

# Assessing the Usability of Georeferenced Tweets for the Extraction of Travel Patterns: A Case Study for Austria and Florida

Hartwig HOCHMAIR and Sreten CVETOJEVIC

University of Florida, Fort Lauderdale / USA · hhhochmair@ufl.edu

This contribution was double-blind reviewed as full paper.

## Abstract

An understanding of people's travel behavior is important for a functional design of transportation networks. This paper explores the use of georeferenced tweets for extracting aggregated travel patterns, i.e. describing the routes that people travel on a given day from origin to destination. The focus is on terrestrial long-distance travel, expanding over more than 100km. The study uses georeferenced tweets collected over four weeks for a test region in Austria and one in Florida. It applies selection filters to extract tweets that contain potentially useful information about users moving between different cells of the test regions. Further the mean travel direction for each grid cell is computed for different days and analyzed. The study also explores the use of a space-time permutation model to identify spatio-temporal clusters of tweets and their change over time.

## 1 Introduction

Different types of Volunteered Geographic Information (VGI) (GOODCHILD 2007) have been recently used to identify clusters of human activities and travel behavior in urban environments. This includes the analysis of geotagged user generated content, such as shared photos, videos, blog posts or tweets, as well as resources of social networking, such as Foursquare. Most of the studies on VGI based analysis of travel trajectories focus on urban regions (GIRARDIN et al. 2008, NOULAS et al. 2012, RÖSLER & LIEBIG 2013). Metropolitan regions are typically also the focus of travel household surveys that obtain people's travel behavior on a specific day (PARTHASARATHI et al. 2012). Comprehensive transportation planning, however, also requires an understanding of people's regional travel behavior that extends beyond cities. Thus, this study will assess the feasibility of extracting regional travel behavior through the spatio-temporal analysis of tweets. While only about 1 percent of tweets are georeferenced, Twitter data has some advantages over other VGI data sources. As opposed to data contributed to photo-sharing portals, the location of tweets is not tied to scenic places, since tweets are typically also sent from locations of everyday activities. Furthermore, tweets are not restricted to pre-defined check-in locations, such as with Foursquare. The goal of this study is to assess whether the spatio-temporal information associated with georeferenced tweets can be used to describe people's long-distance travel behavior on different days of the study period. This includes the analysis of the direction of

aggregated travel between region grid cells, and the identification of changes in clustering on different days, as found through a space-time permutation model. Besides VGI, cell phone data and GPS based travel surveys have also been previously used to explore travel behavior, which, however, typically requires the purchasing of data, or equipping probe vehicles with GPS receivers.

So far only a few studies used Twitter data for the analysis of travel patterns. For example ANDRIENKO et al. (2013) constructed trajectories from footprints of individual Twitter users for the greater Seattle area and identified their center locations. These were found to be moderately correlated with population density ( $r=0.52$ ). KRUMM et al. (2013) used a machine learning approach to predict the likelihood of candidate locations to be a travel destination based on Twitter data in the US, combined with demographic characteristics of places shown as previously visited through Twitter. CHENG & LEE (2011) use personal check-in information gathered through the public Twitter feed for travel analysis and find that coastal cities and people in inland states have larger travel radii.

## 2 Data Preparation

### 2.1 Study area and data retrieval

Georeferenced tweets were downloaded between December 6, 2013, and January 6, 2014, using an open source library called 140dev phirehose, developed by Adam Green. The client was written in PHP and the data downloaded into a MySQL database using a set of geographic bounding boxes. For this study, tweets were not filtered by device (e.g. desktop computer or mobile device), and tweets from all applications, e.g. Twitter for Android or TweetDeck, were used.

