Dietz, Linus W.; Sen, Avradip; Roy, Rinita; Wörndl, Wolfgang

Article — Published Version

Mining trips from location-based social networks for clustering travelers and destinations

Information Technology & Tourism

Suggested Citation: Dietz, Linus W.; Sen, Avradip; Roy, Rinita; Wörndl, Wolfgang (2020) : Mining trips from location-based social networks for clustering travelers and destinations, Information Technology & Tourism, ISSN 1943-4294, Springer, Berlin, Heidelberg, Vol. 22, Iss. 1, pp. 131-166, https://doi.org/10.1007/s40558-020-00170-6

Information Technology & Tourism (2020) 22:131–166
https://doi.org/10.1007/s40558-020-00170-6

ORIGINAL RESEARCH

Mining trips from location-based social networks for clustering travelers and destinations

Linus W. Dietz[1] · Avradip Sen[1] · Rinita Roy[1] · Wolfgang Wörndl[1]

Received: 23 June 2019 / Revised: 3 January 2020 / Accepted: 21 January 2020 /
Published online: 29 January 2020

Abstract

It is important to learn the characteristics of travelers and touristic regions when trying to generate recommendations for destinations to users. In this work, we first present a data-driven method to mine trips from location-based social networks to understand how tourists travel the world. These trips are quantified using a number of metrics to capture the underlying mobility patterns. We then present two applications that utilize the mined trips. The first one is an approach for clustering travelers in two case studies, one of Twitter and another of Foursquare, where the pure mobility metrics are enriched with social aspects, i.e., the kinds of venues into which the users checked-in. Clustering 133,614 trips from Twitter, we obtain three distinct clusters. In the Foursquare data set, however, six clusters can be determined. The second application area is the spatial clustering of destinations around the world. These discovered regions are solely formed by the mobility patterns of the trips and are, thus, independent of administrative regions such as countries. We identify 942 regions as destinations that can be directly used as a region model of a destination recommender system. This paper is the extended version of the conference article "Characterisation of Traveller Types Using Check-in Data from Location-Based Social Networks" presented at the 26th Annual ENTER eTourism Conference held from January 19 to February 1, 2019 in Nicosia, Cyprus.

Keywords Mobility modeling · Cluster analysis · Spatial clustering · Recommender systems

1 Introduction

Analyzing the mobility of travelers reveals a lot of information about their behavior, preferences, and the destinations they visit. This is interesting for a number of different purposes. Municipalities can obtain information about the popularity of destinations

✉ Linus W. Dietz
linus.dietz@tum.de

[1] Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

within their district to build infrastructure and provide services in an informed way. Destination marketers can learn more about the context of their prospective guests and make improved offers to attract more visitors. Tourist agencies or travel recommender systems can characterize their clients and suggest serendipitous, yet accurate destinations to visit. Finally, prospective travelers can benefit from useful recommendations when planning their trips.

Tourist mobility can be observed in different ways. Analyzing the number of accommodation bookings in a city, tracking ticket sales of flights or trains, or analyzing the congestion of highway connections only captures aggregate travel patterns of one destination or the connections between them. To provide insights into individual travel, we analyze the movement of individual travelers with data from location-based social networks (LBSNs). Our definition of LBSNs follows the one of Roick and Heuser (2013), which includes both social networks that allow the geotagging of contents, such as Twitter or Flickr, as well as geosocial networking sites, such as Foursquare.

The basic idea of our approach is to chronologically sort all of a user's geotagged content into a stream of check-ins and to segment it into periods of being at home and of travel. Consecutive check-ins outside of one's home are combined into a trip that can be characterized using different metrics (Dietz et al. 2018a). Such trips can be used to find the destinations visited together, to derive the durations of stays (Dietz and Wörndl 2019), to cluster them into discernible groups (Dietz et al. 2018b), or to discover larger travel regions (Sen and Dietz 2019).

In this paper, we refine and extend the aforementioned approaches. After reviewing the most relevant literature on the respective topics, we describe three major contributions in the following sections: First, we thoroughly describe the method for deriving trips from various LBSNs, such as Foursquare, Twitter and Flickr. We describe metrics to quantify the quality of trips and compare the mobility metrics of several data sets in Sect. 3. Second, we extend the approach of Dietz et al. (2018b) to cluster the mined trips by social aspects and a second case study on a Twitter data set in Sect. 4. Third, in Sect. 5, we present a novel approach to transform the mobility patterns manifested in trips into a network of tourist flows and use a community detection approach to discover tourist destinations that are defined by actual tourist mobility as opposed to political and administrative boundaries. Finally, we conclude our findings in Sect. 6.

The contributions of this paper can be used to improve the personalization of destination recommender systems. By observing a large number of travelers, we learn how real users travel and should be able to make more realistic recommendations. The spatial clustering of cities to larger travel regions can be directly used as a region model within a destination recommender system that was formerly dependent on political boundaries (Wörndl 2017; Dietz 2018).

2 Related work

The motivation of this work is to improve several aspects of tourist recommender systems. For making good travel recommendations, we need deeper insights as to how people travel, what types of travelers exist, and which regions one should travel

to. This section discusses the literature on tourist recommender systems, the analysis of human mobility, characterizing travelers, and previous approaches to discovering and defining travel regions.

2.1 Tourist recommender systems

Individual tourism is a challenging domain for recommender systems due to the substantial complexities of planning an independent trip and the huge economic importance of the travel and tourism industry (Chaudhari and Thakkar 2019). Big commercial player such as Booking, Tripadvisor, and Skyscanner focus on recommending single items, such as hotels, restaurants, and flights. Nowadays, the academic community is more concerned with recommendations of various touristic items, such as attractions, tourist packages, and composite trips (Borràs et al. 2014). Since these items are often not as well defined as hotels or restaurants, collaborative filtering methods have proven to be less suited. In addition, most people usually travel less frequently than e.g., consuming music or movies, making collaborative recommendations even less reliable due to the cold start problem. Instead, content-based and knowledge-based recommendation techniques are often employed (Burke and Ramezani 2011), or in case it is possible, hybridization (Kbaier et al. 2017).

To facilitate the content-based paradigm, users and items need to be placed in the same feature space that allows for the calculation of a similarity metric. This is usually done using a common categorization, and the similarity measure determines the ranking of the recommended items. This categorization problem is nontrivial, since it requires reliable information about both the user and the candidate items. It can, however, overcome the cold start problem (Burke 2007), since, unlike in a collaborative filtering approach, one can design intelligent systems that efficiently and effectively capture the user's preferences (Braunhofer et al. 2014). For example, it is possible to characterize users based on their past trips (Dietz and Weimert 2018), or to define a more elaborate mapping between classes of users and destinations (Sertkan et al. 2017). Also, the information about a user's social network and previously visited places can be used (Bao et al. 2015; Tsai et al. 2019). Dietz (2018) proposed a data-driven destination recommender system that would suggest composite trips of destinations to users. This paper contributes to the outlined ideas by improving the personalization of the recommendations, especially the duration of stay at the respective destination depending on the traveler type. Furthermore, in Sect. 5, we propose an approach to constructing a hierarchical model of travel regions.

2.2 Human mobility analysis

The analysis of human mobility gives insights into various aspects of everyday life. Before the advent of online social networks on GPS-enabled devices, data sources like mobile phone communication records (González et al. 2008), Wi-Fi usage (Zhang et al. 2012), and raw GPS trajectories (Zheng et al. 2009) have been used to analyze individual human mobility. Given today's availability of LBSN data that enriches a pure location trace with further information, such as user-posted content

and the user's social network, much research has been done analyzing mobility using data from Twitter, Foursquare, and other platforms (Hess et al. 2015).

A prominent research objective is the predictability of human mobility. Song et al. found that individual mobility patterns follow reproducible scaling laws (Song et al. 2010a) and described the limits of the extent to which human mobility can be predicted (Song et al. 2010b). More recently, Ouyang et al. (2016) have analyzed mobility data to predict travel trajectories using a deep learning framework. Similar approaches to predicting the next visited place exist for tourists as well (Zheng et al. 2017) The correlation of locations with a social activity that can be studied with LBSN data promises interesting insights into social behavior. Cheng et al. (2011) found recurring daily and weekly patterns of activity and Wang et al. (2011) found a positive pairwise correlation between social connectedness, i.e., the strength of interactions, and mobility. Noulas et al. analyze activity patterns of Foursquare users, such as the spatial and temporal distances between two check-ins (Noulas et al. 2011). They discover place transitions that could well be used to predict or recommend the future locations of users. The general idea behind their approach is quite similar to ours, however, their motivation was to uncover recurring patterns of human mobility, thus the resulting metrics go in a different direction.

LBSN mobility data have been used to improve recommender systems (Bao et al. 2015). Zheng and Xie (2011) studied spatial co-occurrences that can also be used to identify similar users and generate implicit ratings for collaborative filtering algorithms. Bao et al. (2012) matched the travelers in a foreign city to local experts based on their respective home behaviors to improve the accuracy of a point of interest recommender. LBSN data has also been used to capture cross-border movement (Blanford et al. 2015). The authors demonstrate how the movement dynamics of people in a country can be analyzed, however, this study is not about tourists and is limited to one country, Kenya. Hsieh et al. (2012) used past LBSN data to recommend traveling paths, while Zheng et al. (2019) proposed heuristics to approximate the similarity of tourist trips. For this they present solutions to derive the popularity, the proper time of day to visit, the transit time between venues and the best order to visit the places. In contrast to our scenario, the routes contain single points of interest in urban areas and they leave determining durations of stay at one place to future work. Recently, Dietz et al. (2018a) have proposed a metric-based approach that extracts foreign trips from LBSN data. In this paper, they analyze tourist mobility patterns with the goal of investigating the popularity and co-occurrences of tourist destinations in composite trips.

2.3 Tourist roles

The characterization of tourists has been discussed in literature for decades with an increasing level of complexity. One of the first works was Cohen's four different social roles of tourists: the "organized mass tourist", "individual mass tourist", "explorer", and "drifter" (Cohen 1972). Pearce used fuzzy set theory to define 15 different travel roles (Pearce 1982), while McKercher used an approach motivated by cultural sciences to classify tourists based on the importance of cultural motives

when deciding which destination to visit and the depth of cultural experience gathered by the tourist (McKercher 2002). Finally, Yiannakis and Gibson (1992) took a sociological perspective to observe which roles—they identify 17—are enacted by people when they travel; and associated these with different psychological needs.

