INVESTIGATING THE POTENTIAL OF SOCIAL NETWORK DATA FOR TRANSPORT DEMAND MODELS

Michael A.B. van Eggermond[1,*], Haohui Chen[2], Alexander Erath[1], Manuel Cebrian[2]

[1] Future Cities Laboratory, Singapore ETH Centre, [2] National Information and Communications Technology Australia (NICTA)

[*] Corresponding author: eggermond@ethz.ch

SUMMARY

Shortcomings of travel diaries include the common underreporting of short trips and, more importantly, that it is not feasible to sample from all potential user groups and over a longer time span in the study region due to time and budget limitations.

Social network data offers the possibility to observe users over a larger time span for almost negligible costs. Studies have shown the possibilities of using social network data; however, a **comparison** with travel diaries or other transport related data sources is **lacking**.

In this paper we analyze **geo-referenced Twitter** activities from a large number of users in Singapore and its neighbouring countries. By combining this data, population statistics and travel diaries, and applying clustering techniques, we address questions regarding the **detection of activity locations**, the **spatial separation** between these locations and the **transitions** between these locations.

Descriptive analysis shows that **determining home locations is more difficult** than detecting work locations for most planning zones. The **spatial separation** between detected activity locations from Twitter data and as reported in a travel survey and captured by public transport smart card data are at large **similarly distributed**. This equally holds for the transitions between zones.

Whether the differences between Twitter data and other data sources stem from differences in the population sub-sample, the clustering methodology or whether social networks are being used significantly more at certain locations is to be determined by further research.

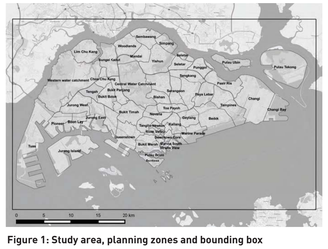


Figure 1: Study area, planning zones and bounding box

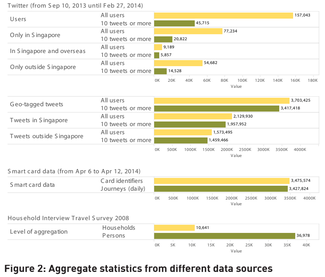


Figure 2: Aggregate statistics from different data sources

STUDY AREA & DATA

As a case-study for this study we consider the city state of **Singapore**. Singapore has a land area of 712 km² and has a population of 5.08 million (2010). GDP per capita amounts to S$ 59,813 (US$ 45,200, 2010), which makes it one of the wealthiest countries in (Southeast) Asia.

As opposed to many other social networking sites, **Twitter** offers the opportunity to download the profile of the users and Twitter messages, or tweets, including the geo-location of the tweet. Data has been collected for a bounding box surrounding Singapore from September 10, 2013 until February 27, 2014.

Trip information is given by the Household Interview Travel Survey 2008. The survey is conducted once every four years and is commissioned by the Singaporean Land Transport Authority.

We use 7 days of **smart card data** from trips made between April 6, 2013 to April 12, 2014. Smart card data records per smart card the boarding station, the boarding time, the alighting station and the alighting time.

Singapore's **populations statistics** have been included as well as estimated work locations in Singapore by planning zone.

Despite a large number of Twitter users present in the data set which we collected over a period of 8 months, only an amount comparable to a travel survey turned out to be useful for further analysis. This is mainly due to the limited number of user tweeting frequently.

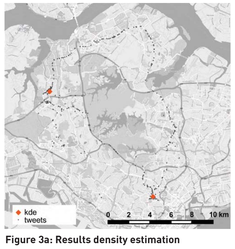


Figure 3a: Results density estimation

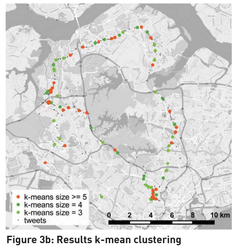


Figure 3b: Results k-mean clustering

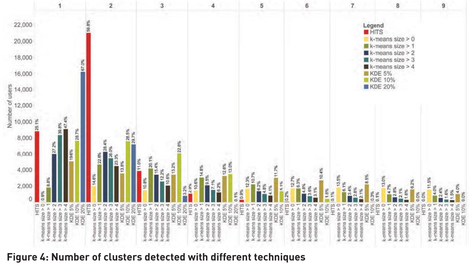


Figure 4: Number of clusters detected with different techniques

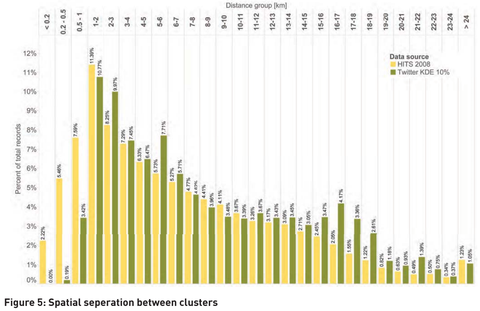


Figure 5: Spatial seperation between clusters

METHODOLOGY

The challenge is to **recognize locations** visited by an individual. With locations, activity locations in a traditional sense are meant: an individual's home location, work location, education locations and locations where discretionary activities are performed. As such, we do not touch upon the fact that activities are also performed en-route.

Each tweet consists of a timestamp, a user identifier, a location and a message. It is assumed that events occurring at activity locations tend to be less geographically dispersed. Partitioning geographically close **events into clusters** should help identify those en-route activities, as their clusters should contain fewer events.