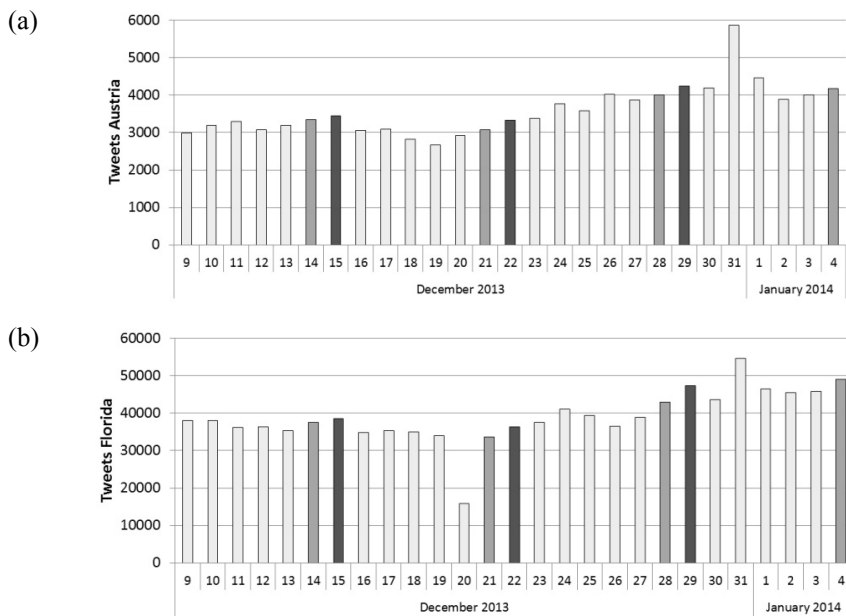
The study uses data from two distinct regions to identify potential differences in user tweeting behavior, and the consequences for travel pattern analysis as a result thereof. The first test region covers Austria and adjacent areas (Fig. 3). It was chosen to explore whether one could observe tweet-based travel patterns associated with winter tourism from the surrounding larger cities, e.g. Munich, Vienna, Graz, or Maribor. The second test region encompasses the Eastern stretch of Florida and southern part of Georgia. Florida was chosen since it is a prominent vacation destination where changes in travel behavior throughout the week or during the winter break could possibly be detected with Twitter data.

### 2.2 Data preparation

The analysis consists of two parts. Firstly, mean travel directions for individual grid zones are computed based on trips originating from these zones. Secondly, space-time clusters of tweets are identified. In terms of data requirement both analyses have in common that only tweets from human users should be used, whereas automated, software generated tweets should be excluded. Each tweet comes with a time stamp provided in UTC, which was in a first step converted to local time for the two regions. Using time stamp and position information, the travel speed between two consecutive tweets of each user was computed, assuming a straight travel trajectory. We removed tweets from users where the travel speed between any two tweets exceeded 150 km/h. Visual inspection of point patterns from

excluded users revealed a random-like sequence of tweet locations spatially far apart, but separated only by seconds. These users, based on the information provided in their user profiles, were, for example, radio stations or online newspapers. Further we checked the remaining users for an excessive frequency of tweets, which would also be indicative of an automated tweet service. However, no such users were observed. This filtered dataset was used as is for the space-time permutation model described in section 4.

