Mobilidade de Turistas Internacionais: Uma Comparação entre Dados Oficiais e de LBSN

Lucas E. B. Skora[1], **Helen C. M. Senefonte**[1,2], **Myriam R. B. S. Delgado**[1], **Ricardo Lüders**[1], **Thiago H. Silva**[1,3]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Curitiba, Brasil

[2]Universidade Estadual de Londrina (UEL)
Londrina, Brasil

[3]University of Toronto
Toronto, Canada

lucasskora@alunos.utfpr.edu.br, helen@uel.br

{myriam,luders,thiagoh}@utfpr.edu.br

Abstract. *Studying the behavior of tourists is strategic for improving services in this competitive economic segment. Current work often explores this issue using traditional data such as questionnaires. This type of source provides valuable information; however, they suffer from scalability and coverage. An alternative source that minimizes these problems is obtained by location-based social networks (LBSNs). Nevertheless, for the proper use of these data, it is necessary to verify whether the behavior captured in these networks satisfactorily reflects the real behavior measured with traditional data. Thus, the present work aims to validate whether the international flow of tourists captured with an LBSN satisfactorily reflects the real behavior measured with traditional data. Initial results suggest that the LBSNs data represent remarkably well the behavior studied and that they can enable research on the mobility of these tourists.*

Resumo. *O estudo do comportamento de turistas é estratégico para melhoria dos serviços nesse competitivo segmento econômico. Trabalhos atuais geralmente exploram essa questão usando dados tradicionais, como questionários. Esse tipo de fonte fornece informações valiosas, no entanto, sofre com escalabilidade e abrangência. Uma fonte alternativa que minimiza esses problemas é obtida pelas redes sociais baseadas em localização (LBSNs). No entanto, para o uso apropriado desses dados é necessário averiguar se o comportamento capturado nessas redes reflete de maneira satisfatória o comportamento real medido com dados tradicionais. Assim, o presente trabalho visa validar se o fluxo internacional de turistas capturado com uma LBSN reflete de maneira satisfatória o comportamento real medido com dados tradicionais. Resultados iniciais sugerem que os dados LBSNs representam notavelmente bem o comportamento estudado e que podem habilitar pesquisas sobre a mobilidade desses turistas.*

1. Introdução

De acordo com a Organização Mundial do Turismo (OMT), as atividades relacionadas ao turismo geram milhões de empregos, direta ou indiretamente, em todo o mundo

[unw 2021]. Sendo assim, esse tipo de estudo é estratégico para a economia de diversos países. Vários trabalhos desenvolvidos nesse campo de atuação utilizam dados tradicionais, como questionários, considerados o *ground truth* em vários casos. Por exemplo, [Zieba 2017] explora dados de *surveys* na análise de características individuais de turistas austríacos e como elas podem influenciar suas motivações de viagem. Com base em outro tipo de fonte de dados tradicional, [Scuderi and Dalle Nogare 2018] apresenta um estudo sobre os padrões de gastos em cartões de turistas. Utilizando dados obtidos de maneira tradicional pela OMT, [Lozano and Gutiérrez 2018] estudam o fluxo mundial de turistas, obtendo vários indicativos importantes para o desenvolvimento de novas estratégias voltadas ao turismo internacional.

Dados oficiais, como os fornecidos pela OMT, são tipicamente difíceis de serem obtidos, pois abordagens tradicionais para obtê-los possuem baixa escalabilidade, ou seja, demoram a atingir uma quantidade relevante de dados. Para amenizar essa questão, estudos recentes vêm explorando fontes alternativas, como dados de redes sociais baseadas em localização ou LBSNs (*Location-Based Social Networks*) [Zheng et al. 2014, Silva et al. 2019]. A grande quantidade de dados de LBSNs disponíveis para estudos, bem como a possibilidade de desenvolver análises em diferentes granularidades, como de indivíduos ou grupos em áreas específicas, são características determinantes com o potencial de complementar fontes tradicionais de diferentes maneiras [Silva et al. 2019].

Contudo, é fundamental averiguar se o comportamento capturado através de dados LBSNs reflete de maneira satisfatória o comportamento real medido com dados tradicionais para o estudo de determinadas questões. Nesse contexto, a principal questão de pesquisa que norteia este trabalho é: dados de LBSNs podem ser usados para modelar satisfatoriamente o fluxo de turismo internacional em comparação com dados tradicionais fornecidos pela OMT?

Para responder essa questão, o presente trabalho investiga as diferenças e similaridades entre as informações obtidas por dados oficiais da OMT no estudo de [Lozano and Gutiérrez 2018], com as informações obtidas usando dados LBSNs extraídas na pesquisa atual. Os resultados preliminares indicam que dados LBSNs são comparáveis a dados oficiais e representam satisfatoriamente a realidade do fluxo de turismo internacional.

O restante do trabalho está organizado da seguinte forma. Na Seção 2, são apresentados os principais trabalhos correlatos disponíveis na literatura. A metodologia de avaliação e comparação utilizadas nesse estudo são descritas na Seção 3. Na Seção 4, os resultados são apresentados e discutidos. Por fim, as considerações finais são apresentadas na Seção 5.

2. Trabalhos correlatos

Esta seção apresenta alguns dos principais trabalhos relacionados da literatura. [Miguéns and Mendes 2008] faz uma análise de dados da OMT sobre turismo para comparar as distribuições de grau em uma rede formada pelas relações de turismo entre países, usando também o grau ponderado de entrada e saída da rede. Os autores concluem que a distribuição de grau é aleatória, enquanto a rede ponderada é do tipo *scale free* (segue uma distribuição de lei de potência), demonstrando a importância de considerar o peso das arestas em uma rede. Além disso, busca a correlação entre grau e grau ponderado,

mostrando que, para a rede analisada, o grau de entrada tem correlação forte com grau de entrada ponderado, enquanto o grau de saída ponderado cresce quase quadraticamente em relação ao grau de saída.