With such diversity of tourist categorizations in the literature, the best grouping of tourist preferences and needs to improve destination recommendation is unclear. More importantly, none of the existing categorizations have been validated with observational data (Neidhardt et al. 2014), so it is unclear whether the categories apply to real travelers. To address this challenge, Neidhardt et al. developed the *Seven Factor Model* of tourist behavioral patterns (Neidhardt et al. 2014) based on the *Big Five Factor Model* (McCrae and John 1992) from psychology and a factor analysis of the 17 tourist roles proposed by Yiannakis and Gibson (1992); Gibson and Yiannakis (2002). With a destination recommender system in mind, they elicited user preferences through an image classification task, where the users are to pick the most appealing travel-related photos from a collection. The classification of these pictures along the *Seven Factors* has been previously determined using a questionnaire. Thus, the user's selection of images constitutes a personalized mixture of taste model, allowing for content-based recommendation of points of interests that were rated by experts along the *Seven Factors* in the design stage. Continuing this line of research, Sertkan et al. used unsupervised learning to cluster 561 tourist destinations from a rich commercial data set based on 18 motivational and 7 geographical attributes (Sertkan et al. 2017). Using an expert mapping of the *Seven Factors* to these destinations, they could distill associations between destination attributes and the *Seven Factor Model* that indicate travel behaviors. The *Seven Factor Model* relies heavily on expert knowledge, which is a drawback if this information is not available or costly to obtain. To overcome these limitations, Dietz et al. (2018b) propose trips mined from LBSNs to cluster trips into distinct groups using mobility metrics. To obtain good cluster quality, they perform a correlation analysis of the mobility features and identify four important features: the number of countries visited, the duration of travel, the radius of gyration, and the displacement from home. The resulting clusters are "Vacationers", "Explorers", "Voyagers", and "Globetrotters". We improve upon this research by also analyzing domestic trips and compare a pure mobility-based cluster analysis with social aspects, i.e., to which kinds of establishments the travelers have checked-in during their trip.

2.4 Touristic region discovery via community detection

Researchers have already attempted to define regions based on human mobility data for various purposes such as administrative region discovery (del Prado and Alatrista-Salas 2016), topical region discovery (Taniguchi et al. 2015), and political redistricting (Joshi et al. 2009). Closest to our region discovery approach is the work of Hawelka et al. (2014), who aim to find larger regions of mobility by combining several countries. We aim to find touristic regions that are smaller and potentially independent of countries.

There are various algorithms to perform community detection in networks, such as the Louvain method (Blondel et al. 2008), GDBSCAN (Orman et al. 2011), and Infomap (Rosvall et al. 2009). The complexity of these methods is $\mathcal{O}(n \log n)$. GDBSCAN is less flexible compared to the two others, since it requires to use the distance between spatial points to form clusters that are geographically contiguous. This is a limitation that the other two methods do not have, since the weights of the edges can be chosen at the analyst's will. In the end, we decided to use Infomap, since it has been reported that it outperforms the Louvain method in the quality of the communities (Fortunato and Hric 2016), and there is an up-to-date implementation available.[1] This implementation of Infomap can recursively apply the algorithm to the detected clusters to detect hierarchies of clusters. This mitigates the resolution limit problem, where the size of communities depends on the size of the graph. Thus, the Infomap implementation was our choice to be used without modification in the spatial clustering of tourist destinations of Sect. 5.

3 Trip mining

In this section, we explain our method to mine trips from various LBSNs, namely Foursquare, Twitter and Flickr. Geo-tagged posts in LBSNs provide an incomplete view of a user's mobility, since a user's location is only recorded when she decides to share it. However, given the prevalent use of LBSNs on mobile devices, users often leave a nearly continuous spatio-temporal trace behind them. For example, if a user tweets using a mobile device and decides to enable the "Tweet with location" feature, her location will be recorded with every sent tweet. Similarly, if a user checks in at Foursquare venues, her presence at the venue at a particular time is recorded.

These posts can be seen as a continuous stream of check-ins: a check-in is a tuple of the unique identifier of a user, a location, and a timestamp. The precision of a location's coordinates does not have to be exact, but can also be on the granularity of destinations, such as cities or small islands. Since the data set does not include additional metadata, such as user profiles, users' home countries must be solely determined from the check-in stream. Literature lists several strategies for that, such as *Plurality*, the *geometric median*, or *nDays* (Kariryaa et al. 2018). Segmenting users' check-in stream into trips by periods of travel before returning home can then be done; however, the derived trips need to be checked for data quality, as some users might check in rarely, thus, their true location might be concealed.

3.1 Data sets

Human mobility has been of great interest to the scientific community, as it explains a lot about people's habits; however, location data is inherently privacy-sensitive and

[1] https://www.mapequation.org/code.html.

Table 1 Characteristics of the data sets

Feature	Foursquare	Flickr	Twitter
Number of users	266,909	214,204	2,662,741
Number of check-ins	33,263,633	48,469,177	263,926,396
Observation period	2012-04-03–2013-09-17	2001-07-22[a]–2014-04-26	2011-05-16[a]–2019-04-28

[a]Left 0.1% quantile

anonymizing it for research purposes is challenging, since correlating trajectories with single data points introduces many de-anonymization opportunities (de Montjoye et al. 2013). For this reason, location-based social networks are usually quite restrictive towards querying user location and enforce more or less strict API limits. In this paper, we analyze three data sets: a Foursquare data set from 2012/13 (Yang et al. 2015), the YFCC100M Flickr data set (Thomee et al. 2016), and a self-crawled Twitter data set from 2018/19. Bao et al. (2015) list further data sets stemming from LBSNs.

The raw data sets of Foursquare and Flickr are available following the respective references. Furthermore, we published the mined trips from all three data sets with redacted user identifiers and dates to protect the users' privacy.[2] Table 1 shows an general overview of the data sets. In the case of Flickr and Twitter, we sorted the check-ins chronologically and discarded the first 0.1%, since they were very sparse and potentially wrong, such as 1970-01-01 (the Unix timestamp `0`).

3.1.1 Foursquare

Yang has published a check-in data set[3] stemming from Foursquare (Yang et al. 2015). It contains check-in data spanning 18 months (April 2012–September 2013) and 266,909 users at 3,680,126 venues in 77 countries; however, the data set only contains check-ins from the 415 most popular cities on Foursquare and, therefore, does not include data from travelers seeking recreation in the countryside. Table 2 shows how the distribution of the travelers' origin is influenced by the original data collection. The data set is interesting, because it features many users not residing in the Western countries. The large number of users in countries like Turkey and Indonesia is in line with reports on the regional popularity of Foursquare.

3.1.2 Flickr

The YFCC100M data set is described as the "largest public multimedia collection ever released" (Thomee et al. 2016). It comprises 100 million media objects, less than half of them enriched with geotags. The images were uploaded to Flickr

[2] https://github.com/LinusDietz/JITT2020-Mining-Trips-Replication.
[3] https://sites.google.com/site/yangdingqi/home/foursquare-dataset.

Table 2 Distribution of travelers' home country Foursquare

Home country	Fraction of travelers in %
Turkey	15.15
Brazil	14.28
USA	11.14
Japan	10.16
Indonesia	8.31
Chile	5.59
Malaysia	5.20
Mexico	4.71
Russia	2.85
Thailand	1.95
Other	20.66

Table 3 Distribution of travelers' home country Flickr

Home country	Fraction of travelers in %
USA	27.49
Great Britain	8.00
Spain	5.38
France	4.64
Canada	4.03
Germany	3.83
Japan	3.63
Australia	3.56
Italy	2.82
Brazil	2.49
Other	34.14

between 2004 and 2014 and have been published under a Creative Commons license. We only analyze the metadata of geotagged images for our purpose, since we are not interested in the images themselves. The distribution of user origin, cf. Table 3, is more diverse than in the other two data sets, with most users coming from highly developed countries.

3.1.3 Twitter

Twitter has been frequently used to analyze the individual mobility of the platform's users. The reasons for this are that—in contrast to other social media platforms for private communication—the content on Twitter is mostly public. Twitter also offers APIs to query information about it's users, including the approximate location of

Table 4 Distribution of travelers' home country Twitter

Home country	Fraction of travelers in %
USA	60.38
Great Britain	11.39
Japan	4.20
Canada	2.91
Brazil	2.23
Germany	1.96
Mexico	1.72
Netherlands	1.62
Australia	1.51
Spain	1.32
Other	10.77

their tweets if they have enabled sharing the geolocation of their tweets. By querying the timelines of these users, we can follow their movement patterns.

We have continuously collected timelines of Twitter users since mid-2018 to build up a database of 267,853 timelines. These users tweet in all regions of the world, and the individual check-ins are matched to 24,186 cities with over 15,000 inhabitants each using the GeoNames Gazetteer.[4] As can be seen in Table 4, most users come from the United States, where Twitter is highly popular.

3.2 Method details

The chronologically sorted list of the check-ins of each user is segmented into periods of being at home and periods of travel. To determine the home location of the user, we use the plurality strategy, i.e., choosing the city with the highest number of check-ins. While this is the simplest heuristic to compute, literature shows that its accuracy is on par with more sophisticated methods such as the geometric median (Kariryaa et al. 2018). It may, however, be susceptible to the effects of commuting and users who predominately use social media when traveling. To reduce such false classifications, we discard travelers whose check-ins at home are fewer than a predefined threshold, in our case 50%. This threshold is an aggressive reduction of the data set, discarding about 95% of the users whose home city is unclear to us. It is, however, necessary, since incorrect classification of the user's home would have severe consequences on the forthcoming analyses. The 50% cut-off could be lowered for analyses that are not so much dependent on the correct classification of the home location or the home location can be retrieved using other channels, such as a field in the user profile.

[4] http://download.geonames.org/export/dump/readme.txt.

$$\overbrace{\text{Munich } d_0}^{Home} \rightarrow_2 \underbrace{\overbrace{d_3\ Paris\ d_{11}}^{\text{Block}}}_{\text{Trip 1}} \rightarrow_0 \overbrace{d_{12}\ \text{Munich } d_{100}}^{Home} \rightarrow_0$$

$$\rightarrow_0 \underbrace{\overbrace{d_{101}\ Paris\ d_{101}}^{\text{Block}} \rightarrow_0 \overbrace{d_{101}\ \text{New York City } d_{105}}^{\text{Block}} \rightarrow_3 \overbrace{d_{109}\ \text{Washington, D.C. } d_{118}}^{\text{Block}}}_{\text{Trip 2}} \rightarrow_1$$

$$\rightarrow_1 \overbrace{d_{120}\ \text{Munich}}^{Home} \rightarrow \ldots$$

Fig. 1 Example of a user's check-in stream with two trips

Figure 1 exemplifies a check-in stream of a user from Munich. We define a number of continuous check-ins at one location as a block, which can be while traveling or while at home. In our example, the user's check-in stream starts on day 0 (d_0) in Munich and is followed by a block of 9 days (d_3–d_{11}) in Paris. In this case, the block is terminated by a check-in in Munich on the next day (d_{12}). Since Munich is the user's home location, the first trip is considered completed. This trip, thus, only consists of one block.

Staying at home for 83 days, the user is then observed checking-in in Paris on d_{101} and a few hours later in New York City. Since she was located in Munich the day before, it seems quite probable that she traveled from Munich to New York City with a stopover in Paris. The check-in stream shows several check-ins in New York City until d_{105} and continues with check-ins in Washington, D.C. from d_{109}–d_{118}, before the trip is again terminated by the return to Munich on d_{120}. Thus, this trip has a duration of 18 days and consists of three blocks.