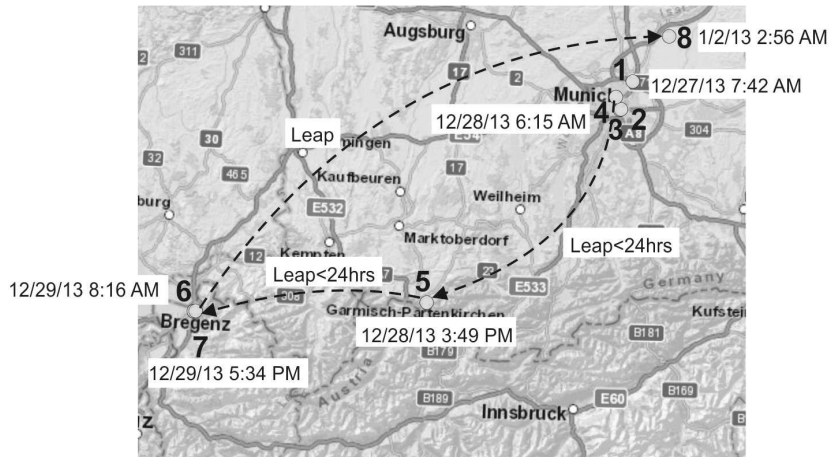
Fig. 1 shows the daily numbers of tweets remaining after this first filtering step for the Austria (a) and Florida (b) test regions. Weekdays are shown in yellow, Saturdays in orange, and Sundays in brown. On December 20, 2013, some unexpected difficulties in connecting to the Twitter API for the Florida region led to a drop in collected data, as can be seen in Fig. 1b. In general the number of tweets is relatively consistent throughout the week, with slightly higher activities on Sundays. Both regions show increased Twitter activity from the end of December onwards, probably due to more frequent vacation trips during the winter break. The peak activity is found on December 31 for both tested regions.



**Fig. 1:** Daily tweets for Austria (a) and Florida (b) after the removal of automated tweets

The analysis of directions for long distance travel behavior requires the identification of users whose sequence of tweets contains at least one long-distance leg (leap). For this second filter step we used a minimum direct distance threshold of 70 km between two subsequent tweets of a user, which corresponds to a network distance of approximately 100 km when considering the circuitry of road networks. Furthermore, to be able to identify the approximate day of a leap, we kept only leaps for which the time stamps of the first and second tweet were less than 24 hours apart. Fig. 2 clarifies these principles by visualizing a user's sequence of tweet locations in the Austria test region. The trip sequence starts with a tweet in Munich on 12/27/13. Three more local tweets follow, with the last one being sent

on 12/28 in the morning. The fifth tweet follows about eight hours later in Füssen. The step between tweets #4 and #5 denotes a leap spanning less than 24 hours. This is followed by another leap completed within 24 hours to Bregenz (tweet #6). The last leap from tweet #7 to #8 ends at the Munich airport, but spans more than 24 hours. Thus the history of this user would contribute two trips for the grid based direction analysis (#4->#5, #5->#6). This filtering method assumes that users do not tweet frequently during their long-distance travel, at least not more frequently than every 100 km, since otherwise the leap would not be detected.



**Fig. 2:** Leaps in a user's travel history in the Austria analysis region

Tab. 1 provides statistics about the collected and processed tweets from both test regions. The middle section of the table illustrates the effects of the three aforementioned filter steps on the number of retained tweets. Removal of users with automated tweets reduces the number of tweets to about one half for Austria and one fifth for Florida. The removal of tweets from users that do not have a leap in their tweet sequence reduces tweet numbers to 5.3% and 3.2%, respectively. The third filter reduces the number of remaining tweets to about 0.5% and 1.3%, respectively, of the original georeferenced tweets for Austria and Florida.

**Tab. 1:** Tweet statistics between 12/9/13 and 1/4/14

		Austria	%	Florida	%
<i>Region</i>	Reference area [km <sup>2</sup> ]	141841	–	147975	–
	Georeferenced tweets	191894	100	5085750	100
<i>Data filtering</i>	Step 1: Manual tweets	96973	50.5	1072339	21.1
	Step 2: Tweets with leaps	10102	5.3	160395	3.2
	Step 3: Leaps in < 24 hrs	907	0.5	67496	1.3
	Manual tweets/km <sup>2</sup> /day	0.025	–	0.27	–
<i>User statistics</i>	Real users tweeting >1 time	11629	100	97608	100
	Users with leaps	1236	10.6	13555	13.9
	Users with leaps in < 24hrs	144	1.2	3908	4.0

While the test regions are of approximately equal size, Florida has about 11 times as many manual tweets as Austria. This relative data abundance makes Twitter data more suitable for travel pattern analysis in Florida than in Austria. Florida has 3908 users with leaps spanning less than 24 hours, as opposed to Austria with only 144 (Tab. 1 bottom).