[Provenzano et al. 2018] estuda estatísticas do turismo na Europa considerando dados da OMT e dados geolocalizados do Twitter. Os autores concluem que há grande sobreposição entre os dois tipos de fonte. [D'Agata et al. 2013] analisa rotas de turistas entre destinos turísticos na Sicília a partir de dados de *surveys* usando várias métricas de análise de redes buscando encontrar as principais regiões e rotas. [Zhou et al. 2016] estuda a rede de comércio internacional construindo um grafo usando apenas o maior parceiro econômico de cada país. A justificativa para esse filtro é que cada país tem poucos relacionamentos comerciais importantes e alguns países concentram grande parte do comércio internacional. Portanto, faz sentido focar apenas nas relações centrais.

[Hawelka et al. 2014] usa dados geolocalizados do Twitter[1] para estudar a mobilidade internacional, alertando para o enviesamento existente nas populações mais jovens. Com o intuito de analisar a relevância dos dados para cada país considerado, foram calculadas taxas de penetração (razão entre número de usuários que vem daquele país pela população do país). Os autores destacam que a penetração não é diretamente proporcional à mobilidade dos usuários. Por exemplo, Estados Unidos tem a maior penetração, mas com pouquíssimos usuários indo para outros países, enquanto a Bélgica e Áustria têm altíssima mobilidade e pouca penetração. Os únicos dois países encontrados com alta penetração e mobilidade foram Singapura e Kuwait. Uma conclusão destacada pelos autores é que, ao analisar a topologia da rede gerada entre nações, os grupos que emergem respeitam em algum nível as fronteiras geopolíticas reais.

[Belyi et al. 2017] defende que a mobilidade humana é muito complexa por existirem diferentes tipos de mobilidade, como visitas turísticas únicas ou migrações permanentes. Além disso, cada fonte de dados pode influenciar um aspecto específico da mobilidade, enfatizando diferentes vieses. O artigo utiliza 3 fontes de dados: Flickr[2] (representando atividades de lazer e visita de pontos turísticos), Twitter (qualquer atividade em um ambiente com acesso à internet, sejam visitas de negócios ou lazer) e dados oficiais de migração da ONU. Os resultados desse modelo "multicamadas" mostram padrões que não são visíveis quando analisados individualmente. Para comparar a rede gerada com outros relacionamentos internacionais relevantes, foram usados grafos que representam dependências coloniais, linguagens comuns e relações de comércio entre países. Os autores concluem que o peso normalizado das arestas segue uma distribuição de frequência com parâmetros similares aos dados do *Twitter* e *Flickr*, mas os dados de migração seguem uma distribuição de frequência diferente.

O estudo da combinação entre dados tradicionais e de mídias sociais também é explorado na literatura. Por exemplo, [Al Baghal et al. 2021] estuda o impacto da assimetria numérica dos dados disponíveis em cada fonte na capacidade de representação das conclusões obtidas pelo uso de bases combinadas. Os autores utilizam a técnica de análise de sentimos para estudar como a diferença na quantidade de dados pode influenciar os resultados de uma pesquisa e concluem que o uso combinado de diferentes fontes tem grande potencial para pesquisas sociais.

[1]https://twitter.com
[2]https://flickr.com

[Lozano and Gutiérrez 2018] estudam a rede global de turismo formada com dados da OMT a fim de obter uma visão sobre sua estrutura e as interações entre os países de origem e de destino. Foram desenvolvidas várias análises nessa direção, proporcionando uma rica visão do fluxo internacional de turistas com dados tradicionais. Os autores mostram que os indicadores calculados pela análise de rede dos fluxos globais de turismo também podem ser usados para enriquecer as informações fornecidas pelas estatísticas de turismo atuais.

O presente trabalho se difere dos demais, pois o objetivo é comparar se informações de fluxo de mobilidade internacional de grande abrangência (considerando todos os continentes), obtidas com dados tradicionais em [Lozano and Gutiérrez 2018], são equiparáveis com dados de LBSNs. Este estudo também é mais amplo na quantidade de métricas consideradas nas comparações, fornecendo uma visão mais rica dessa comparação sob diversos aspectos. Embora existam outras comparações entre essas duas fontes de dados, pelo o conhecimento dos autores, nenhuma delas é extensiva, utilizando métricas diferentes nas análises. Além disso, as comparações existentes não avaliam o turismo internacional de maneira ampla, normalmente focando em alguma região específica.

3. Metodologia de avaliação

Na comparação realizada nesse trabalho, utilizam-se duas bases de dados: i) uma formada por *check-ins* de LBSNs obtidos na rede social geolocalizada Foursquare[3]; ii) outra que engloba dados oficiais dos *surveys* "*Tourism statistics- Arrivals of non-resident visitors at national borders, by nationality*" e "*Tourism statistics-Outbound tourism- rips abroad by resident visitors to countries of destination (basis: arrivals in destination countries)*" da OMT, baseando-se em uma análise feita em [Lozano and Gutiérrez 2018]. A proposta deste trabalho envolve a utilização de dados do Foursquare, sendo que a abordagem de referência utiliza dados tradicionais (considerados o *ground truth*) obtidos do levantamento realizado pela OMT. *Check-ins* geolocalizados formam o conjunto de dados brutos, e a partir destes há um agrupamento por usuário. Este agrupamento servirá de base para a obtenção dos grafos conforme detalhado na próxima seção.

