{1,2,3} In and Out subgraphs. An average matrix of distances is then computed element by element over these six matrices. Each row of the resulting matrix thus contains the average distance between one particular country and the others. Each dataset (Foursquare and WTO) yields to its own average matrix.

The Pearson correlation $\rho$ for a particular country is then computed between its associated rows in Foursquare and in WTO average matrices. It represents how similar the country is related to the others in both datasets. Figure 10 shows the results for all countries that appear in both datasets. First, note that for most of these countries, $\rho \geq 0.6$, confirming the indication of similarity we observed in the previous sections. The figure also highlights countries with $\rho \in [0.2, 0.4]$, i.e., moderate positive correlations: Canada, China, France, Malaysia. One reason that might help explain these results is that those countries receive a high volume of tourists from different countries. Because social media data tend to be biased toward certain individuals in certain countries [6], this bias might impact a comprehensive analysis. For instance, France and Malaysia being more represented in the Foursquare dataset than their neighbors might also affect their distances to other countries. In addition, the fact that China restricts the use of certain web services might also play an important role in this result. Canada’s strong ties to the US across all subgraphs might skewer its results. Only one result corresponding to Turkey has a negative correlation. As we observed in the results from the previous sections, Turkey’s disproportionate popularity in the Foursquare dataset has influenced the results, impacting the analysis as well. Moreover, differently from what is observed in WTO data, few African countries appear in the Foursquare analysis. Therefore, unbalanced country information in each dataset is an important issue, even considering countries that appear in both datasets.

5. Conclusions

According to the World Tourism Organization (WTO), the urban tourism industry has become essential to several economies, producing, according to the WTO, more than US$1 billion only in 2019. Another benefit is the ability of such an industry to generate and sustain millions of jobs and companies around the world. In this context, studying tourists and their behavior is essential to facilitate the growth and improvement of the urban tourism industry. The specific field of tourist mobility is still poorly studied, especially on large scale. The difficulty in building suitable data sets can be cited as one of the main reasons for this gap.

Traditional data sources for studying tourist movements, such as surveys, do not scale easily and sometimes do not provide necessary details. Therefore, there is a need to investigate alternative sources to study the phenomena. In this work, we identified that Location-Based Social Networks (LBSN), specifically Foursquare, are comparable to some extent to official data and can satisfactorily represent the reality of the global tourism network. This happens even with some known limitations of LBSNs, such as the predominance of some demographic groups - mainly young people with regular internet access [6].

As expected, the results pointed out differences between survey and Foursquare data analyses. Many of the greatest differences occurred in the case of individual country analyses such as PageRank centrality. They differed mainly in the case of smaller and less central countries for international tourism, for which there are fewer data. Analyses considering the graph as a whole, such as the motif analysis, had highly similar results. A multidimensional comparison between the results obtained from Foursquare and WTO datasets showed that most countries have correlation values higher than 0.6, which means that both datasets share similar information. However, Canada, China, France, and Malaysia have moderated positive correlations between 0.2 and 0.4, and only Turkey has a negative correlation, probably due to the biased popularity of Turkey in Foursquare. These results suggest that Foursquare data can be useful for representing

![Figure 10: Pearson correlation coefficient expressing the relationship between distances from one country to all others using Foursquare and WTO. The x-axis, artificial displacement, is a controlled value just for visualization purposes; every point has one value that does not overlap another.](Image)
tourism behaviors with limitations due to the different popularity of social media in certain countries. This point should be considered in the future.

We would also like to point out that the strengths of LBSN data are not necessarily in macro analyses, thinking in terms of countries or regions as units. With these sources, it is possible to work at the level of individual addresses, going into more detail.

The results obtained so far have opened up a range of new opportunities for expanding and complementing studies on tourist movement that rely on traditional sources, such as data from WTO. With LBSN data, we can reach a global scale with a relatively low cost; in addition, we have the opportunity to explore a fine granularity in available data. Thus, for example, new applications and services could benefit from LBSN data to develop innovative solutions to meet the demands of the competitive global tourism.

Acknowledgements

This study is also partially supported by the projectGoodWeb (Grant 2018/23011-1 and 2020/07494-2 from São Paulo Research Foundation - FAPESP), and CNPq (grants 439226/2018-0, 314699/2020-1, 310998/2020-4).

References

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

[2] M. Zieba, Cultural participation of tourists—evidence from travel habits of Austrian residents, Tourism Economics 23 (2) (2017) 295–315.

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

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

[5] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: Concepts, methodologies, and applications, ACM Trans. Intell. Syst. Technol. 5 (3) (2014) 38:1–38:55.

[6] T. H. Silva, A. C. Viana, F. Benevenuto, L. Villas, J. Salles, A. Loureiro, D. Quercia, Urban computing leveraging location-based social network data: A survey, ACM Comput. Surv. 52 (1) (2019) 17:1–17:39.

[7] J. Murphy, M. W. Link, J. H. Childs, C. L. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, P. Harwood, Social media in public opinion research: report of the aapor task force on emerging technologies in public opinion research, American Association for Public Opinion Research, Deerfield (2014) 1–57.