Next, both test regions were subdivided into squared cells approximately 90 km wide. While it is possible to build an origin-destination matrix that counts the number of trips between zone pairs, leaps spanning less than 24 hours were scarce. Thus, instead we tried to capture the mean direction of travel originating from each cell for each day. All leaps spanning less than 24 hours (last row in Tab. 1) were extracted from the database, and the direction of travel between leap origin and destination was computed as an azimuth, i.e. an angle measured in clock-wise direction from north. Next, the mean azimuth was computed over all leaps originating from the cell under consideration, separated by date as found in the time stamp of the destination tweet. The computation of the mean azimuth follows the method in (ROGERSON & YAMADA 2009):

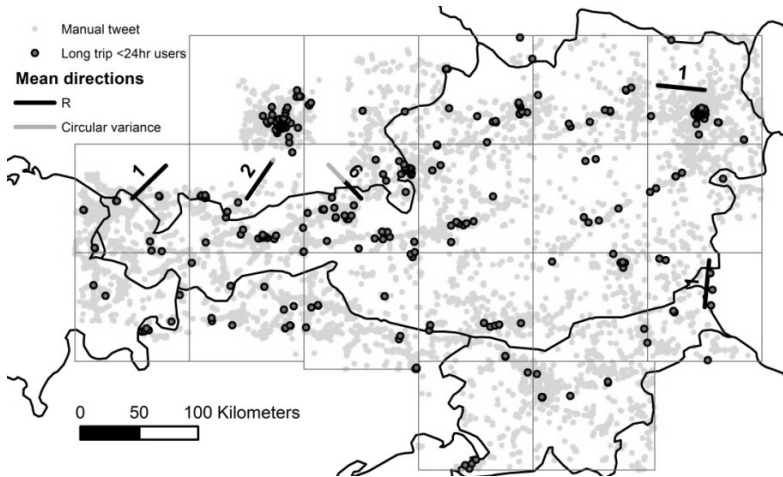
1. Compute the mean of cosines (C) and sines (S) of the azimuths;
2. Find  $R = \sqrt{C^2 + S^2}$ ;
3. The mean azimuth A is the angle that satisfies  $A = \sin^{-1}(S/R)$  and  $A = \cos^{-1}(C/R)$ .

Further, the variance of the azimuth, termed the *circular variance*, can be computed as  $1 - R$ . A high R means that the circular variance is low and that, in this context, travel directions do not show much variability. Thus the circular variance is a measure of dispersion of azimuths.

### 3 Travel Analysis: Directions and Frequencies

This first part of the analysis explores potential changes in the mean azimuth of a zone between different days. Fig. 3 shows the Austrian test region with its grid cells. Orange dots indicate the location of manual tweets, and green dots the location of tweets from users that have at least one leap spanning less than 24 hours in their history. It shows further mean azimuths for December 15, 2013, expressed as bars anchored at the respective center of grid cells. Each bar has the same length since only travel direction is considered for this analysis. The proportion of black coloring on each bar shows the R value, ranging between 0 and 1, as described above. A high R, and thus a large black proportion of a bar, indicates a high consistency of travel azimuths and leaps between different users originating from that zone on a given day. As opposed to this, gray expresses circular variance. Thus a bar with a large gray proportion indicates high dispersion of azimuths between users on that day for that cell. The number next to each bar shows how many leaps originated from a given zone on a given day. Cells without a bar indicate that there were no leaps originating from that cell on a given day. The pattern of bars in Fig. 3 shows that some trips originating from western Austria were heading towards Munich. In general it can be expected that the directions of the bars tend to be oriented towards the center of the test region, where the mean azimuth would point if leap destinations were randomly distributed throughout the test region. This is because cells on the edges do not have neighboring cells on each side, limiting the range of possible azimuths of trips originating from these cells (edge effect). As opposed to this, for zones surrounded by all nine cells (which is the case for three cells in Fig. 3), trips are possible in any direction. As a consequence it can be expected that mean azimuths will be more dispersed around the region center.