The design decision to include short stopovers in the main trip was made due to the fact that stopovers can be often extended for several days without an increase in the flight ticket price. In fact, some travelers actively choose a city as a destination for the stopover; thus making it part of the trip. Having said that, there are some other uncertainties. As previously described, we only know the location of the user if she decides to check-in, tweet, or take a photo with a GPS tag. During a block, this is not much of an issue, since we assume that sequential check-ins at the same location mean that the user has not moved. The transition time, i.e., the time between two blocks, is more important, as it determines the duration the user has been at a location. In Fig. 1, we denote the transition time with a $\rightarrow_t$, where t is the days between the last check-in of the first block and the first check-in of the subsequent block. Trip 2 starts with perfect information: We know where the user was on all days from d_{100} to d_{105}. What we do not know is where she was on d_{106}–d_{108}, because the next check-in was just on d_{109} in Washington, D.C. Regarding the temporal segmentation of the blocks and trips, we follow a conservative strategy, which means that the block is terminated with the last check-in without the transition time. We think this strategy is sound, because it does not involve any speculation about the traveler's location. For example, at the end of trip 2, it is not clear when the traveler flew back from the United States to Munich. All we know that she was in Washington D.C. on d_{118} and

in Munich on d_{120}. Since we do not have any evidence of the user's location, we do not add d_{119} to any of these blocks. The drawback of this is that the sum of the durations of a trip's blocks can be shorter than the duration of the trip.

3.3 Trip quality assurance

Using the aforementioned trip heuristic, one would potentially get a lot of trips comprising of only a single check-in. To filter out typical business trips, we only analyze trips with a minimum duration of 7 days. The maximum duration of the trips is set to 365 days, since longer durations are not considered a "visit" anymore in the recommendations for tourism statistics by the United Nations Department of Economic and Social Affairs (2010). Furthermore, we require the user to display relatively steady check-in behavior during travel. Thus, this section is about metrics that ensure a minimum quality of the check-in behavior.

The check-in frequency shown in Eq. (1) is not robust against a multitude of check-ins on 1 day, which makes it unsuitable for assessing the reliability of the check-in stream. In this regard, the better measure is the check-in density (Eq. 2), as it captures the fraction of days with a check-in during a trip. Thus, it captures how steady the check-in stream is, which is more important than having several check-ins at the same location on 1 day. We exclude trips that fall under the minimum check-in density of 0.2, which means that the user must have checked-in at least once in 5 days on average.

$$\text{Check-in frequency} = \frac{\text{check-ins}}{\text{days}} \tag{1}$$

The minimum value of check-in density should be chosen depending on the use case. For the purpose of analyzing global mobility patterns, we analyzed the consequences of enforcing a minimal check-in density. Figure 2 depicts the cumulative density function of the check-in densities of the trips. Since the curve is smooth and without an obvious "elbow", we set the threshold at 20%, which discards 32.88% of the trips. Recalling our initial goal with this heuristic, we reduced the mean transition time from 9.80 to 3.39 days while still keeping 67.12% of the trips.

$$\text{Check-in density} = \frac{\text{days with check-in}}{\text{days}} \tag{2}$$

3.4 Mobility metrics

Using this data-driven method, we obtain trips from each data set which we summarize in Table 5. The number of trips is the highest in the Twitter data set; the least amount of trips come from Flickr. The ratio of the number of foreign trips to domestic ones is about 1:19. While the metrics from the previous section were all about the quality of the users' check-in stream, the following metrics capture the mobility patterns of the users. We visualize the distribution of all metrics using the empirical cumulative distribution function (ECDF).

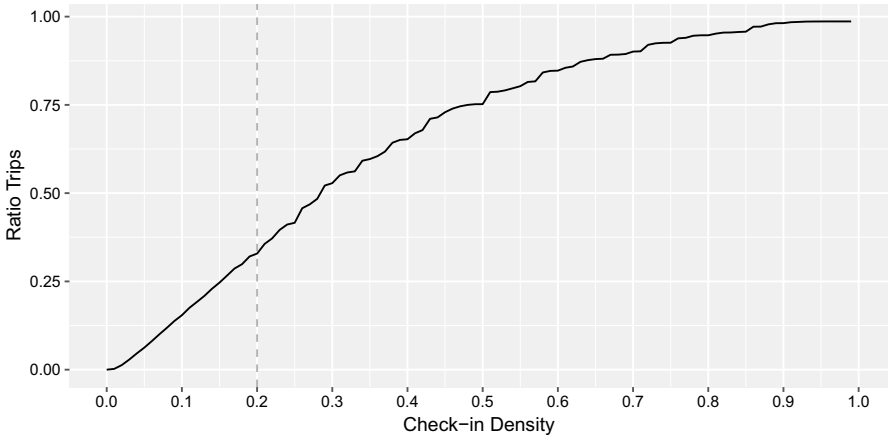


Fig. 2 Empirical distribution function of the check-in densities in the Foursquare data set

3.4.1 Trip duration

The trip duration is the number of days between the first and the last check-in of the trip. Figure 3 shows the cumulative distribution function of the trip durations. One can see the sharp increase in the curves of all three data sets. Overall, 90% of all trips are shorter than 30 days and 62% are shorter than 2 weeks. Flickr has an overall high mean duration of 31.59 days, followed by Twitter with 18.45 and Foursquare. When looking at the median, the data sets are quite similar with a value of 9 Foursquare, 11 for Twitter, and 12 in the case of Flickr. The main reason for the higher mean value of Flickr is that it has more very long trips in comparison to the other data sets. This is an indication that Flickr is used in a different way than the other LBSNs. It might also be that trips are not segmented correctly, possibly due to photographers mostly taking pictures when on travel as opposed to being at home.

3.4.2 Locations, blocks, and countries visited

We also analyzed the number of distinct locations, blocks, and countries within a trip. The number of locations is naturally lower than the number of blocks, since one location can be visited in several noncontiguous blocks of a trip. Figure 4 also reveals that in all data sets most trips span very few countries.

3.4.3 Check-in distance and radius of gyration

Check-in distance measures the mean geographic distance between two consecutive check-ins. This metric is heavily influenced by the check-in frequency. Thus, we prefer to use the radius of gyration for measuring how far the users traveled within a trip. To do so, we follow the definition of González et al. (2008). In simple terms, the radius of gyration measures the mean distance between the mean location of the trip to all other check-ins. It is, thus, more robust against skewed distributions of

Table 5 Trip statistics

Feature	Metric	Flickr	Foursquare	Twitter
Number	Trips	1254	20,317	133,614
Number	Travelers	1254	10,508	23,178
Number	Check-ins	96,111	101,759	2,665,987
Duration	Mean	31.59	11.70	18.45
Duration	SD	54.14	8.14	26.65
Duration	Max	362	222	364
Checkins	Mean	76.64	5.01	19.95
Checkins	SD	138.81	5.20	37.67
Checkins	Max	1063	355	991
Locations	Mean	22.39	3.16	5.27
Locations	SD	58.96	1.92	5.84
Locations	Max	942	25	284
Blocks	Mean	1.83	2.37	1.75
Blocks	SD	3.53	2.08	2.35
Blocks	Max	80	95	165
Countries	Mean	1.43	1.06	1.40
Countries	SD	1.33	0.28	0.93
Countries	Max	22	9	32
Checkin density	Mean	0.41	0.34	0.50
Checkin density	SD	0.22	0.13	0.23
Checkin density	Max	1.00	1.00	1.00
Radius of gyration	Mean	424.46	121.70	703.35
Radius of gyration	SD	1258.09	603.44	1435.82
Radius of gyration	Max	12,735.39	12,871.49	15,834.70
Displacement	Mean	295.19	186.58	1349.55
Displacement	SD	894.25	921.24	2484.86
Displacement	Max	13,183.50	16,279.58	19,430.03

The first three rows showcase the amount of data pruning in comparison to Table 1

check-ins than the check-in distance. In Fig. 5, one can see that the Twitter trips have the largest radius of gyration followed by the Flickr trips and the Foursquare trips.

3.4.4 Displacement

Displacement measures the distance between the user's home location and the mean position of the places visited during the trips. In our data, a similar trend as in the radius of gyration emerges: the Twitter users travel farther than the Foursquare and the Flickr users; however, Twitter shows a clearly slower increase of the distribution function than in the radius of gyration. The reason for this big difference is unclear. Possibly, the socio-economic background of some Twitter users is a different one than of the Foursquare users allowing them to make more intercontinental trips.

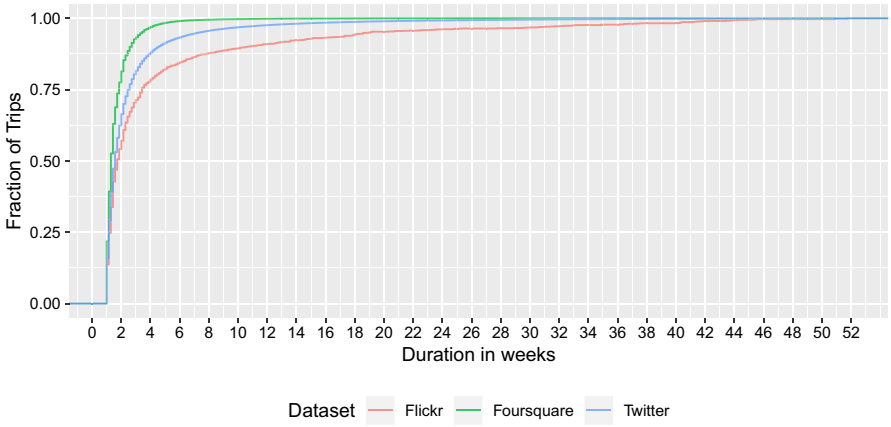


Fig. 3 ECDF of the trip duration

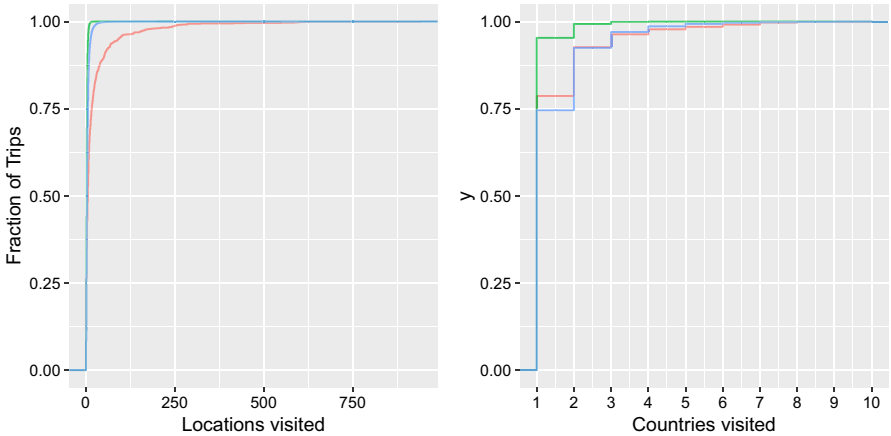


Fig. 4 ECDF of the number of visited locations and countries

3.4.5 Venue information

Finally, we looked at the types of venues checked into according to Foursquare's categorization.[5] Naturally, this information is only available for the Foursquare data set. Out of the ten top-level Foursquare categories, we took a subset of four of the most relevant categories to characterize a trip: Food, which comprises of restaurants and cafés, Nightlife, which are mostly bars and clubs, Arts and Entertainment, which also encompasses all kinds of cultural sites, and Outdoors and Recreation, which are parks and other sports-related sites. As can be seen in Fig. 6, Food is the most

[5] https://developer.foursquare.com/docs/api/venues/categories.