3.1. Geração dos grafos

Para representar o fluxo dos turistas no desenvolvimento desse estudo, são utilizados grafos considerando pares **de-para**, ou seja, **de** um determinado país **para** outro. Para os grafos usados nesse artigo, foram selecionados os países com mais de 1000 *check-ins* registrados, totalizando 117 países na base de dados Foursquare. Nos dados da OMT, há 214 países nos subgrafos que consideram o fluxo de saída de turistas e 148 nos subgrafos do fluxo de entrada de turistas. Para determinar o país de origem de um determinado usuário, utiliza-se o país com o maior número de *check-ins* realizados pelo usuário, que é considerado turista em todos os demais países. Para cada usuário de um determinado país, são contabilizados todos os países diferentes visitados. Um grafo direcionado $G = (V, E)$ é então construído, onde o conjunto V representa os países selecionados e uma aresta direcionada $e_{i,j} \in E$ com peso $w_{i,j} \in \mathbb{N}$ contecta os países $v_i, v_j \in V$ se o número $w_{i,j}$ de turistas vivem em v_i e visitaram v_j (fizeram ao menos um *check-in* em v_j) é maior que zero.

[3]https://foursquare.com.

A partir do grafo G, seguindo a proposta de [Lozano and Gutiérrez 2018], são criados os subgrafos $G_{in,k}$ (entrada de turistas) e $G_{out,k}$ (saída de turistas) segundo as seguintes definições: $G_{in,k} = (V, E_{in,k})$ com $E_{in,k} \subset E$ e $e_{i,j} \in E_{in,k}$ se $e_{i,j}$ está entre as $k \in \{1, 2, 3...\}$ arestas de maior peso entrando em v_j; $G_{out,k} = (V, E_{out,k})$ com $E_{out,k} \subset E$ e $e_{i,j} \in E_{out,k}$ se $e_{i,j}$ está entre as $k \in \{1, 2, 3...\}$ arestas de maior peso saindo de v_i. Nos experimentos, utilizam-se apenas os subgrafos contendo as *top-1*, *2* e *3* arestas de maior peso entrando ou saindo de cada nó (país). Com isso, nesse estudo são obtidos 6 subgrafos dessa rede internacional de turismo.

Cada grafo gerado passa por um processo de classificação de países, considerando algumas grandezas, tais como, centralidade de *PageRank* [Brin and Page 1998] para todos os subgrafos, assim como centralidade de intermediação, força de entrada, força de saída e grau de entrada para o *top-3 out*. Todas essas classificações foram feitas usando as bibliotecas *NetworkX* ou Pandas com a linguagem de programação *Python*, e então comparadas com as do artigo original.

3.2. Clusterização hierárquica

Com o objetivo de encontrar agrupamentos nos subgrafos *top-3 in* e *out*, foi utilizado o algoritmo de clusterização hierárquica com critério de ligação dado pela média das distâncias (*average linkage*) [Tan et al. 2016]. Para tanto, foi utilizada a linguagem de programação *Python* em conjunto com as bibliotecas *numpy*[4], *networkX*[5] e *sklearn*[6].

Entretanto, essas ferramentas não aceitam uma matriz de afinidade como entrada (ou seja, os próprios subgrafos), mas apenas matrizes de distâncias. Portanto, é necessário representar os grafos em matrizes de distâncias. Isso é feito pela normalização das matrizes de adjacência dos grafos. A normalização ocorre por linha, no caso do *top-3 out*, ou por coluna, no caso do *top-3 in*, de forma que a soma dos elementos normalizados $n(w_{i,j})$ das linhas ou colunas da matriz resulte em 1. Posteriormente, cada elemento da matriz de distâncias é atualizado com o valor $1 - n(w_{i,j})$, transformando afinidade normalizada em distância normalizada. Para tornar a comparação mais simples, a linha de corte do dendrograma é definida de modo a gerar o mesmo número de grupos (*clusters*) de [Lozano and Gutiérrez 2018], excluindo grupos com apenas um país.

3.3. Fluxos intercontinentais

Além da análise do fluxo internacional, foi desenvolvido também um estudo sobre o fluxo intercontinental de turismo. Os continentes considerados foram América do Norte/América Central, América do Sul, Europa, África, Ásia e Oceania. Da mesma forma que na análise internacional, utilizaram-se de grafos direcionados, agora com notações correspondentes aos continentes $G_{cont,in} = (V_{cont}, E_{cont})$ onde V_{cont} são os continentes listados e o peso de $e_{i,j} \in E_{cont,in}$ é igual à soma dos pesos das arestas no subgrafo *top-3 in* tal que o país de origem é do continente v_i e o país de destino do continente v_j. As transições dentro do mesmo continente foram mantidas. Para fins de comparação com os dados da OMT, os dados LBSNs foram normalizados dividindo-se o peso de cada aresta pela soma Helen do peso de todas as arestas, obtendo-se assim a porcentagem do fluxo de turismo que ocorre entre cada par de continentes.

[4]https://numpy.org.

[5]https://networkx.org.

[6]https://scikit-learn.org/stable.

3.4. Motifs

O censo de *motifs* é uma análise que, diferente das outras feita nesse trabalho, revela os padrões locais de um grafo, considerando apenas 3 nós e a organização de suas arestas. Nos experimentos, utilizou-se o programa *mfinder* para escolher os *motifs* mais relevantes no grafo de entrada. Este programa usa o *Z-value* da frequência de um *motif* particular em relação à frequência desse *motif* em um conjunto de redes aleatórias de mesmo tamanho. Assim, uma comparação é feita entre os resultados do censo de *motifs* de 3 nós das duas bases de dados através da diferença percentual dos *Z-values* de cada *motif*. Essa métrica foi escolhida pois suaviza as diferenças entre os diferentes tamanhos dos subgrafos gerados pelos dados da OMT (214 nós para os subgrafos *top-3 out* e 148 para os *top-3 in*) e do Foursquare (117 nós).