Despite the edge effect we were hoping that changes in mean azimuths between different days for a given cell would provide insight into trip flow directions between days. However, the small sample size of leaps in each zone for the Austria test region makes the mean azimuths unreliable, leaving many cells even without a mean direction. Thus it can be concluded that due to data scarcity tweets cannot be successfully applied for this kind of aggregated travel analysis for the Austrian network. To gain higher data reliability, one could, for example, consider leaps for the same day of the week (e.g. Wednesdays) aggregated over several months, as long as travel patterns can be expected to be stationary within the assessed time period. While this is an option, this approach was not tested in this study since the data collection time was too short.

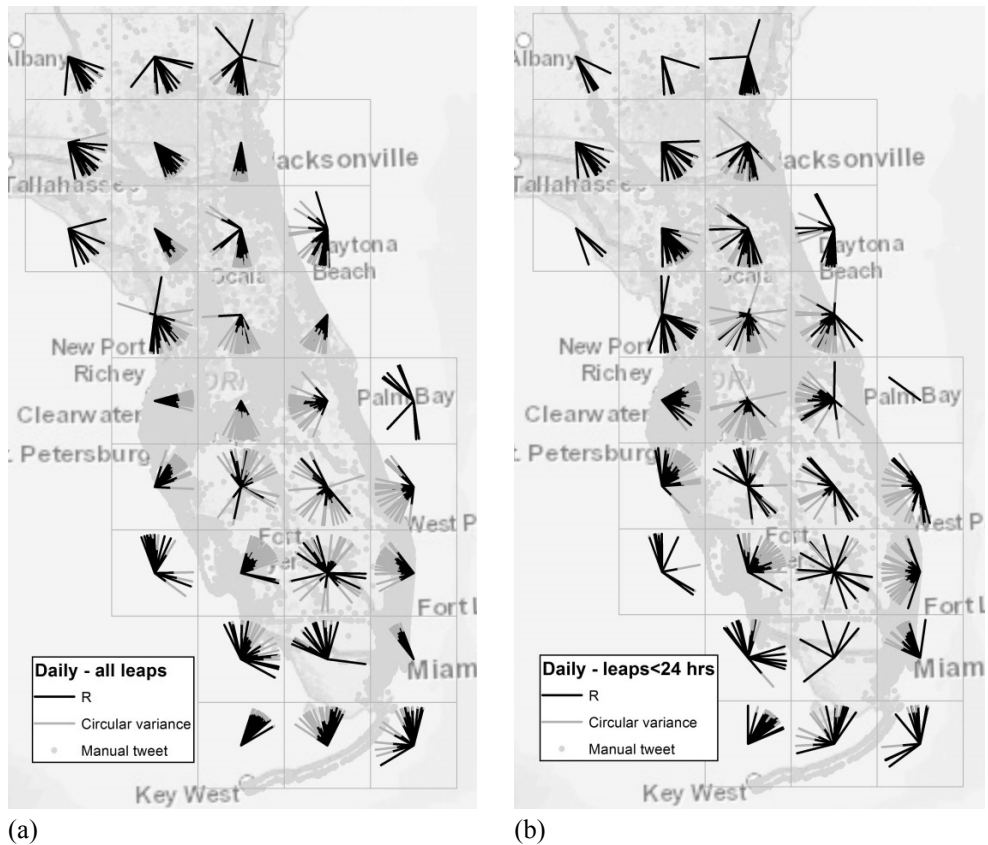


**Fig. 3:** R and circular variance for the mean in directional observations for 12/15/2013

Fig. 4 visualizes mean azimuth bars for zones in the Florida test region for all days of the four week test period. Fig. 4a considers all leaps, whereas Fig. 4b shows mean azimuths only for leaps spanning less than 24 hours, which leads to a more scarce directional pattern. The figure also shows that azimuth directions have a larger variability in the region center both on individual days (indicated through the longer gray portions on the bars), and between days (shown through a wider range of azimuths of bars between different days), both resulting from the edge effect.