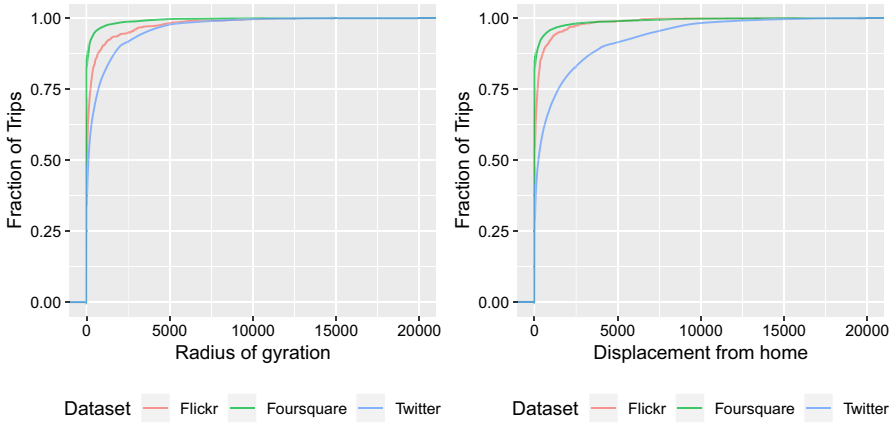


Fig. 5 Left: ECDF of the radius of gyration. Right: ECDF of the mean displacement

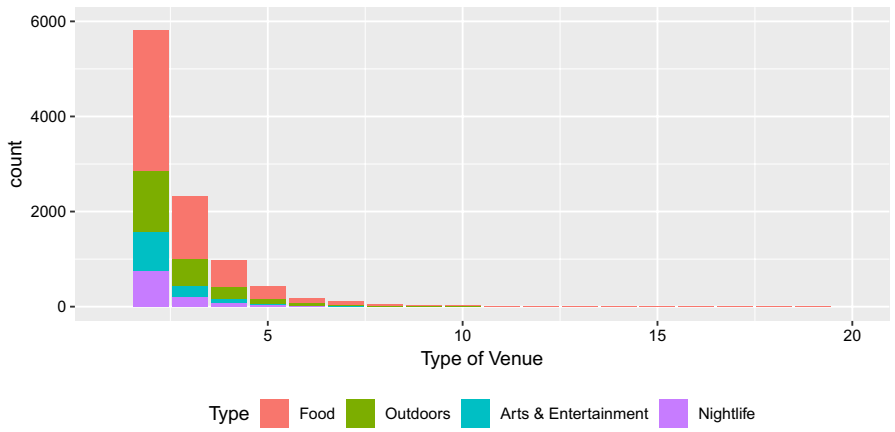


Fig. 6 Types of venues

common venue type, followed by Outdoors, Nightlife, and Arts and Entertainment check-ins. Furthermore, each trip typically has few check-ins that fall under these categories.

3.5 Summary

The proposed approach makes it possible to mine trips from the check-in stream of LBSN users. We have derived a number of metrics for the trips and have distinguished two types of metrics: first, metrics that capture the quality of the data, and second, metrics that capture the underlying mobility of the travelers. Making use of the former ones, we could determine which trips are plausible and sound, whereas the latter ones enable us to do further analyses in this domain.

The algorithms described in this section are implemented in a Python module, published under the permissive MIT License on Github.[6] It provides functionality for parsing the aforementioned data sets and can be extended for parsers of other data sets in about 30–50 lines of Python code.

We learned that not all LBSNs are equally good for such analysis, since users of a geosocial network like Foursquare comprise a different population than Twitter. As a consequence, Twitter users travel further and more often compared to users of Foursquare. Flickr is not well suited for this kind of analysis, since we observe that few people post geotagged pictures of their home. This makes it hard to determine their home locations with the information we have at hand and, thus, making this data set not well-suited for our further applications. Naturally, this sampling of trips based on the respective LBSN limits the generalizability of the findings. The proposed trip mining approach is tailored to LBSN data in large quantities. Since most of the data is not travel-related per se, the heuristics to filter the check-in information for touristic trips will inevitably throw away most of the data points. Furthermore, the data is not suitable for all analyses. For example, short holiday trips of 3–4 days are hard to distinguish from business trips of the same duration. For these reasons, we opted to only analyze long trips of a duration of at least 7 days.

The mined trips can serve as starting points for various improvements to recommender systems. First of all, they show the relative popularity of cities throughout the year. This can be used to increase the diversity of recommendations and, thus, avoid peak season visits for travelers who are sensitive to mass tourism. Furthermore, it shows patterns of destinations that are often visited together. In the composite destination recommendation scenario (Herzog et al. 2019; Dietz 2018), this can provide information on which cities should be combined. Finally, the trip data gives cues on how many destinations should be visited within a trip of a given length and also how long one should stay at each destination (Dietz and Wörndl 2019).

In the next section, we perform a cluster analysis to identify different kinds of trips. This can be useful for distinguishing traveler types in the preference elicitation phase of a travel recommender system. Another application for which we use the mined trips is touristic region discovery, which we describe in Sect. 5.

4 Clustering of traveler types

The first application of the mined trips for use in touristic information systems is a cluster analysis. Cluster analysis is the task of finding groups of data objects, where each group comprises similar objects, whereas the groups themselves are dissimilar to each other. This technique can uncover a structure within unlabeled data and is therefore categorized as unsupervised machine learning (Jain and Dubes 1988).

By revealing different kinds of tourist trips, we can offer insights about the general characteristics of different types of travelers. While this is an analytic result on its own, it can be directly used as part of user modeling within a recommender

[6] https://github.com/LinusDietz/tripmining.

system. But what kind of travelers are there? To answer this, we analyze trips from two data sources: Twitter and Foursquare. In the former, we have a pure mobility trajectory with a check-in granularity of cities, whereas in the latter, the check-ins are attributed to specific venues within a city with further information about the type of the venue.

4.1 Method

In both analyses, we follow the method introduced by Dietz et al. (2018b). The trips are characterized by several features derived from the check-in stream; however, not all metrics are useful for the analysis. Since the goal is to capture the underlying phenomena of the users' travel behavior, we only use the metrics from Sect. 3.4 that capture the users' mobility instead of the quality of data.

Among the remaining metrics, we perform a correlation analysis and remove redundant features, i.e., those whose correlation to another feature is very high. The threshold for this was set at a Pearson correlation coefficient > 0.75. The reason for this exclusion is that highly correlated features will not improve the segregation in the clustering algorithm.

Having decided upon the metrics, we normalize all features using min–max normalization and then run the K-means clustering algorithm with a different number of expected clusters. We evaluate the quality of the determined clusters in terms of the within-cluster sums of squares and the average silhouette (Rousseeuw 1987). The silhouette width measures how well a data object fits into its labeled cluster as opposed to all other clusters. Therefore, it is a robust and easy to interpret method that gives a broad overview of the overall solution quality, as well as information about each data object.

4.2 Case study 1: Twitter trips

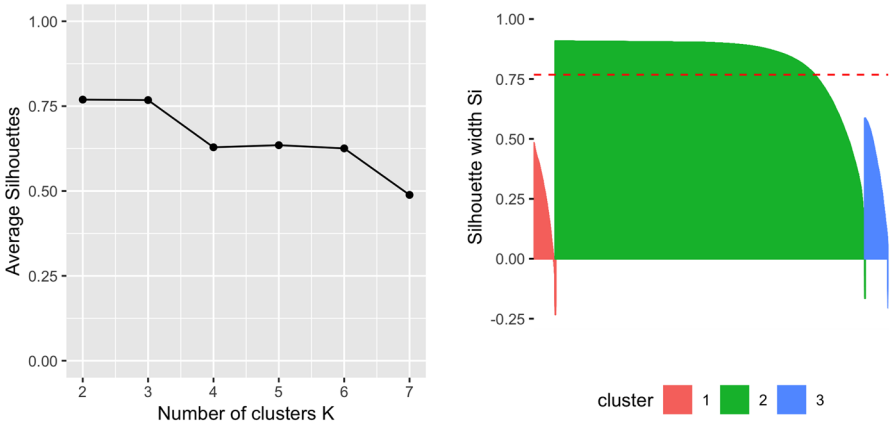
As already mentioned, the trips from Twitter are the most numerous. Due to memory limitations, we drew a random sample of 40,000 trips to run the clustering using K-means clustering. Since this is almost half of the trips, the sample is representative for the overall number of trips.

4.2.1 Features

As already mentioned, this case study is about the pure mobility of the travelers. After the correlation analysis, we could retain the four metrics for capturing the mobility: the duration of the trip, the number of locations, the number of blocks, the radius of gyration, and the displacement from home.

4.2.2 Results

Analyzing the results of the clustering from $K = 2$ to $K = 7$ clusters, there is always one dominant cluster and several smaller ones. According to the silhouette



(a) Average silhouette width (b) Silhouette plot for the three clusters

Fig. 7 Twitter: choosing the best number of clusters

width in Fig. 7a, the solution with two or three clusters has a higher quality, before it decreases to a lower plateau for $K \geq 4$. Thus, we chose the result of $K = 3$ as our final result.

Since we do not discriminate between international and domestic trips, unsurprisingly, most trips from the dominant cluster in green (cf. Fig. 7b) are domestic trips, with a low mean displacement of 565.57 km and only 1.22 countries on average. This result is potentially an outcome of most Twitter users residing in the USA. The two other clusters are smaller in number and more specialized. The Globetrotters travel further, visit the most countries, and display the highest radius of gyration. With 35 days duration, these trips are also the longest. Finally, the Distant Vacationers travel furthest from home, but are not so active during their travel. Their radius of gyration is only one third of the Globetrotters, despite visiting nearly as many distinct locations. Note that these names are only one way to name these clusters. We gave the clusters names to make it easier to refer to them and did our best to choose names that reflect the nature of the trips according to the clustering result (Table 6).

4.3 Case study 2: Foursquare trips

As opposed to the Twitter study and a previous analysis of Foursquare trips (Dietz et al. 2018b), we want to analyze what clusters are formed when taking the activities of the travelers into account.

4.3.1 Features

For this analysis, we use the mobility features of the trips enriched with the type of venues the travelers checked into on Foursquare. This results in the following

Table 6 Twitter: resulting clusters

	Domestic	Globetrotters	Distant vacationers
Ratio	87.1%	6.4%	6.5%
Silhouette width	0.83	0.25	0.38
Duration	19.65/53.45	34.99/96.03	25.05/59.18
Locations	5.13/6.6	9.73/12.35	8.18/7.92
Blocks	1.46/1.83	4.89/6.93	3.34/4.07
Countries	1.22/0.6	3.04/2.01	2.39/1.48
Radius of gyration	329.2/541.36	5172.25/2435.53	1677.74/1428.58
Displacement	565.57/887.83	5262.2/2323.11	8733.89/2834.18

Mean value/standard deviation

features: duration, countries visited, the displacement, the radius of gyration, and the number of check-ins in the categories Food, Nightlife, Arts and Entertainment, and Outdoors and Recreation. No features had to be removed after the correlation analysis.

4.3.2 Results

With this data set the results were more nuanced and a bigger number of clusters was found. The determination of the number of clusters using the silhouette width in Fig. 8a suggested six clusters. Analyzing the results summarized in Table 7 more closely, again a dominant cluster of "Short Domestic" trips arises with 76% of all trips residing in this group. These trips are on average the shortest, have the smallest radius of gyration, and are almost exclusively in the home country of the traveler, since the displacement is on average as small as 40 km. The other clusters are low in number and highly specialized.

