Use of Geolocated Twitter Data in Mobility Pattern Analysis of Foreign Tourists

Transport Engineering Laboratory
Isaac Alfonso Garcia Garcia

1. INTRODUCTION

Explaining leisure travel is challenging the field of transport planning in both quantitative and qualitative senses [1]. This is mainly because it is driven by different, less constant factors.

Tourism related studies have relied either on aggregated and temporally-sparse official statistics or on selective small-scale observations and surveys [2]. However, with the surge of smartphones and social networks like Twitter, the use of the digital trace that people leave behind while travelling has become possible. This type of big data is called "tourism big data" and has the potential to provide higher detail into patterns of tourists at a lower cost than traditional surveys.

Previous research has found that mobility measures obtained from geolocated Twitter data have the same behavior as other human mobility related datasets. Moreover, trips extracted from this type of data were found to be closely related to leisure activities, validating the use of Twitter as a data source for tourism related statistics.

Additionally, previous studies on filter methods for identifying foreign tourists in Twitter data have concentrated in calculating external parameters as a form to detect foreign tourists. However, the current research explores the use of existing parameters within the data in order to achieve a higher accuracy.

2. METHODOLOGY

The study proposed a new data filter method to identify foreign tourists by the utilization of internal parameters from geolocated Twitter data. The method was tested with data from Japan, and validated with official tourism statistics. Several spatial-temporal analyses were conducted to investigate: travel patterns, trip purpose, and tourist group behavior. Finally, a case study was performed by using data from Mexico.

3. RESULTS AND DISCUSSION

3.1 Proposed Method for Data Filter

The data filter method was tested with summer season data from Japan. From the results of studying the composition of each subgroup, shown in Figure 1, user language parameter was found to be the best indicator of whether a user is foreigner. The same trend was found when analyzing the total geolocated tweets by each subgroup.

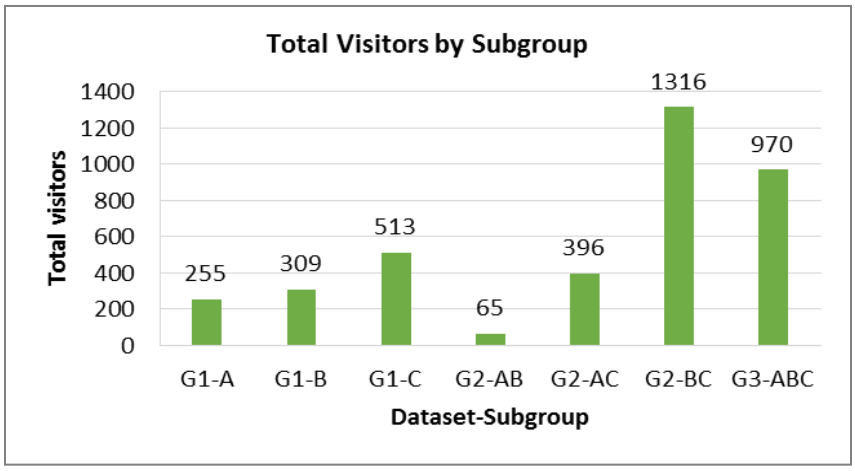

Figure 1. Total Identified Visitors by Subgroup

3.2 Validation Results

The accuracy level of the filter method was calculated by evaluating a random sample of each dataset. Accuracy results are shown in Table 1. Additional validation was performed by correlating the most accurate datasets (G2 and G3) to official statistics, presented in Table 2.

Table 1. Accuracy of Filter Method

Accuracy rate per data group			
Dataset	**Subgroups**	**Parameters fulfilled**	**Accuracy**
G1	A	Tweet lang.	54%
	B	Time zone	
	C	User lang.	
G2	A	Tweet lang. & Time zone	**92%**
	B	Tweet lang. & User lang.	
	C	Time zone & User lang.	
G3	A	All parameters	**94%**

Table 2. G3 and G2 Correlation to Official Statistics

Comparison to G3 Results (r^2 values)		
Group	G3	G2
Foreigner hotel guests	0.862	0.8803
Foreigner hotel guests (Excluding top 3 prefectures)	0.7952	**0.6868**

Comparison to G3 Results (r^2 values)		
Group	G3	G2
Foreigner tourist visits	0.7156	0.7925
Foreigner tourist visits (Excluding top 3 prefectures)	0.8548	**0.7583**

3.3 Spatial-Temporal Analysis

Travel Patterns

From the visitor Trajectory Analysis, flows of tourists to and from Narita International Airport were found to follow the trends in departure and arrival international flight times. An additional study in Nara prefecture investigated a trend for tourists to visit during the day and avoid staying the night. This had already been identified in Japanese tourists by local authorities. However, from the results the phenomenon was confirmed to also be present with foreigner tourists. Therefore, the movements of visitors were confirmed to be captured by the filtered data and were validated.

Trip Purpose

Land use data from Tokyo was use to infer the trip purpose of visitors. From the results, on average 77% of the activity was identified to happen in commercial districts. Nevertheless, insight on trip purpose was found to be limited from this data. From further studying the spatial distribution of the identified visitors, a strong correlation was found to tourist attractions. Moreover, popularity was identified to be a factor. Finally, the preferences to certain types of tourist attractions could also be observed in the data.

Tourist Group Behavior

In the language distribution analysis, the correlation between the filtered data and different types of tourist attractions was studied. In the case of the English group no particular trends were observed. On the contrary, in the Tagalog group specific trends were found. This is considered to be due to English being widely used by people of different nationalities, whereas Tagalog is a more regional language.

3.4 Case Study in Mexico

The adaptability of the proposed method was studied by applying it to one week of data from Mexico. From the results, the data filter can be applied. However, improvements to filter out bot accounts are necessary to increase the obtained accuracy (71%).

4. CONCLUSIONS

The findings of the present study are summarized as follows:

- User language was found to be the best indicator of whether a user is foreigner.
- Accuracy was found to increase with the increase in the parameters fulfilled by the data. Additionally, filtered data was highly correlated to official tourism statistics.
- Movements and trends in mobility were found to be captured by the filtered data and can be validated.
- Land use data provided limited insight on the trip purpose of visitors. However, trends were identified when correlating to tourist attractions.
- Region-specific languages were found to better provided insight on the trip purpose of visitors.
- Finally, the data filter method can be applied to other cases. Nonetheless, improvements in the bot account filter are suggested.

REFERENCES

1. UrryJohn. (2011). Mobilities: New Perspectives on Transport and Society.
2. BartoszHawelka, Izabela Sitko, EuroBeinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, CarloRatti. (2012). Geo-located Twitter as the proxy for global mobility patterns.