A closer look at selected days suggests some travel pattern change between days. Fig. 5a shows the mean azimuth bars for two days before the winter break. Bars in black/gray show azimuths for Wednesday, 12/10, and those in blue for Saturday, 12/14. Leap numbers next to bars are generally higher for Saturday than for Wednesday. This difference in leap numbers, combined with the fact that the total number of tweets is similar on these two days (Fig. 1b) shows that a higher percentage of Twitter users is traveling on the weekend. Besides trip numbers, some indication about a change in average travel direction may be obtained from this analysis as well. The two circled zones show examples where the average azimuth on Saturday tends to point more towards larger cities, i.e. Tampa and Miami, compared to Wednesday, suggesting preferred destinations for weekend trips in Metropolitan areas.

Fig. 5b juxtaposes the mean azimuth bars for Saturday, 12/14 (blue), and Tuesday, 12/31 (black/gray). While the total number of tweets in the Florida region is about 45% higher for 12/31 than for 12/14, some zones show more than twice the number of leaps on New Year's Eve than on 12/14, indicating an increased number of trips on 12/31. An example is the zone around Miami (circled), where the number of leaps increases from 14 to 39.

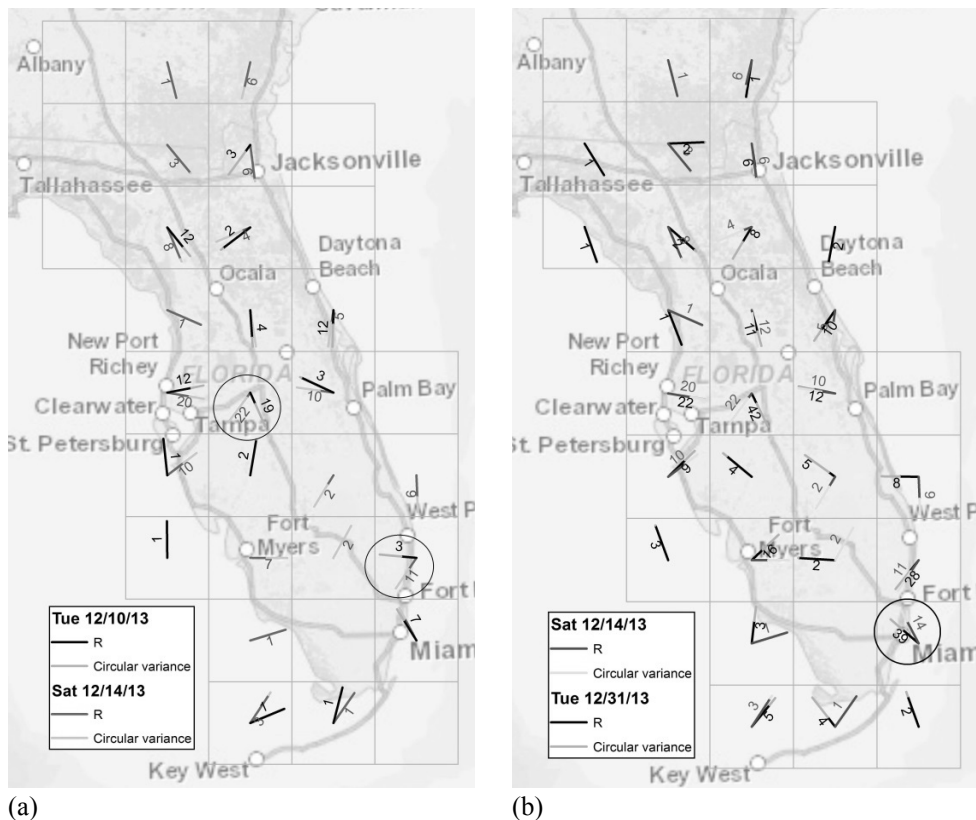


**Fig. 4:** Daily mean azimuths using all leaps (a) and leaps spanning less than 24 hours (b)

## 4 Cluster Analysis

In the second part of the analysis a space-time permutation model (KULLDORFF et al. 2005) was run in the SaTScan 9.2 software package. This model applies a spatial scan statistic to detect significant spatio-temporal clusters in a set of discrete case events, i.e., manual tweets (compare Tab. 1). The goal was to identify a potential change in hot spot locations over time, and thus potential travel destinations for Twitter users. For Austria the complete set of approximately 97000 tweets was used, whereas for Florida a random sample of 50000 out of the  $\sim 1$  mio. tweets was used instead. This analysis resulted in numerous significant ( $p < .05$ ) clusters with a radius between a few meters and a few kilometers for