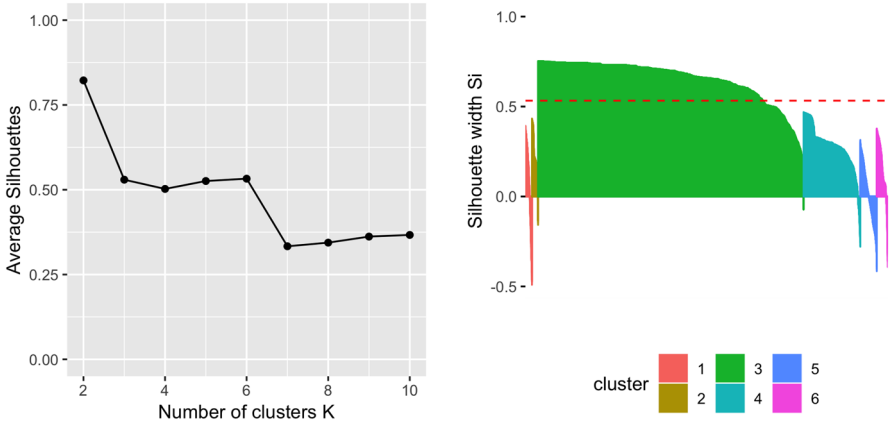
The "Party" trips are about 2-week long trips that visit around four cities in one country. They are distinguished by their high number of food check-ins and very high number of nightlife check-ins.

The "City" trips are quite similar to the domestic short trips, however, they visit more cities and these destinations are more distant to their home town.

The "Foreign" trips are about 2 week long trips to several cities located about 1700 km away from home. People travel quite extensively, as the radius of gyration of about 1300 km indicates.

The "World" trips are quite similar to the "Globetrotters" from Twitter. They visit the highest number of locations, travel the farthest, and also have the highest radius of gyration with nearly 4000 km.

Finally, there is the cluster of the "Long Domestic" trips that last about 6 weeks, which corresponds pretty well to the summer holidays of students at school or university. The small radius of gyration and the high number of outdoors and food



(a) Average silhouette width (b) Silhouette plot for the six clusters

Fig. 8 Foursquare: choosing the best number of clusters

check-ins indicates that these trips might be monothematic vacations, e.g., at a beach resort during summer holidays.

4.4 Summary

This section described a method for finding tourist types using LBSN data. We presented two case studies of international and domestic trips stemming from Twitter and Foursquare. On Twitter, only three clusters emerged, whereas on Foursquare most trips resided in two clusters, with four more very specialized ones.

This method could be used to characterize prospective travelers from data about their past trips. Thus, it can be applied for preference elicitation and user modeling within recommender systems in tourism. Moreover, the analysis requires no user interaction, which is good for the user experience and is also computationally cheap; however, it requires access to the user's check-in history. This can be achieved through an app permission by which the user grants access to their timeline on an LBSN that they have been using, e.g., through a third-party Facebook or Twitter application. Obtaining the data in such a way, a classifier trained with this paper's approach can be used to classify the current user and, thus, be a foundation for providing personalized recommendations.

We have noticed that the cluster analysis results strongly depend on the input data. Developers of recommender systems should carefully evaluate only to include features that are useful for the preference elicitation and the recommendation outcome. Otherwise, the approach is at risk to overfit the data and reports outlier groups as happened in the Foursquare analysis. Finally, it seems to us that this kind of analysis is more suitable for analyzing international trips, as reported by Dietz et al. (2018b).

Table 7 Foursquare: resulting clusters

	Party	City	Foreign	World	Domestic long	Domestic short
Ratio	1.6%	14.7%	3.2%	1.2%	3.2%	76.2%
Silhouette	0.08	0.28	0.01	0.16	0.23	0.65
Duration	16.01/11.85	11.46/5.06	12.87/6.72	17.72/14.89	40.76/22.41	10.3/3.87
Locations	4.36/2.81	3.7/1.8	4.89/2.49	6.48/3.88	6.07/3.52	2.78/1.47
Blocks	3.24/2.8	2.53/1.8	4.14/2.26	5.71/3.67	5.5/6.06	2.07/1.45
Countries	1.06/0.24	1.02/0.14	1.85/0.54	2.39/1.05	1.03/0.17	1.01/0.1
Radius of gyration	107.65 /289.05	47.01 /138.24	1304.51 /963.4	3987.41 /2725.93	75.12 /216.55	27.85 /100.1
Displacement	192.69 /596.07	65.08 /204.52	1692.77 /1184.4	7224.18 /2678.58	95.05 /232.25	39.5 /146.52
Food check-ins	1.92/6.64	0.97/1.6	1.55/2.1	1.94/2.34	3.22/5.07	0.94/1.24
Arts check-ins	0.48/0.94	0.33/0.67	0.35/0.79	0.8/1.31	0.9/2.15	0.28/0.66
Outdoors check-ins	0.72/1.99	0.45/0.94	0.54/0.96	0.87/1.47	2.28/3.54	0.46/0.92
Nightlife check-ins	3.67/1.13	1.23/0.42	0.21/0.49	0.49/1	0.26/0.56	0/0

Mean value/standard deviation

5 Region discovery

The mobility of travelers manifested in the trips can be used for further applications. In this section, we describe a methodology to obtain a map of the world's travel regions that is entirely based on tourist travel behavior instead of political regions. With this approach we aim to uncover implicit tourist regions that are independent of administrative boundaries, e.g., in areas where travel can occur irrespective of national borders, such as the Schengen Area of Europe.

To achieve this, we construct a graph of flows from the trips and use a community detection algorithm to cluster single destinations into coherent travel regions (Sen and Dietz 2019). We use the Twitter data set described in Sect. 3.1.3, as it is the largest, most widespread, and most recent.

5.1 Method

As the Infomap algorithm (Rosvall et al. 2009) uses a weighted graph for community detection, we convert the trips into a graph of flows. Transforming the tourist trips into a graph is relatively straightforward, however, there are several options for quantifying the weights between the nodes.

5.1.1 Community detection

Infomap is a graph community detection algorithm that is designed to discover the underlying structure of the nodes and edges (Rosvall et al. 2009). It can be applied to large directed or undirected graphs and can yield multi-level hierarchies for communities. The algorithm accounts for weights of the edges and, thus, seems to be quite suitable for our application to the weighted flows and distances of tourist movement between cities. The algorithm tries to optimize communities to have more flows within themselves than other communities by using a random walker that traverses the graph. Since it uses a probabilistic model to find communities, the algorithm runs ten times to reduce the probability of obtaining a local minimum. Infomaps picks the best solution according to it's internal quality measure, the description length (Rosvall et al. 2009). In our approach, a community corresponds to a set of cities that form a region. Since the results of Infomap are hierarchical, it will return a tree of regions and subregions, depending on a graph-theoretic termination criterion. Choosing the right granularity of regions depends on the use case.

5.1.2 Graph creation

We transform the trips into an undirected graph, where each city is a node. To form the edges, we try to map the *traveled-together* relation, i.e. that two cities have been visited within the same trip, as closely as possible. The flow between two cities is computed by summing up the co-occurrences of the two nodes in a clique formed by all cities in a trip, over all trips. For example, if a trip consisted of travel from

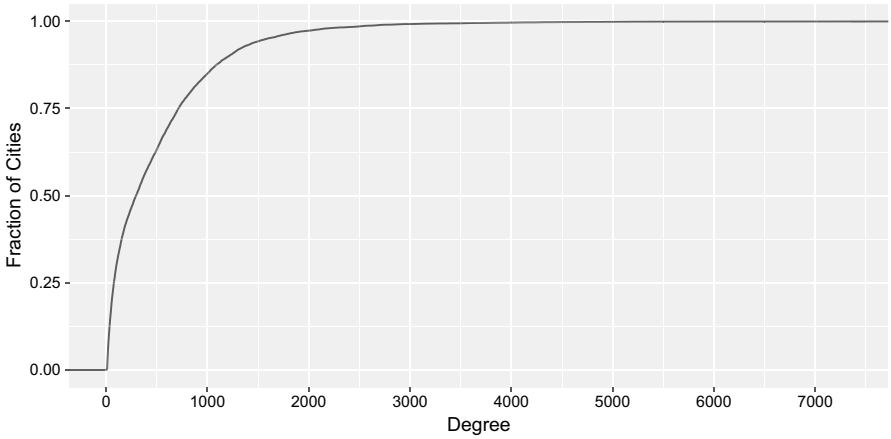


Fig. 9 ECDF of the node degrees

Munich to Berlin via Nuremberg, we would also count the flow from Munich to Berlin. The alternative of not adding this transitive connection to the weight is not appealing for us, since we think the *traveled-together* relation more accurately models the underlying mobility than one-to-one connections. In our example, this would mean that we lose the information that Munich and Berlin have been visited within the same trip. The final weight of each edge is the amount of flow divided by the Euclidean Distance between the two cities. Including the distance in the edge weight reduces the noise in the flow graph introduced by distant traffic hubs, such as airports. Transforming the mobility patterns into this graph-based representation enables us to run the Infomap community detection algorithm to see which cities form coherent clusters.

5.1.3 Graph description

The resulting graph consists of 14,558 nodes and 3,624,909 edges. The degree distribution depicted in Fig. 9 is long-tailed with very few high degree nodes and 87% of nodes having a degree of less than 1000. The graph density of 0.034 also indicates a sparse graph.

5.2 Results

The communities computed by the Infomap algorithm show four top-level regions that align well with existing continental boundaries. These are then further subdivided into a region hierarchy of up to five levels; however, for our purposes, level three and four are the interesting ones. On the second level, many regions are analogous to countries but there are a few interesting variations from this rule. The next level of regions tends to align mostly with travel destinations within federal states. The hierarchy and the discovered regions are discussed in detail in the following.

Table 8 Numerical description of the four top-level regions

Region name	Cities	2nd-level regions	3rd-level regions	4th-level regions
South America	1873	9	193	19
North and Central America	4193	17	254	145
Europe and West Africa	6591	14	381	116
Asia and Oceania	3201	13	114	196
World	15,858	53	942	476

Table 8 gives an overview of the hierarchy of the discovered regions. The South American region consists of the fewest cities, the European cluster has most. The number of second-level regions is small with the average number of cities in each region varying from 208 in South America to 470 in Europe. The third-level regions are clusters of about ten cities in South America, while the mean number of cities in the North American cluster and the European clusters is 17, and even 28 in Asia and Oceania. Most third-level clusters do not contain any sub-regions and most regions at lower-hierarchy levels are very small.

5.2.1 Level 1: continents

The four major regions found at the first level are loosely aligned with existing continental and cultural boundaries (cf. Fig. 10). The division between the Americas is a perfect cut between North and Central America and South America. Africa is under-represented in the data because it has only a few check-ins in Morocco, Algeria, Ghana, Nigeria, Kenya, and South Africa. These countries are merged in the European cluster with the exception of Kenya, which is in the Asian region. The European cluster is merged with all of Russia, Turkey, and the Arabian Peninsula. The Asian region comprises the Indian subcontinent, South East Asia, South Korea, Japan, Australia, and New Zealand.

On this level, the geography and the accessibility of the cities plays a dominant role. This explains the Arabian countries, which belong in the European cluster due to the major aviation hubs. The distance factor introduced to the edge weights seems to have a lower impact than the actual flows within the regions, since even the easternmost parts of Russia are clustered with Europe. Unfortunately, the lack of data from Africa and selected countries in Central Asia hinders the formation of clusters in these areas. We will discuss this limitation at the end of this section.