4. Resultados

Esta seção apresenta os resultados dos experimentos realizados, buscando destacar aspectos dos subgrafos gerados na Seção 4.1, da métrica *PageRank* na Seção 4.2, da classificação dos países sob diferentes métricas na Seção 4.3, da clusterização hierárquica na Seção 4.4, do fluxo intercontinental de turistas na Seção 4.5 e da análise de *motifs* na Seção 4.6. Todos os resultados apresentados nesta seção foram obtidos com a base de dados do Foursquare. A comparação com a base de dados OMT é feita usando os resultados de [Lozano and Gutiérrez 2018].

4.1. Subgrafos gerados

Os subgrafos gerados foram preliminarmente analisados através do uso de ferramentas visuais para grafos. A Figura 1 mostra o resultado para o subgrafo *top-2 out*. Nota-se que os Estados Unidos e a Turquia têm enorme importância na rede. A observação referente ao Estados Unidos é intuitiva, pois espera-se um elevado número de turistas americanos em outros países. Porém, a importância da Turquia não é óbvia. Isso se deve à popularidade desproporcional da rede social nesse país durante o período estudado. Além disso, ainda é possível identificar na Figura 1, a existência de centros "locais" de turismo, tais como Arábia Saudita (SA), Rússia (RU), Reino Unido (GB), Emirados Árabes Unidos (AE), Malásia (MY), França (FR), Alemanha (DE) e Brasil (BR). Comparando essas representações com os dados da base da OMT (resultados de [Lozano and Gutiérrez 2018]), a principal diferença observada é a super-representação da Turquia.

4.2. Centralidade Pagerank

Como descrito na Seção 3, a métrica de centralidade *PageRank* [Brin and Page 1998] foi utilizada para analisar a importância relativa dos subgrafos estudados. A Tabela 1 apresenta os 20 países com maiores *PageRanks*. Os países destacados em negrito estão entre os 20 primeiros países de todos os 3 subgrafos *top-1*, *2* e *3 in* ou todos os 3 subgrafos *top-1*, *2* e *3 out*. Os países em negrito e itálico aparecem em todas as 6 análises.

Os países em negrito dos subgrafos *top-1*, *2* e *3 out* da base Foursquare são EUA, México, Turquia, Chipre, Holanda, Arábia Saudita, Bahrain, Malásia, Tailândia, Alemanha e Reino Unido (11 dos 20 países), e para o banco de dados da OMT (obtidos em [Lozano and Gutiérrez 2018]), EUA, México, África do Sul, Tailândia, Botsuana,

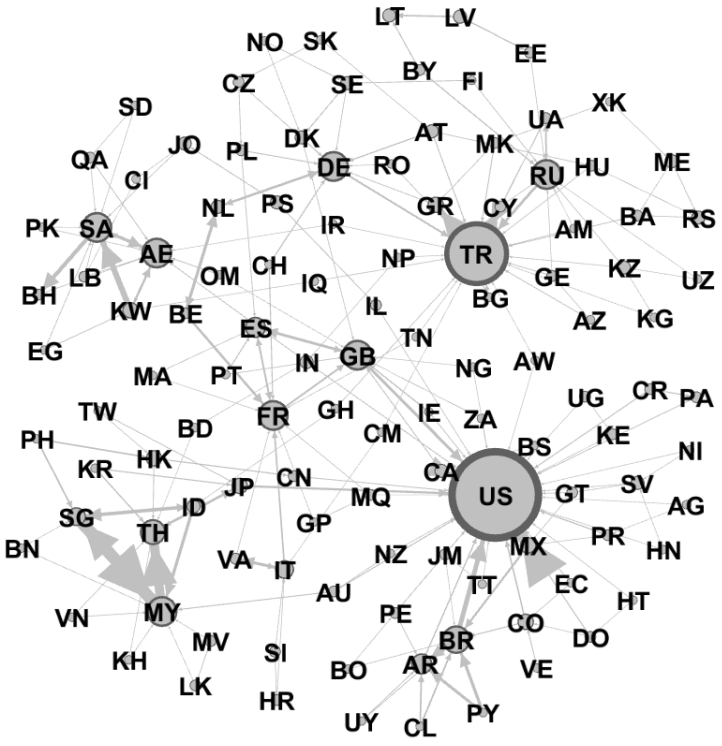

Figura 1. Representação gráfica do subgrafo top 2 out para os dados do Foursquare. Observa-se que o grau de saída de todo nó é 2, mas os grau de entrada variam.

Tabela 1. Países com maior centralidade *PageRank*.

Subgrafos					
Top-1 Out	***Top-2 Out***	***Top-3 Out***	***Top-1 In***	***Top-2 In***	***Top-3 In***
EUA	***EUA***	***EUA***	**Maldivas**	Guadalupe	Martinique
México	**Turquia**	**Reino Unido**	***Holanda***	Martinique	Guadalupe
Turquia	**México**	França	**Bélgica**	**Sri Lanka**	Macedônia
Chipre	Canadá	**Turquia**	**Sri Lanka**	**Maldivas**	Kosovo
Holanda	**Chipre**	Espanha	Peru	Macedônia	**Sri Lanka**
Arábia Saudita	**Arábia Saudita**	**México**	Chile	Kosovo	**Maldivas**
Bélgica	Grécia	Canadá	Portugal	Letônia	Cazaquistão
Bahrain	EAU	**Alemanha**	***Malásia***	Lituânia	Quirguistão
Malásia	**Reino Unido**	**Arábia Saudita**	Brasil	***Holanda***	Quênia
Tailândia	***Malásia***	Itália	***EUA***	**Bélgica**	Uganda
Itália	Espanha	EAU	Turquia	***EUA***	***Holanda***
Vaticano	**Alemanha**	Bélgica	Omã	Sérvia	***Malásia***
Letônia	**Bahrain**	***Holanda***	Arábia Saudita	***Malásia***	***EUA***
Palestina	França	***Malásia***	Áustria	Montenegro	Sérvia
Israel	**Tailândia**	**Chipre**	Argentina	Alemanha	Costa Rica
Lituânia	Brasil	Grécia	Eslováquia	Palestina	Letônia
Alemanha	***Holanda***	**Tailândia**	Nova Zelândia	Singapura	**Bélgica**
Reino Unido	Singapura	Vaticano	Vaticano	Peru	El Salvador
Sri Lanka	Kuwait	Kuwait	Hong Kong	Estônia	Panamá
Maldivas	Japão	**Bahrain**	Palestina	Arábia Saudita	Eslováquia