both test regions, but no larger clusters. Therefore it was explored whether these small clusters group within certain regions on a given day. For Austria, the hypothesis was that during winter break and weekends some Twitter users travel into the mountains, more specifically to ski resorts. Thus, in the first step, all ski resorts with at least 5 ski lifts in the Austria test region were digitized from <http://www.bergfex.com>. Fig. 6 shows these ski resorts as well as clusters identified for Saturday, 12/21/13, as an example.



**Fig. 5:** Mean in directional observations for Dec 11 and 14 (a), and Dec 14 and 31 (b)

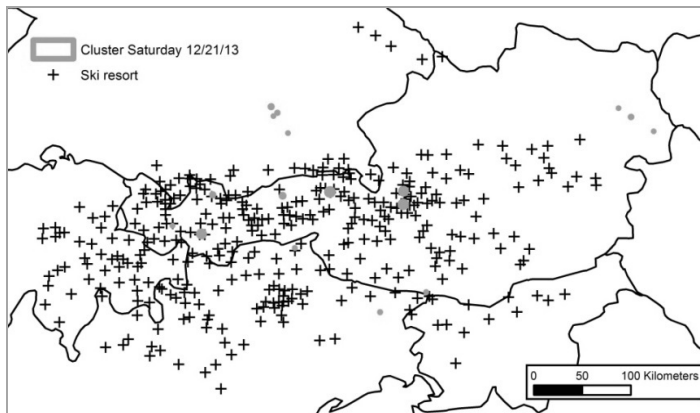
Next, the percentage of clusters within a 5 km buffer around ski resorts was determined for different days. Tab. 2 lists this percentage, together with the total number of clusters that were found for a given date, for Wednesday through Saturday for the four weeks of the data collection period. Week 2 and 3 show the highest average percentage of clusters near ski resorts (26% and 24%, respectively), whereas the lowest average percentage is in Week 4. Thus, during the peak skiing season (end of December to early January), there is no evidence for a higher proportion of tweets in ski resorts. The percentage for the first week peaks on Saturday. Since 12/14 is not yet holiday season, this peak may be potentially attributed to workers or students returning to home towns, some of them ski resorts, from larger cities, such as Munich. Week 2 also peaks on the weekend, i.e. Friday and Saturday, even more so than the first week. This could also indicate the returning of residents to their home town at the beginning of the holidays. The third week, which falls completely within



the winter break, shows no specific peak day for percentage, indicating no major shift from or to ski resorts. Week 4 has the lowest percentage of tweet clusters around ski resorts among all four analyzed Saturdays. This may indicate tourists' arrivals in their home towns at the end of holidays, away from the ski resort. It must be noticed that the clustering is not necessarily caused by the travel of people between cells in the analyzed area, but potentially also by visitors from outside areas, e.g. other countries.