5.2.2 Level 2: countries

At the second level of the region hierarchy, we found that many regions align with national boundaries; however, there are exceptions observed in each of the top-level clusters.

In Europe a large second-level cluster is found spanning the countries of Germany, Austria, Switzerland, Hungary, the Czech Republic, Poland, and

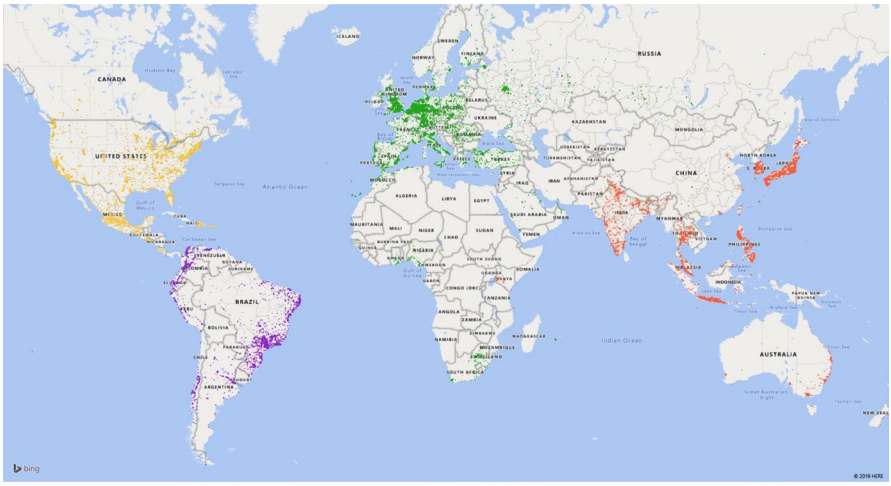


Fig. 10 The top-level regions

Romania (cf. Fig. 11). The Scandinavian countries are clustered together with Russia and the Baltic countries. Italy forms one region with Serbia, however, the unavailability of data from Croatia possibly influenced this result in an unpredictable manner. The Iberian countries are clustered with Morocco, which could be attributed to immigration patterns and the very cheap flight and ferry prices between these countries. Belgium and the Netherlands form another region, and also the British Isles are clustered together. France, Turkey and Greece, however, form regions identical to their national borders.

The second-level regions formed in North America in Fig. 12 mostly disregard national boundaries. Mexico is in one region with other Central American countries, while the USA and Canada are divided into fourteen clusters. The western Canadian states are merged together with Oregon and Washington, while California is split into two major clusters with the southern cluster expanding down to Tijuana and Mexicali in Mexico. Mexican cities on the borders of Arizona and Texas are also members of predominantly American clusters. Several other well-known regions, such as the New England states, Florida and the Great Lakes area form their own clusters.

In Asia (see Fig. 13), the Indian subcontinent and Pakistan are grouped together, which is surprising given the geopolitical context. Australia and New Zealand are clustered together, while the countries in South East and East Asia form their individual regions.

The second-level regions reveal that there are some countries that are traveled to exclusively, but other countries are more frequently traveled to together. In very large countries like Brazil and the USA, we observe a subdivision into multiple subregions at the second level. This is proof that the approach works well for domestic tourism regions.

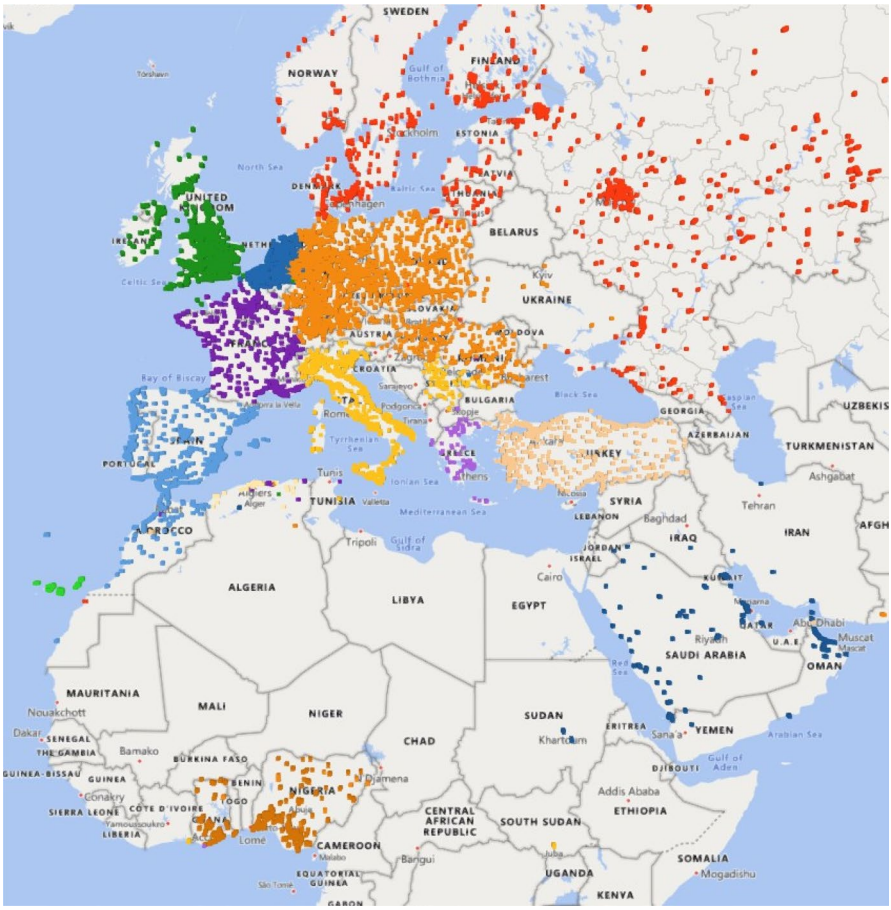


Fig. 11 The second-level community structure of Europe

5.2.3 Level 3: destinations

The regions formed at the third level of the hierarchy can be already seen as tourism destinations; however, the results show varying granularities in different parts of the world with some regions containing further subregions.

The third-level clusters of the big Central European cluster in Fig. 14 are varied in terms of the size and density of cities. The dense regions are typically very contiguous and are centered around a major city. For example, the region containing Munich is comparatively large and includes southern Bavaria. Large areas of the Czech Republic, Poland, and Hungary form homogeneous clusters with no further subregions.

Figure 15 shows that Belgium forms two regions at the third level, however, the Netherlands is divided into six regions that align well to the local divisions.

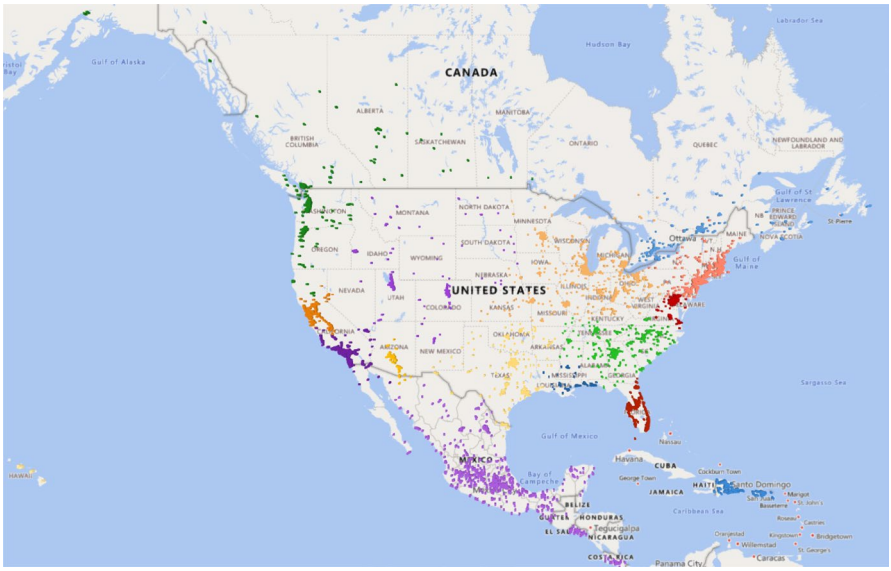


Fig. 12 The second-level communities of North America

Fig. 13 The second-level community structure of South Asia

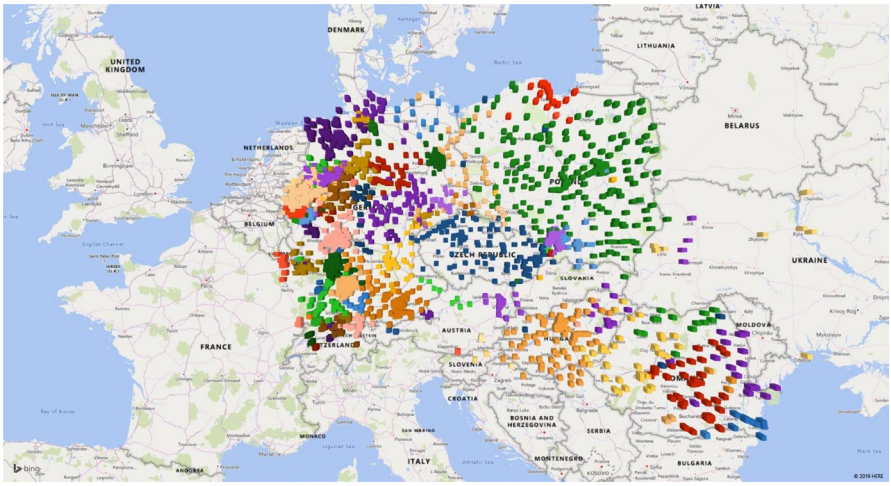


Fig. 14 The third-level community structure of the Central European cluster

Similar contiguous subdivisions are found in the British Isles, Spain, and Italy. The clustering of Morocco, cf. Fig. 16, with Gibraltar and Andalusia is interesting.

The third-level clusters in North America and South America show similar patterns to those in Europe with clusters being centered around cities, as can be seen in Fig. 17.

Pakistan and India are separated at this level with India forming four subregions (Fig. 18). In Thailand, one region is formed by places along the touristically very active coast, while the inland regions are divided into numerous smaller regions. The regions formed in Japan are similar to the political Japanese regions. Australia consists of three regions, one in Western Australia and two in the South East, while New Zealand forms it's own region.

The third-level hierarchical result generally provides regions that can be seen as coherent tourist destinations. At this level, they become small enough to visit them exhaustively within few days and most do not contain further subregions.

5.2.4 Further levels

Some regions are further subdivided, which we discuss for the sake of completeness. In the third-level region of New York State, a fourth-level region with the Burroughs Manhattan, Bronx, Brooklyn, Queens, Staten Island, and Jersey City, which is not a part of New York City, is formed (cf. Fig. 19). Long Island contains two more regions, while four other regions surround New York City. This shows that if the Infomap algorithm obtains sufficient data, it is capable of discovering very fine-grained regions. This example of New York city is an artifact of the high-population density, the municipality structure, and the large amount of Twitter data in this area.