[8] T. Al Baghal, A. Wenz, L. Sloan, C. Jessop, Linking twitter and survey data: asymmetry in quantity and its impact, EPJ Data Science 10 (32) (2021) 1–20.

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

[10] Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from gps trajectories, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 791–800.

[11] A. Lew, B. McRercher, Modeling tourist movements: A local destination analysis, Annals of tourism research 33 (2) (2006) 403–423.

[12] D. A. Fennell, A tourist space-time budget in the shetland islands, Annals of Tourism Research 23 (4) (1996) 811–829.
A. Countries Considered in this Study

These are the countries/territories from each continent considered in the present study (in parenthesis, we show the number of countries/territories considered in relation to the total number in that continent):

- **North America** (19 out of 45): the US, Mexico, Costa Rica, Canada, Dominican Republic, Martinique, Panama, El Salvador, Puerto Rico, Nicaragua, Guatemala, Honduras, Guadeloupe, Jamaica, Haiti, Aruba, Antigua and Barbuda, Trinidad and Tobago;

- **South America** (10 out of 15): Brazil, Chile, Paraguay, Colombia, Peru, Argentina, Venezuela, Ecuador, Uruguay, Bolivia;

- **Europe** (38 out of 49): Russia, UK, Spain, Latvia, Ukraine, Belgium, Germany, Italy, France, Netherlands, Cyprus, Belarus, Greece, Portugal, Hungary, Serbia, Poland, Sweden, Austria, Czech Republic, Finland, Romania, Switzerland, Ireland, Bulgaria, Denmark, Macedonia, Kosovo, Croatia, Vatican, Norway, Lithuania, Georgia, Estonia, Bosnia and Herzegovina, Slovakia, Slovenia, Montenegro;

- **Africa** (11 out of 55): Egypt, South Africa, Kenya, Morocco, Tunisia, Sudan, Uganda, Ghana, Nigeria, Cameroon, Côte d’Ivoire;

- **Asia** (37 out of 48): Turkey, Malaysia, Japan, Thailand, Indonesia, Philippines, Saudi Arabia, Kuwait, Singapore, South Korea, India, UAE, China, Sri Lanka, Hong Kong, Jordan, Lebanon, Taiwan, Maldives, Azerbaijan, Qatar, Oman, Kazakhstan, Vietnam, Pakistan, Bahrain, Israel, Iran, Kyrgyzstan, Uzbekistan, Brunei, Bangladesh, Palestine, Armenia, Iraq, Nepal, Cambodia;

- **Oceania** (2 out of 14): Australia, New Zealand;

B. Strongly Connected Components Found for Foursquare Data

We use Tarjan’s algorithm [31] with Nuutila’s modifications [32] (default parameters of NetworkX library). For the Top-3 Out subgraph obtained with Foursquare data, the strongly connected components with more than one element are:

- Argentine, Poland, Greece, Cyprus, Spain, Canada, Brazil, the UK, Italy, Paraguay, France, Chile, Mexico, Vatican, Germany, Russia, Ukraine, the US, Netherlands, Turkey, and Belgium; Denmark and Sweden; Belarus, Estonia, Lithuania, Latvia, and Finland; Panama and Costa Rica; Guatemala and El Salvador; Slovenia, Croatia, Montenegro, Serbia, and Bosnia and Herzegovina; South Korea, Thailand, Indonesia, Singapore, Malaysia, and Japan; Kazakhstan, Kyrgyzstan, and Uzbekistan; Jamaica, and Trinidad and Tobago; Oman, UAE, Saudi Arabia, Kuwait, and Bahrain; Palestine, Jordan, and Israel; China, Hong Kong, and the Philippines; Georgia, and Armenia; Guadeloupe and Martinique; the Maldives and Sri Lanka; Australia, and New Zealand; Uganda, and Kenya.

For Top-3 In for Foursquare data, they are:

- Macedonia and Kosovo; Bosnia and Herzegovina, Serbia, and Montenegro; Sri Lanka and Maldives; Singapore, Indonesia, Thailand, Japan, Malaysia, and the Philippines; Kyrgyzstan and Kazakhstan; El Salvador and Guatemala; Kuwait, UAE, and Saudi Arabia; Martinique, and Guadeloupe; Uganda and Kenya; Germany, Spain, Lithuania, Ukraine, Latvia, France, Netherlands, Russia, Belarus, Belgium, Poland, United Kingdom, and Cyprus; Costa Rica and Panama; Colombia, Canada, Chile, Paraguay, Mexico, Argentina, Brazil, and the US.

The components (ignoring those composed of only one country) for the Foursquare subgraphs are shown in Figures 11 and 12.