Malásia, França, Espanha, Ucrânia, Israel, ilhas Maurício, Hong Kong, Benim, Grécia, Etiópia e Filipinas (16 dos 20 países). Com isso, percebe-se que a classificação das centralidades dos subgrafos *top-1*, *2* e *3 out* são mais homogêneos nos dados da OMT do que nos dados do Foursquare.

Além disso, nos 3 subgrafos das duas bases de dados, os Estados Unidos é o país com maior centralidade. A popularidade desproporcional da Turquia na base de dados Foursquare fez com que esse país aparecesse no alto dos 3 índices, além de ter sido suficiente para fazer o Chipre, destino turístico importante para os turcos, também aparecer nos 3 índices. Por fim, nenhum país africano aparece nesses índices, diferente do que acontece nos dados da OMT.

Para os subgrafos *top-1*, *2* e *3 in*, os países em negrito da base Foursquare são Maldivas, Holanda, Bélgica, Sri Lanka, Malásia e EUA (6 de 20), e os da base OMT são EUA, Canadá, China, Hong Kong, Alemanha, Argentina, França, África do Sul, Federação Russa, Ucrânia, e Uzbequistão (11 de 20). Curiosamente, os países com maior centralidade para os subgrafos *top-1*, *2* e *3 in* Foursquare são, respectivamente, Maldivas, Guadalupe e Martinique, com os EUA próximo da décima posição, enquanto os EUA continuam dominando os *top in* da OMT. Isso pode indicar que, pensando em países individuais, os dados Foursquare representam os países que agem principalmente como origens de turistas de forma mais precisa do que os que agem como destinos.

4.3. Classificação de países

A Tabela 2 apresenta os países classificados de acordo com algumas grandezas relacionadas ao destaque desse país no subgrafo *top 3 out*. Os países das 20 primeiras posições em 3 das 4 classificações estão em itálico e em negrito os que estão em todas as 4 classificações.

Tabela 2. Classificações dos países no subgrafo top 3 out.

Posição	grau de entrada	grau de entrada ponderado	grau de saída ponderado	intermediação normalizada
1	**EUA**	**EUA**	*Malásia*	**EUA**
2	**Turquia**	*Malásia*	*Singapura*	**Reino Unido**
3	**Reino Unido**	**Tailândia**	**Turquia**	*França*
4	*França*	**Turquia**	**México**	**Turquia**
5	**Alemanha**	*Singapura*	**EUA**	**Alemanha**
6	*EAU*	**Alemanha**	**Brasil**	**México**
7	*Rússia*	*França*	Kuwait	Bélgica
8	**Tailândia**	**Reino Unido**	*Arábia Saudita*	**Brasil**
9	*Itália*	*Arábia Saudita*	**Reino Unido**	*Rússia*
10	*Malásia*	**Brasil**	Indonésia	**Espanha**
11	*Arábia Saudita*	*Argentina*	**Tailândia**	*Itália*
12	**México**	**México**	Canadá	*EAU*
13	**Espanha**	**Espanha**	Chipre	Omã
14	**Brasil**	*EAU*	Paraguay	**Tailândia**
15	Kuwait	Grécia	Bélgica	Grécia
16	*Argentina*	Chipre	Holanda	Chipre
17	Sérvia	Indonésia	*Itália*	Ucrânia
18	*Singapura*	Canadá	*Rússia*	*Argentina*
19	Japão	Bahrain	**Alemanha**	Japão
20	Polõnia	Holanda	**Espanha**	Índia

Algo em comum entre a base de dados Foursquare e OMT é o destaque dos EUA, mesmo que não ocupe a primeira posição no ranking Foursquare de grau ponderado de

saída. Além disso, mais uma vez, a Turquia tem um destaque desproporcional na base de dados Foursquare. Para aprofundar a comparação, a Tabela 3 mostra as classificações dos países no banco de dados OMT, com a posição nas classificações Foursquare entre parênteses. As células marcadas com NA indicam países que foram ignorados no grafo, pois possuem menos de 1000 *check-ins* registrados em seu território no Foursquare.

Tabela 3. Comparações das classificações dos países no subgrafo top 3 out.

Posição OMT	grau de entrada	grau de entrada ponderado	grau de saída ponderado	intermediação normalizada
1	EUA (1)	EUA (1)	EUA (5)	EUA (1)
2	Malásia (11)	Espanha (13)	Alemanha (19)	França (3)
3	África do Sul (77)	França (7)	Canadá (12)	Grécia (15)
4	Canadá (27)	Ucrânia (32)	China (38)	Chipre (16)
5	Ucrânia (76)	Tailândia (3)	Reino Unido (9)	Espanha (10)
6	Tailândia (8)	Malásia (2)	Singapura (2)	Malásia (26)
7	Grécia (22)	Hong Kong (29)	México (4)	Ucrânia (17)
8	Espanha (13)	México (12)	Rússia (18)	Andorra (NA)
9	Benim (NA)	Canadá (18)	França (21)	Filipinas (64)
10	França (4)	Grécia (15)	Itália (17)	Tailândia (14)
11	Israel (66)	África do Sul (79)	Holanda (16)	África do Sul (54)
12	Brasil (14)	Irlanda (96)	Suíça (31)	Canadá (49)
13	Colômbia (24)	Indonésia (17)	Espanha (20)	Sri Lanka (110
14	Angola (NA)	Brasil (10)	Japão (22)	México (6)
15	Etiópia (NA)	Uzbequistão (82)	Moldávia (NA)	Brasil (8)
16	Mali (NA)	Andorra (NA)	Malásia (1)	Ilhas Maurício (NA)
17	Peru (37)	Peru (51)	Coréia do Sul (29)	Guadalupe (100)
18	Barbados (NA)	Cambodja (87)	Indonésia (10)	Domminica (NA)
19	Botsuana (NA)	Filipinas (42)	Belarus (30)	Hong Kong (30)
20	Ilhas Maurício (NA)	Botsuana (NA)	Portugal (46)	Antígua e Barbuda (87)