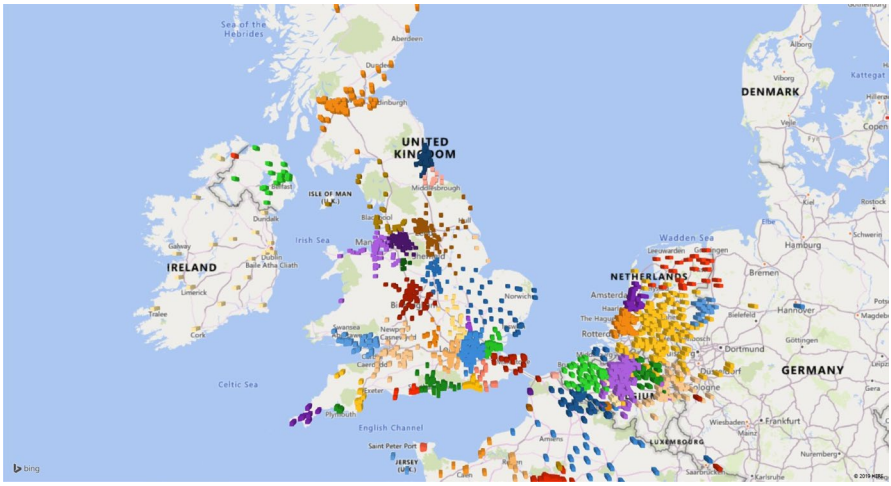


Fig. 15 The third-level community structure of the British and the Benelux cluster

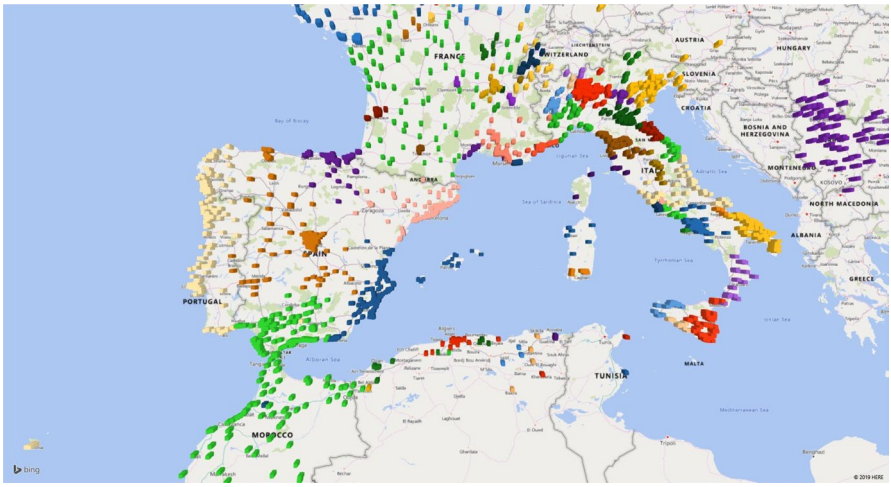


Fig. 16 The third-level community structure of the Iberian and Italian clusters

5.3 Summary

The first three levels of the cluster hierarchy roughly align to continents, countries, and travel destinations. At the second level of the hierarchy, we find that many countries form their own region, while larger and more populous countries, such as the USA and Brazil are subdivided at this level. India stands out, possibly due its lower per capita Twitter usage, and is only subdivided at the third level. Belgium and Netherlands as well as the Iberian countries forming common

Fig. 17 The third-level community structure of south-east USA and Central America

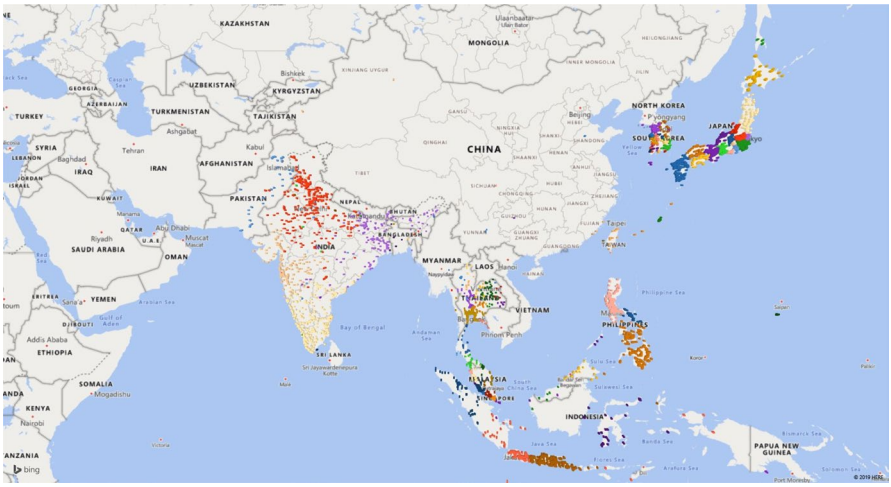


Fig. 18 The third-level community structure of South Asia

regions indicates a tendency for people to travel within countries that have close cultural ties. A similar grouping is formed by the German speaking countries of Austria, Switzerland, and Germany; however, the inclusion of Poland, Hungary, and Romania in the same cluster also underlines the high mobility within the European Union.

Sometimes, we would have expected different borders to be drawn, such as a clear separation at the outer border of the European Union. This is not the case in the countries around the Baltic Sea (cf. Fig. 11), where the Baltic States and

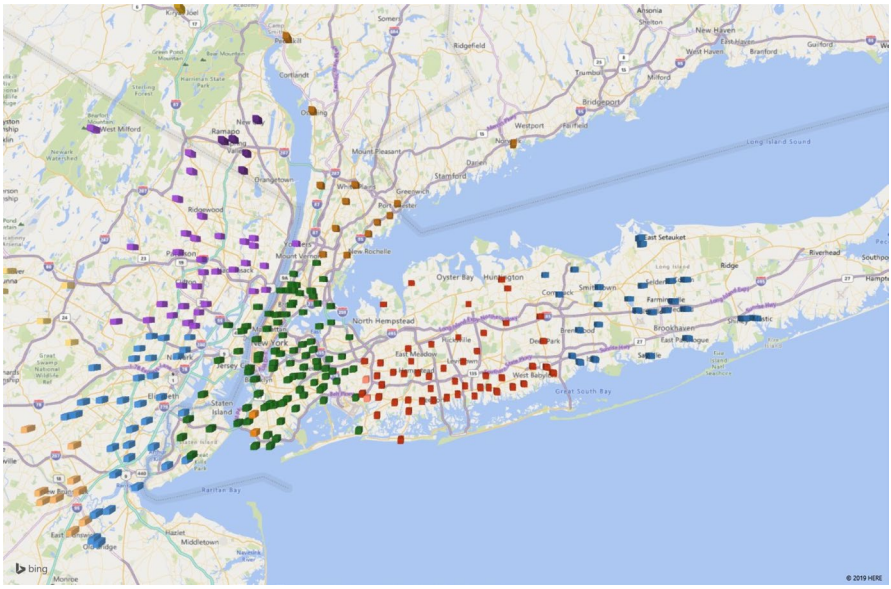


Fig. 19 The fourth-level community structure of New York

Russia are merged together on the second level of Europe. We attribute this to the high number of Russians living in these countries.

In some cases, also a fourth or fifth level exists in the hierarchy. The destination regions containing cities such as Los Angeles and London are found at the third level, while the third-level cluster containing New York consists of multiple fourth-level clusters, with the city of New York forming a cluster that is quite well aligned to the city burroughs. In India, the popular tourist states of Kerala and Rajasthan form fourth level clusters, while the rest of the country is decomposed to tourism destinations. These observations make a termination criterion for subdividing the regions an important problem. Additionally to this, a useful criterion would take into account whether the resulting areas fulfill a certain threshold for with regards to the area, the number of cities, and other metrics relevant to the purpose for the clustering. In our opinion, this cannot be decided with the current data, but requires a use-case-specific analysis of the regions; however, the third-level clusters are already very well defined and form understandable regions.

The stability of the algorithm's output is quite high over several runs. We experimented a lot with Infomap and have not experienced noteworthy changes in the clustering given the same input parameters. This is mostly due to the fact that the algorithm does not recurse to deeper levels if the results become unstable due to insufficient data.

An important limitation of this approach is missing data. If the underlying data source is missing check-ins from a country for any reason, the algorithm does not have good means to counter this. In the case of China, where Twitter is a target of censorship (Bamman et al. 2012), independent clusters simply form around the large

country. In the case of small countries with missing data such as Croatia or Belarus, the algorithm can ignore the missing data resulting in clusters that encompass the area, such as a sea.

Thus, this approach provides a fine-grained map of touristic travel regions in any information system concerning travel. Since the region model is hierarchical, its application scenarios are flexible and developers can pick the hierarchy that suits their needs best.

6 Conclusions

This paper presented three major contributions in the field of tourist recommender systems, mobility analysis and user modeling. The first is a metric-driven method to mine trips from various location-based social networks. We show how to extract domestic and international trips and ensure that the quality is sufficient for further analyses. By comparing several data sets, we find that the users display different mobility behavior on different platforms and that not all platforms are equally suited for this kind of analysis. For example, Flickr users typically have too few geotagged images, which results in only 1254 trips out of over 48 million check-ins.

Second, we present two case studies of cluster analyses of trips from Twitter and Foursquare. The purely mobility-based data set from Twitter revealed three clusters, while the Foursquare data that contained information about the type of venues was segmented into six clusters. This shows that the result is highly dependent on feature selection, which should be accounted for when using this method to classify users.

Finally, we presented an approach for the spatial clustering of touristic regions from Twitter trips. To the best of our knowledge, this is the first application of geo-located tweets to find travel regions with data spanning the whole world. The analysis of results reveals a hierarchy of regions, with tourist destinations residing on the third level. These results confirm that the use of volunteered geographic information to find traveler mobility patterns and define regions based on the patterns is a feasible approach.

The findings of this paper reveal much about how different user groups travel throughout the world. We have established a methodology that extracts travel trajectories from incomplete information sources. Naturally, not all LBSN sources are equally good for different use cases, but we have provided researchers and tourism analysts tools to evaluate this. Working with imperfect knowledge about the travelers' mobility has some limitations. Since we had to filter out many trips due to data quality issues, the results might be biased towards the behavior of travelers who continuously share their location on LBSNs. A generalization of the results should, thus, be done with care. Furthermore, the availability of spatio-temporal user data for independent research purposes is on the decline (Freelon 2018). For this reason, we had to work with two out-of date static data sets in the case of Foursquare and Flickr and had to put a lot of effort into building our own data set of Twitter trips using the official APIs. Other popular LBSNs, such as Facebook, Instagram, or Snapchat do not permit independent content extraction.

The cluster analyses of the trips provide tools to classify users in any tourism information system. This is useful, since knowing the type of traveler can be used to filter which items should be shown to the user. Performing the classification of a user can be done via the analysis of her past travel data, such as booking information or LBSN profiles, but also using a self-assessment of the traveler type. In the future, we plan to implement this in a global destination recommender system for composite trips, thereby extending previous approaches (Dietz 2018; Dietz et al. 2019). The results of the trip clustering approach showed a high dependence on choosing the right features for the given use case. For this reason, the resulting clusters should not necessarily be taken as a generally valid segmentation of traveler types; instead, the proposed method can be applied to determine the groups of users of one's own information system.