**Tab. 2:** Clusters around ski resorts for the Austria test region

	Start date	Day of week	Clusters total	% $\leq 5\text{km}$ from ski
Week 1	12/11/2013	Wednesday	16	19
	12/12/2013	Thursday	17	24
	12/13/2013	Friday	15	13
	12/14/2013	Saturday	16	31
Week 2	12/18/2013	Wednesday	14	14
	12/19/2013	Thursday	4	25
	12/20/2013	Friday	26	35
	12/21/2013	Saturday	24	29
Week 3	12/25/2013	Wednesday	25	28
	12/26/2013	Thursday	33	15
	12/27/2013	Friday	23	26
	12/28/2013	Saturday	33	27
Week 4	1/1/2014	Wednesday	17	12
	1/2/2014	Thursday	31	29
	1/3/2014	Friday	20	25
	1/4/2014	Saturday	15	13



**Fig. 6:** Spatio-temporal tweet clusters on Saturday before winter break

For Florida, this method revealed only 65 significant clusters compared to 500 for Austria. Visually, no obvious pattern in the shift of clusters between days could be observed. A possible explanation is that Florida offers numerous vacation sites, metropolitan areas, and universities throughout the state, balancing the trip flows between these locations.

## 5 Conclusions and Future Work

This study assessed the suitability of Twitter data for travel pattern analysis at the regional level. Only up to 5% of georeferenced tweets reveals information on long distance travel. Data scarcity was a severe problem for the direction based analysis method, especially for Austria. The approach provided some insight for the Florida test region, where the manual tweet density is ten times higher than for Austria. The analysis revealed changes in the number of trips and flow direction on different days. For future work we will consider aggregating several days to increase the sample size of leaps, as well as expanding the study area to mitigate the edge effect. It is also necessary to identify the inherent biases of Twitter data, e.g. data scarcity or limitation to certain demographic groups, for this kind of analysis.

## References

- ANDRIENKO, G., ANDRIENKO, N., BOSCH, H., ERTL, T., FUCHS, G., JANKOWSKI, P. & THOM, D. (2013), Thematics Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science & Engineering*, May/June, 72-82.
- CHENG, Z. & LEE, J. C. (2011), Exploring Millions of Footprints in Location Sharing Services. *Proceedings of Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.
- GIRARDIN, F., BLAT, J., CALABRESE, F., FIORE, F. D. & RATTI, C. (2008), Digital Footprinting: Uncovering Tourists with User-Generated Content. *Pervasive Computing*, 7 (4), 36-43.
- GOODCHILD, M. F. (2007), Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0 (Editorial). *International Journal of Spatial Data Infrastructures Research (IJSDIR)*, 2, 24-32.
- KRUMM, J., CARUANA, R. & COUNTS, S. (2013), Learning Likely Locations. In: CARBERRY, S., WEIBELZAHN, S., MICARELLI, A. & SEMERARO, G. (Eds.), *User Modeling, Adaption, and Personalization – Proceedings of UMAP 2013 (LNCS 7899)*. Berlin, Springer, 64-76.
- KULLDORFF, M., HEFFERNAN, R., HARTMAN, J., ASSUNÇÃO, R. & MOSTASHARI, F. (2005), A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. *PLOS Medicine*, 2, 216-224.
- NOULAS, A., SCELLATO, S., LAMBIOTTE, R., PONTIL, M. & MASCOLO, C. (2012), A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE*, 7 (5), e37027.
- PARTHASARATHI, P., HOCHMAIR, H. H. & LEVINSON, D. M. (2012), Network structure and spatial separation. *Environment and Planning B, Planning and Design*, 39 (1), 137-154.
- ROGERSON, P. & YAMADA, I. (2009), *Statistical Detection and Surveillance of Geographic Clusters*. Boca Raton, FL, CRC Press.
- RÖSLER, R. & LIEBIG, T. (2013), Using Data from Location Based Social Networks for Urban Activity Clustering. In: VANDENBROUCKE, D., BUCHER, B. & CROMPVOETS, J. (Eds.), *Geographic Information Science at the Heart of Europe (Lecture Notes in Geoinformation and Cartography)*. Berlin, Springer.