Nota-se que alguns países estão sempre em posições próximas nas classificações, usando uma ou outra base de dados, como os Estados Unidos e a Tailândia no *top 3 out*, e outros em posições muito diferentes, como Israel (posição 11 na classificação de grau de entrada no subgrafo *top 3 out* OMT, e 66 no Foursquare), Irlanda (posição 12 no grau ponderado de entrada *top 3 out* OMT, e 96 no Foursquare). Isso provavelmente é causado pela menor quantidade de *check-ins* feitos na base de dados por habitantes desses países menores, tornando o modelo menos preciso nestes casos.

4.4. Clusterização hierárquica

De acordo com a Seção 3.2, os seguintes *clusters* de países foram obtidos:

- **Subgrafo *top 3 in***: {México, EUA, Porto Rico, República dominicana, Costa Rica, Canadá, Jamaica, Haiti, Panamá},{Reino Unido, Espanha, Irlanda, Norway, África do Sul, Portugal}, {Filipinas, Hong Kong}, {Rússia, Cazaquistão, Belarus, Ucrânia, Uzbequistão, Letônia, Lituânia}, {França, Martinique, Guadalupe, Tunísia, Marrocos}, {Trinidade e Tobago, Antígua e Barbuda}, {Argentina, Uruguai, Paraguai, Brasil, Chile, Bolívia}, {Bósnia e Herzegóvina, Montenegro, Sérvia}, {Singapura, Malásia, Brunei, Japão, Indonésia, Vietnam, Tailândia, Taiwan, Coréia do Sul}, {Suécia, Dinamarca}, {Austrália, Nova Zelândia}, {Colômbia, Venezuela, Equador}, {Turquia, Geórgia, Armênia, Kosovo, Romênia, Irã, Chipre, Azerbaijão, Macedônia do Norte, Bulgária, Grécia}, {Uganda, Quênia}, {Kuwait, Iraque, Arábia saudita}, {El Salvador, Guatemala}, {Emirados Árabes Unidos, Maldivas, Omã, Índia, Paquistão, Sri Lanka}, {Croácia, Eslovênia}, {Qatar, Sudão}, {Alemanha, Áustria}, {Bélgica, Países Baixos};
- **Subgrafo *top 3 out***: {Trinidade e Tobago, México, Colômbia, Venezuela, Equador, Aruba, Antígua e Barbuda, Bahamas, EUA, Canadá, Jamaica}, {Singapura, Cambodja, Nepal, Malásia, Japão, Bangladesh, Indonésia, Vietnã, Tailândia, Índia, Coréia do Sul}, {Emirados Árabes Unidos, Egito, Kuwait, Qatar, Omã, Sudão, Bahrain, Paquistão, Arábia saudita}, {Bósnia e Herzegóvina, Quirguistão, Turquia, Cazaquistão, Chipre}, {Finlândia, Geórgia, Armênia, Noruega, Rússia, Suécia, Dinamarca, Ucrânia}, {Peru, Argentina, Uruguai, Paraguai, Brasil, Chile}, {Reino Unido, Costa do Marfim, França, Líbano, Espanha}, {Uganda, Quênia}, {Martinique, Guadalupe}, {Maldivas, Sri Lanka}, {Croácia, Eslovênia, Vaticano, Itália, Montenegro, Sérvia}, {Hungria, Eslováquia, Áustria}, {Suíça, Polônia, Bélgica, República Tcheca, Romênia, Alemanha, Países Baixos}, {Kosovo, Macedônia do Norte, Bulgária, Grécia}, {Palestina, Jordânia, Israel}, {Estônia, Belarus, Letônia, Lituânia}, {Nigéria, África do Sul}, {Porto Rico, República Dominicana}, {China, Filipinas, Taiwan, Hong Kong}, {Austrália, Nova Zelândia}, {Nicarágua, Costa Rica, Panamá}, {Honduras, El Salvador, Guatemala}

Como outros trabalhos já observaram, muitos *clusters* se formam em regiões geograficamente próximas [Hawelka et al. 2014], mesmo que a distância física não seja uma das variáveis usadas para calcular a clusterização. Além disso, as clusterizações encontradas com nossos dados são bastante similares às obtidas com base nos dados da OMT. Existem algumas divergências pontuais nos *clusters* encontrados neste trabalho e em [Lozano and Gutiérrez 2018]. Entretanto, estas divergências não deixam de ser, na maior parte, geopoliticamente consistentes com o esperado: são grupos de países geograficamente próximos e que possuem afinidades, como mesma raiz da língua e hábitos culturais mais fortes. As exceções que ocorrem são esperadas, pois são originadas de fontes de dados diferentes e podem ser atribuídas à falta de dados para países menos centrais no turismo internacional e com menos penetração do Foursquare.

4.5. Fluxos de turistas entre continentes

Para proporcionar uma visão macroscópica do subgrafo *top-3 in*, a Tabela 4 descreve o fluxo de turismo entre continentes em vez de países. Entre parenteses, está o número de arestas (sem peso) correspondentes a esse fluxo.