The discovered regions provide a hierarchical model of touristic regions. The advantage of this region model is that it is specific to the travel domain and is, thus, the preferable choice for visualizing regions in a travel recommender system over e.g., administrative boundaries. This resolves a problem, where previous systems had to make ad-hoc decisions on how to reasonably split large countries into smaller areas (Herzog and Wörndl 2014; Wörndl 2017). We think that this region model can help users to select their preferred travel destinations, especially in a composite trips scenario (Herzog et al. 2019) and to visualize various trends of global travel in more meaningful ways. While the output of the community detection algorithm itself were quite stable, the results might change with other data sources. Again, by using Twitter as the sole data source, people not active on this platform do not contribute to the region model. This is an important limitation, since this threatens the generalization of the results. In future, the results of different data sets should be compared systematically and also contrasted to official statistics about tourism movement.

Acknowledegments Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bamman D, O'Connor B, Smith N (2012) Censorship and deletion practices in Chinese social media. First Monday. https://doi.org/10.5210/fm.v17i3.3943

Bao J, Zheng Y, Wilkie D, Mokbel M (2015) Recommendations in location-based social networks: a survey. GeoInformatica 19(3):525–565. https://doi.org/10.1007/s10707-014-0220-8

Bao J, Zheng Y, Mokbel MF (2012) Location-based and preference-aware recommendation using sparse geo-social networking data. In: 20th international conference on advances in geographic information systems, ACM, New York, NY, USA, SIGSPATIAL '12, pp 199–208. https://doi.org/10.1145/2424321.2424348

Blanford JI, Huang Z, Savelyev A, MacEachren AM (2015) Geo-located tweets. enhancing mobility maps and capturing cross-border movement. PLoS One 10(6):1–16. https://doi.org/10.1371/journal.pone.0129202

Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 10:1–12. https://doi.org/10.1088/1742-5468/2008/10/P10008

Borràs J, Moreno A, Valls A (2014) Intelligent tourism recommender systems: a survey. Expert Syst Appl 41(16):7370–7389. https://doi.org/10.1016/j.eswa.2014.06.007

Braunhofer M, Elahi M, Ricci F (2014) Techniques for cold-starting context-aware mobile recommender systems for tourism. Intelligenza Artificiale 8(2):129–143. https://doi.org/10.3233/IA-140069

Burke RD (2007) Hybrid web recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web: methods and strategies of web personalization. Springer, Berlin, pp 377–408. https://doi.org/10.1007/978-3-540-72079-9_12

Burke RD, Ramezani M (2011) Recommender systems handbook, chap matching recommendation technologies and domains. Springer, Boston, pp 367–386. https://doi.org/10.1007/978-0-387-85820-3_11

Chaudhari K, Thakkar A (2019) A comprehensive survey on travel recommender systems. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-019-09363-7

Cheng Z, Caverlee J, Lee K, Sui DZ (2011) Exploring millions of footprints in location sharing services. In: Fifth international conference on weblogs and social media, AAAI, Palo Alto, CA, USA, ICWSM '11, pp 81–88

Cohen E (1972) Towards a sociology of international tourism. Soc Res 39(1):164–182

de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the crowd: the privacy bounds of human mobility. Sci Rep 3(1):1–5. https://doi.org/10.1038/srep01376

del Prado MN, Alatrista-Salas H (2016) Administrative regions discovery based on human mobility patterns and spatio-temporal clustering. In: 13th international conference on mobile ad hoc and sensor systems, IEEE, MASS'16, pp 65–74. https://doi.org/10.1109/mass.2016.019

Dietz LW (2018) Data-driven destination recommender systems. In: 26th conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, UMAP '18, pp 257–260. https://doi.org/10.1145/3209219.3213591

Dietz LW, Weimert A (2018) Recommending crowdsourced trips on wOndary. In: RecSys workshop on recommenders in tourism, Vancouver, BC, Canada, RecTour'18, pp 13–17

Dietz LW, Wörndl W (2019) How long to stay where? On the amount of item consumption in travel recommendation. In: ACM RecSys 2019 late-breaking results, pp 31–35

Dietz LW, Herzog D, Wörndl W (2018a) Deriving tourist mobility patterns from check-in data. In: WSDM workshop on learning from user interactions, Los Angeles, CA, USA

Dietz LW, Roy R, Wörndl W (2018b) Characterisation of traveller types using check-in data from location-based social networks. In: Pesonen J, Neidhardt J (eds) Inf Commun Technol Tour. Springer, Cham, pp 15–26

Dietz LW, Myftija S, Wörndl W (2019) Designing a conversational travel recommender system based on data-driven destination characterization. In: ACM RecSys workshop on recommenders in tourism, pp 17–21

Fortunato S, Hric D (2016) Community detection in networks: a user guide. Phys Rep 659(11):1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Freelon D (2018) Computational research in the post-API age. Political Commun 35(4):665–668. https://doi.org/10.1080/10584609.2018.1477506

Gibson H, Yiannakis A (2002) Tourist roles: needs and the lifecourse. Ann Tour Res 29(2):358–383

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271. https://doi.org/10.1080/15230406.2014.890072

Herzog D, Wörndl W (2014) A travel recommender system for combining multiple travel regions to a composite trip. CBRecSys@RecSys. Foster City, CA, USA, pp 42–48

Herzog D, Dietz LW, Wörndl W (2019) Tourist trip recommendations—foundations, state of the art and challenges. In: Augstein M, Herder E, Wolfgang W (eds) Personalized human–computer interaction. de Gruyter Oldenbourg, Berlin, pp 159–182

Hess A, Hummel KA, Gansterer WN, Haring G (2015) Data-driven human mobility modeling. ACM Comput Surv 48(3):1–39. https://doi.org/10.1145/2840722

Hsieh HP, Li CT, Lin SD (2012) Exploiting large-scale check-in data to recommend time-sensitive routes. In: ACM SIGKDD international workshop on urban computing, ACM, New York, NY, USA, UrbComp '12, pp 55–62. https://doi.org/10.1145/2346496.2346506

Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Upper Saddle River

Joshi D, Soh LK, Samal A (2009) Redistricting using heuristic-based polygonal clustering. In: Ninth IEEE international conference on data mining, IEEE, pp 830–835. https://doi.org/10.1109/ICDM.2009.126

Kariryaa A, Johnson I, Schöning J, Hecht B (2018) Defining and predicting the localness of volunteered geographic information using ground truth data. In: Conference on human factors in computing system, ACM, CHI'18. https://doi.org/10.1145/3173574.3173839

Kbaier MEBH, Masri H, Krichen S (2017) A personalized hybrid tourism recommender system. In: 2017 IEEE/ACS 14th international conference on computer systems and applications (AICCSA), pp 244–250. https://doi.org/10.1109/AICCSA.2017.12

McCrae RR, John OP (1992) An introduction to the five-factor model and its applications. Personality 60(2):175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McKercher B (2002) Towards a classification of cultural tourists. Int J Tour Res 4(1):29–38. https://doi.org/10.1002/jtr.346

Neidhardt J, Schuster R, Seyfang L, Werthner H (2014) Eliciting the users' unknown preferences. In: 8th ACM conference on recommender systems, ACM, New York, NY, USA, RecSys '14, pp 309–312. https://doi.org/10.1145/2645710.2645767

Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in Foursquare. In: Fifth international conference on weblogs and social media, AAAI, Palo Alto, CA, USA, ICWSM '11, pp 570–573

Orman GK, Labatut V, Cherifi H (2011) On accuracy of community structure discovery algorithms. J Converg Inf Technol 6(11):283–292. https://doi.org/10.4156/jcit.vol6.issue11.32

Ouyang X, Zhang C, Zhou P, Jiang H (2016) Deepspace: an online deep learning framework for mobile big data to understand human mobility patterns. CoRR abs/1610.07009

Pearce PL (1982) The social psychology of tourist behavior. In: International series in experimental social psychology, vol, 3. Pergamon Press

Roick O, Heuser S (2013) Location based social networks—definition, current state of the art and research agenda. Trans GIS 5(17):763–784. https://doi.org/10.1111/tgis.12032

Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. Eur Phys J Spec Top 178(1):13–23. https://doi.org/10.1140/epjst/e2010-01179-1

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20(1987):53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sen A, Dietz LW (2019) Identifying travel regions using location-based social network check-in data. Front Big Data. https://doi.org/10.3389/fdata.2019.00012

Sertkan M, Neidhardt J, Werthner H (2017) Mapping of tourism destinations to travel behavioural patterns. In: Stangl B, Pesonen J (eds) Information and communication technologies in tourism. Springer International Publishing, Cham, pp 422–434. https://doi.org/10.1007/978-3-319-72923-7_32

Song C, Koren T, Wang P, Barabási AL (2010a) Modelling the scaling properties of human mobility. Nat Phys 6(10):818–823. https://doi.org/10.1038/nphys1760

Song C, Qu Z, Blumm N, Barabási AL (2010b) Limits of predictability in human mobility. Science 327(5968):1018–1021. https://doi.org/10.1126/science.1177170

Taniguchi Y, Monzen D, Ariestien LS, Ikeda D (2015) Discover overlapping topical regions by geo-semantic clustering of tweets. In: 29th international conference on advanced information networking and applications workshops, IEEE, pp 552–557. https://doi.org/10.1109/waina.2015.85

Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) YFCC100M: the new data in multimedia research. Commun ACM 59(2):64–73. https://doi.org/10.1145/2812802

Tsai CY, Paniagua G, Chen YJ, Lo CC, Yao L (2019) Personalized tour recommender through geotagged photo mining and LSTM neural networks. MATEC Web Conf. https://doi.org/10.1051/matecconf/201929201003

United Nations Department of Economic and Social Affairs (2010) International recommendations for tourism statistics 2008. https://unstats.un.org/unsd/tradekb/Knowledgebase/50551/IRTS-2008

Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL (2011) Human mobility, social ties, and link prediction. In: 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD'11, pp 1100–1108. https://doi.org/10.1145/2020408.2020581

Wörndl W (2017) A web-based application for recommending travel regions. In: Adjunct publication of the 25th conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, UMAP '17, pp 105–106. https://doi.org/10.1145/3099023.3099031

Yang D, Zhang D, Chen L, Qu B (2015) NationTelescope: monitoring and visualizing large-scale collective behavior in LBSNs. J Netw Comput Appl 55:170–180. https://doi.org/10.1016/j.jnca.2015.05.010

Yiannakis A, Gibson H (1992) Roles tourists play. Ann Tour Res 19(2):287–303. https://doi.org/10.1016/0160-7383(92)90082-z

Zhang Y, Wang L, Zhang YQ, Li X (2012) Towards a temporal network analysis of interactive WiFi users. Europhys Lett. https://doi.org/10.1209/0295-5075/98/68002

Zheng Y, Xie X (2011) Learning travel recommendations from user-generated GPS traces. ACM Trans Intell Syst Technol 2(1):1–29. https://doi.org/10.1145/1889681.1889683

Zheng W, Huang X, Li Y (2017) Understanding the tourist mobility using GPS: where is the next place? Tour Manag 59:267–280. https://doi.org/10.1016/j.tourman.2016.08.009

Zheng Y, Zhang L, Xie X, Ma WY (2009) Mining interesting locations and travel sequences from GPS trajectories. In: 18th international world wide web conference, ACM, New York, NY, USA, WWW'09. https://doi.org/10.1145/1526709.1526816

Zheng W, Zhou R, Zhang Z, Zhong Y, Wang S, Wei Z, Ji H (2019) Understanding the tourist mobility using GPS: how similar are the tourists? Tour Manag 71:54–66. https://doi.org/10.1016/j.tourman.2018.09.019