C. Clusters Found for Foursquare Data

These are the clusters found with the hierarchical clustering algorithm for the Foursquare subgraphs:

- Top-3 In subgraph clusters — Mexico, the US, Puerto Rico, Dominican Republic, Costa Rica, Canada, Jamaica, Haiti, and Panama; the UK, Spain, Ireland, Norway, South Africa, and Portugal; Philippines and Hong Kong; Russia, Kazakhstan, Belarus, Ukraine, Uzbekistan, Latvia, and Lithuania; France, Martinique, Guadeloupe, Tunisia, and Morocco; Trinidad and Tobago and Antigua, and Barbuda; Argentina, Uruguay, Paraguay, Brazil, Chile, and Bolivia; Bosnia and Herzegovina, Montenegro, and Serbia; Singapore, Malaysia, Brunei, Japan, Indonesia, Vietnam, Thailand, Taiwan, and South Korea; Sweden and Denmark; Australia and New Zealand; Colombia, Venezuela, and Ecuador; Turkey, Georgia, Armenia, Kosovo, Romania, Iran, Cyprus, Azerbaijan, Northern Macedonia, Bulgaria, and Greece; Uganda, and Kenya; Kuwait, Iraq, and Saudi Arabia; El Salvador, and Guatemala; United Arab Emirates, Maldives, Oman, India, Pakistan, and Sri Lanka; Croatia, and Slovenia; Qatar and Sudan; Germany, and Austria; Belgium, and the Netherlands.

- Top-3 Out subgraph clusters — Trinidad and Tobago, Mexico, Columbia, Venezuela, Ecuador, Aruba, Antigua and Barbuda, Bahamas, the US, Canada, and Jamaica; Singapore, Cambodia, Nepal, Malaysia, Japan, Bangladesh, Indonesia, Vietnam, Thailand, India and South Korea; United Arab Emirates, Egypt, Kuwait, Qatar, Oman, Sudan, Bahrain, Pakistan, and Saudi Arabia; Bosnia and Herzegovina, Kyrgyzstan, Turkey, Kazakhstan, and Cyprus; Finland, Georgia, Armenia, Norway, Russia, Sweden, Denmark, and Ukraine; Peru, Argentina, Uruguay, Paraguay,
Comparing Global Tourism Flows Measured by Official Census and Social Sensing

Table D.1
Triad census results. Number of occurrences of each pattern found in Top-k Out and Top-k In subgraphs.

| Triad | Top-1 Out | Top-2 Out | Top-3 Out | Top-1 In | Top-2 In | Top-3 In |
|-------|-----------|-----------|-----------|---------|---------|---------|
| A,B,C | 248,425   | 237,931   | 228,246   | 248,187 | 237,330 | 227,622 |
| A→B,C | 10,003    | 17,804    | 24,175    | 10,493  | 19,002  | 25,419  |
| A→B →C| 1,217     | 3,083     | 5,255     | 743     | 1,826   | 3,469   |
| A→B←C | 0         | 23        | 82        | 604     | 1,556   | 2,720   |
| A↔B→C | 47        | 85        | 150       | 41      | 111     | 189     |
| A↔B←C | 48        | 188       | 411       | 0       | 6       | 22      |
| A↔B↔C | 0         | 11        | 39        | 62      | 199     | 399     |
| A↔B↔C,A↔C| 0     | 31        | 61        | 0       | 35      | 110     |
| A↔C,A↔C| 0         | 0         | 0         | 0       | 0       | 0       |
| A↔B→C,A↔C| 0   | 8         | 18        | 0       | 0       | 7       |
| A↔B↔C,A↔C| 0   | 17        | 50        | 0       | 4       | 20      |
| A↔B↔C,A↔C| 0   | 7         | 27        | 0       | 21      | 59      |
| A↔B↔C,A↔C| 0   | 5         | 6         | 0       | 2       | 7       |
| A↔B↔C,A↔C| 0   | 4         | 13        | 0       | 5       | 13      |
| Brazil, and Chile; the UK, Côte d’Ivoire, France, Lebanon and Spain; Uganda, and Kenya; Martinique, and Guadeloupe; the Maldives, and Sri Lanka; Croatia, Slovenia, Vatican, Italy, Montenegro, and Serbia; Hungary, Slovakia, and Austria; Switzerland, Poland, Belgium, Czech Republic, Romania, Germany, and the Netherlands; Kosovo, Northern Macedonia, Bulgaria, and Greece; Palestine, Jordan, and Israel; Estonia, Belarus, Latvia, and Lithuania; Nigeria, and South Africa; Puerto Rico, and the Dominican Republic; China, the Philippines, Taiwan, and Hong Kong; Australia, and New Zealand; Nicaragua, Costa Rica, and Panama; Honduras, and El Salvador. |

D. Triad Census

Triad census consists of counting and classifying the patterns of edges between each set of three nodes in the graph. The analysis is done with UCINET 6.0 [33], Table D.1 shows the results of the census for Foursquare data. As mentioned in Section 3, to compare the results of Foursquare with those provided by WTO, subgraphs density must be considered.

The values distribution is similar between Foursquare and WTO data, especially regarding null elements. A clear difference is the higher frequency of "complete" triads (A↔B↔C) on the Foursquare network. This could be only related to the difference in densities; however, the hypothesis that it is due to data’s characteristics is reinforced since the pattern also emerges in the motif analysis.