Neste trabalho, foram considerados 117 países e territórios, dos quais 19 são norte/centro-americanos, 10 são sul-americanos, 38 são europeus, 11 são africanos, 37 são asiáticos e 2 são da Oceania. Como pode-se observar, a África é o continente menos representado na base de dados, seguido da Oceania, embora a população dos outros países do segundo continente seja realmente muito menor.

Tabela 4. Fluxos de turismos entre continentes para o subgrafo top 3 in.

	América do Norte	**América do Sul**	**Europa**	**África**	**Ásia**	**Oceania**
América do Norte	6862(42)	1172(9)	1917(8)	50(4)	1059(9)	116(2)
América do Sul	1215(9)	3057(21)	113(1)	85(1)	3(1)	0(0)
Europa	41(6)	0(0)	8266(77)	269(12)	1687(10)	0(0)
África	0(0)	0(0)	0(0)	19(3)	0(0)	0(0)
Ásia	0(0)	0(0)	7093(28)	458(13)	17389(91)	295(3)
Oceania	0(0)	0(0)	0(0)	0(0)	0(0)	48(1)

A partir da Tabela 4, nota-se que os fluxos de turismo intracontinentais são mais fortes que os extracontinentais. Além disso, a menor quantidade de dados correspondentes a turistas da Oceania e África também se confirmou nos dados da OMT, embora em proporções diferentes. Uma comparação entre os dados do Foursquare e OMT é feita com as diferenças percentuais entre os fluxos com peso e sem peso (entre parênteses) da Tabela 5. A média das diferenças percentuais para o número de arestas é 1,59%, e para o número de turistas é 1,52%.

Tabela 5. Diferença percentual dos fluxos de turismos entre continentes para o subgrafo top 3 in.

	América do Norte	**América do Sul**	**Europa**	**África**	**Ásia**	**Oceania**
América do Norte	0,19% (-1,96%)	2,04% (0,99%)	3,15% (1,60%)	0,03% (-0,43%)	1,32% (-0,13%)	0,20% (-1,00%)
América do Sul	2,29% (1,66%)	4,16% (0,36%)	0,22% (0,28%)	0,16% (0,28%)	0,00% (0,28%)	0,00% (0,00%)
Europa	-0,87% (-3,00%)	-0,05% (-0,67%)	-11,34% (10,47%)	-0,06% (-3,54%)	0,86% (0,6%)	-0,11% (-0,67%)
África	0,00% (0,00%)	0,00% (0,00%)	0,00% (0,00%)	-3,02% (-8,58%)	-0,37% (-0,44%)	0,00% (0,00%)
Ásia	-0,07% (-2,69%)	0,00% (0,00%)	11,75% (6,4%)	0,50% (2,13%)	-11,07% (3,45%)	0,46% (-0,49%)
Oceania	0,00% (0,00%)	0,00% (0,00%)	0,00% (0,00%)	0,00% (-0,22%)	-0,16% (-0,44%)	-0,21% (-4,20%)

4.6. Análise de *motifs*

A análise de *motifs* busca reconhecer os padrões mais comuns na rede, comparando as suas densidades com a densidade dos mesmos padrões em redes aleatórias de mesmo tamanho. A Tabela 6 mostra os resultados produzidos pelo programa *mfinder*. O valor de

C na Tabela 6 é a densidade por mil de cada padrão, e o valor de Z é o *Z-value* em relação a um conjunto de grafos aleatórios de mesmo tamanho. Os motifs marcados como NA não foram identificados como relevantes para aquele subgrafo específico pelo *mfinder*.

Tabela 6. Análise de *motifs* dos subgrafos.

Motif	top-2 out	top-3 out	top-2 in	top-3 in
A $\leftarrow$ B $\rightarrow$ C, A $\rightarrow$ C	C=23,63 Z=7,4 $\pm$ 2,9	C=24,83 Z=20,9 $\pm$ 4,7	C=17,75 Z=12,5 $\pm$ 3,6	C=30,39 Z=3,2 $\pm$ 8,1
A $\leftrightarrow$ B $\rightarrow$ C, A $\rightarrow$ C	C=5,34 Z=0,8 $\pm$ 0,9	C=10,58 Z=3,6 $\pm$ 1,7	C=10,65 Z=4,2 $\pm$ 3,0	C=16,30 Z=12,1 $\pm$ 4,7
A $\leftarrow$ B $\leftrightarrow$ C, A $\rightarrow$ C	NA	C=2,44 Z=2,0 $\pm$ 1,5	NA	C=1,93 Z=12,1 $\pm$ 4,7
A $\rightarrow$ B $\leftrightarrow$ C, A $\rightarrow$ C	C=12,96 Z=2,0 $\pm$ 1,7	C=20,35 Z=9,5 $\pm$ 3,9	NA	C=5,52 Z=3,9 $\pm$ 1,8
A $\leftrightarrow$ B $\leftrightarrow$ C, A $\rightarrow$ C	NA	C=5,29 Z=2,2 $\pm$ 1,4	C=2,54 Z=0,3 $\pm$ 0,5	C=3,59 Z=2,0 $\pm$ 1,2
A $\leftrightarrow$ B $\leftrightarrow$ C, A $\leftrightarrow$ C	NA	C=2,04 Z=0,2 $\pm$ 0,5	NA	NA

Há uma correspondência elevada entre os *motifs* mais relevantes da rede Foursquare e da rede OMT, mas existem apenas 13 padrões possíveis de *motifs* de 3 nós. O único padrão que foi identificado como relevante para um subgrafo Foursquare e para nenhum subgrafo OMT é a tríade completa (última linha da Tabela 6), o que pode indicar que ele tem uma relevância maior na rede Foursquare.