Two techniques are applied: (1) **recursive k-means clustering** and (2) **kernel density estimation**. The results for one single user can be seen in Figure 3a and Figure 3b. The selected user has tweeted 1,405 times. While the data might look similar to GPS data in terms of detected trajectories, these tweets are not necessarily ordered by time. By means of KDE two clusters are recognized; by means of k-means clustering more than 100 clusters are recognized containing two tweets or more.

NUMBER OF ACTIVITY LOCATIONS

To determine the merits of both the k-means clustering and KDE both methods are **evaluated by the number of clusters recognized** per user and the strength of each cluster. The results of this comparison are shown in Figure 4.

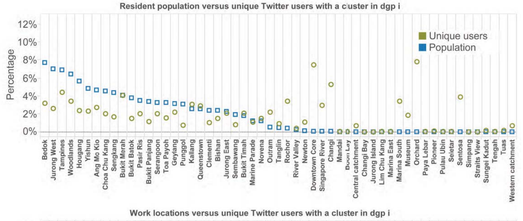
For clusters recognized by k-means clustering the strength is calculated as the number of tweets belonging to each cluster; the size of the cluster. For clusters recognized by kernel density estimation the strength is calculated as the contribution of a single cluster to the sum of the levels of each cluster.

If the goal is to determine the number of **frequently visited locations a threshold** will need to be set. If the goal is determine a users activity space it is possible not to set thresholds and by doing so, not deleting user information.

SPATIAL SEPERATION

To assess whether the distances between different data sources correspond for both data sources, Euclidean distances between all **unique reciprocal locations** per user are calculated. In Figure 5 the results of the **distance comparison** are presented.

Overall, the **distance distribution is similar.** In the household interview travel survey a higher number of cluster-pairs is reported being separated less than 1 kilometer. A closer analysis of HITS reveals that clusters being separated less than 1 kilometer concern the activity pairs 'home-education' (44%), 'home-pick up drop' (11%) and 'home-work' (10%).



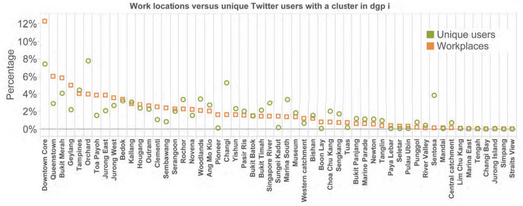


Figure 6: Comparison with aggregate statistics

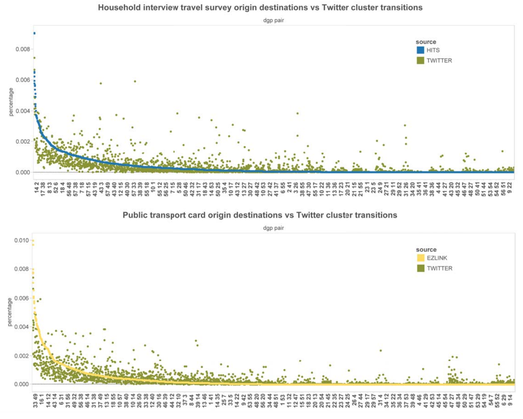


Figure 7: Comparison of origin-destination relationships. Travel survey (top) and smart card data (bottom).

GENERAL STATISTICS

In Figure 6 the number of users with one or more clusters in each respective planning zone against the population (top) and the number of work locations (bottom) are compared. As the **order of magnitude** differs from the number of users found in Twitter, the percentage of the population residing and working in each zone is shown and compared against the number of unique Twitter users with a cluster in this zone.

This comparison highlights that **clusters** where users tweet **are not limited to home or work activity locations**. Zones with outliers include the popular shopping district Orchard, the nightlife district Singapore River and the airport Changi.

DETECTING TRANSITIONS

In Figure 7 the transitions between zones are shown from survey data, smart card data and detected activity locations. Intra-zonal and weekend trips have been excluded.

The relative flow per origin-destination pair is shown. Records are sorted by the percentage per od-pair from smart card data and HITS data respectively. This approach makes it possible to compare the trends between both data sources and detect differences between both data sources.

It can be observed that in both cases **transitions derived from Twitter follow a similar trend** to both smart card data and HITS. The correlation coefficient between HITS and smart card data is 0.88, the correlation coefficient between HITS and Twitter is 0.71 and the correlation coefficient between smart card data and Twitter is 0.76.

KEY FINDINGS

The application of **kernel density estimation** for the detection of clusters yields **more promising** results than k-means clustering.

An important input for transport demand models is **trip distance**. The spatial separation between detected locations and reported activity location **corresponds well**. Short trips under 1 kilometer, 44% of which are home-school trips, are under-estimated.

Location-based **social network data** provides a **promising data source** for the detection of activity locations and the analysis of mobility patterns, especially considering the potential to track users over a longer span of time against negligible costs.

ACKNOWLEDGEMENTS

The research conducted at the Future Cities Laboratory is co-funded by the Singaporean National Research Fund and the ETH Zurich, and located at the Campus for Research Excellence And Technological Enterprise (CREATE). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

ETH zürich CREATE Campus for Research Excellence And Technological Enterprise NICTA

(FCL) FUTURE CITIES LABORATORY 未来城市实验室 (SEC) SINGAPORE-ETH CENTRE 新加坡－ETH研究中心