Buscando fazer uma comparação mais objetiva entre os resultados com as duas bases de dados, foi feita uma comparação dos *Z-values* para os *motifs* nos grafos *top-3 in* e *out* das duas bases, apresentados na Tabela 7. A média dos valores absolutos das diferenças percentuais para o grafo *top-3 in* é 32,86% e para o *top-3 out* é 216,77%. Isso indica que, pensando em relações entre poucos países, as redes sociais sugerem representar melhor os países que são visitados do que os países que visitam outros países.

Tabela 7. Diferenças percentuais entre os *Z-values* das duas bases de dados.

Motif	top-3 in	top-3 out
A $\leftarrow$ B $\rightarrow$ C	-12,6%	-103,33%
A $\rightarrow$ B $\rightarrow$ C	-27,90%	-224,29%
A $\leftrightarrow$ B $\rightarrow$ C	15,90%	55,92%
A $\rightarrow$ B $\leftarrow$ C	-38,92%	-792,67%
A $\rightarrow$ B $\leftarrow$ C, A $\rightarrow$ C	10,19%	-818,66%
A $\rightarrow$ B $\leftarrow$ C, A $\leftrightarrow$ C	46,28%	-61,11%
A $\leftrightarrow$ B $\leftarrow$ C	-14,55%	-67,18%
A $\leftrightarrow$ B $\leftrightarrow$ C	-31,84%	71,74%
A $\rightarrow$ B $\rightarrow$ C, A $\leftarrow$ C	-25,00%	-166,67%
A $\rightarrow$ B $\rightarrow$ C, A $\leftrightarrow$ C	40,74%	-80,00%
A $\leftarrow$ B $\rightarrow$ C, A $\leftrightarrow$ C	28,21%	-258,95%
A $\leftrightarrow$ B $\leftrightarrow$ C, A $\rightarrow$ C	35,00%	-4,55%
A $\leftrightarrow$ B $\leftrightarrow$ C, A $\leftrightarrow$ C	-100,00%	100,00%

5. Conclusão

A indústria de turismo urbano se tornou essencial para várias economias, produzindo renda de mais de 1 bilhão de dólares americanos só em 2019, segundo a OMT. Nesse contexto, estudar os turistas e seus comportamentos é essencial para incentivar e aprimorar esse segmento industrial. O campo específico da mobilidade de turistas ainda é pouco estudado, especialmente em larga escala devido a dificuldade em construir bases de dados apropriadas. Fontes tradicionais para o estudo de movimentação de turistas, como *surveys* e pesquisas, sofrem com a baixa escalabilidade, e muitas vezes não têm um nível de abrangência relevante. Sendo assim, surge a necessidade de investigar fontes alternativas para estudar o mesmo fenômeno.

Neste trabalho, identificou-se que dados de redes sociais baseadas em localização (LBSNs), especificamente do Foursquare, são comparáveis a dados oficiais, e representam satisfatoriamente a realidade do fluxo de turismo internacional. É importante ressaltar que os pontos fortes dos dados LBSNs não são necessariamente as análises macro, pensando em termos de países ou regiões como unidades. Com essas fontes, é possível trabalhar no nível de indivíduos e/ou endereços específicos, permitindo maior detalhamento quando comparado ao obtido com outras fontes. Isso se dá mesmo com algumas limitações já conhecidas de LBSNs, como predominância de alguns grupos demográficos - principalmente jovens com acesso regular à internet [Silva et al. 2019]. Embora sejam comparáveis em várias aspectos, há diferenças esperadas entre os resultados do *survey* e do Foursquare. No estudo realizado, as diferenças mais acentuadas foram observadas nas análises que classificam cada país individualmente, como a centralidade *PageRank*. Essas diferenças ocorreram principalmente para países menores e menos centrais para o turismo internacional, para os quais há menos dados. As análises que levam em conta o grafo como um todo, como o censo de *motifs*, tiveram resultados bem mais próximos. Assim, os resultados preliminares podem indicar que, embora não seja capaz de representar completamente um país, o caráter geral da rede de turismo internacional é replicada nos dados de redes sociais.

Uma conclusão desse trabalho indica que, além de prover excepcional acesso aos detalhes, dados de *LSBNs* não deixam de ser úteis para análises macroscópicas, levando em conta o viés para países mais centrais e/ou com maior penetração pela rede social em questão.

Esse resultado abre um leque de novas oportunidades para a ampliação e complementação dos estudos de movimentação de turistas utilizando fontes tradicionais, como os dados da OMT. Novas aplicações e serviços, por exemplo, podem se beneficiar desse recurso para o desenvolvimento de soluções inovadoras para atender ao competitivo turismo global.

Referências

(2021). *International Tourism Highlights, 2020 Edition*. World Tourism Organization, Madrid, Spain.

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

Belyi, A., Bojic, I., Sobolevsky, S., Sitko, I., Hawelka, B., Rudikova, L., Kurbatski, A., and Ratti, C. (2017). Global multi-layer network of human mobility. *International Journal of Geographical Information Science*, 31(7):1381–1402.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

D'Agata, R., Gozzo, S., and Tomaselli, V. (2013). Network analysis approach to map tourism mobility. *Quality & Quantity*, 47(6):3167–3184.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

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

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

Provenzano, D., Hawelka, B., and Baggio, R. (2018). The mobility network of european tourists: a longitudinal study and a comparison with geo-located twitter data. *Tourism Review*.

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

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

Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining*. Pearson Education.

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

Zhou, M., Wu, G., and Xu, H. (2016). Structure and formation of top networks in international trade, 2001–2010. *Social Networks*, 44:9–21.

Zieba, M. (2017). Cultural participation of tourists–evidence from travel habits of austrian residents. *Tourism Economics*, 23(2):295–315.