Investigating the potential of social network data for transport demand models

Investigating the potential of social network data for transport demand models

Date of submission: 2014-08-01

Michael A.B. van Eggermond
Future Cities Laboratory, Singapore ETH Centre,
1 CREATE Way, Singapore 138602
phone: +65 6601 4053
fax:
eggermond@ivt.baug.ethz.ch

Haohui Chen
National Information and Communications Technology Australia (NICTA), ,
115 Batman Street, West Melbourne VIC 3003
phone: +61 3 9903 4371
fax:
CaronHaohui.Chen@nicta.com.au

Alexander Erath
Future Cities Laboratory, Singapore ETH Centre,
1 CREATE Way, Singapore 138602
phone: +65 6601 4053
fax:
erath@ivt.baug.ethz.ch

Manuel Cebrian
National Information and Communications Technology Australia (NICTA), ,
115 Batman Street, West Melbourne VIC 3003
phone: +61 3 9903 4371
fax:
manuel.cebrian@nicta.com.au

Words: 5199 words + 7 figures + 2 tables = 7449 word equivalents

ABSTRACT

Location-based social network data offers the promise of collecting the data from a large base of users over a longer span of time at negligible costs. While several studies have applied social network data to activity and mobility analysis a comparison with travel diaries and general statistics is lacking.

In this paper we analyze geo-referenced Twitter activities from a large number of users in Singapore and it's 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.

Despite a large number of Twitter users present in the dataset which we collected over a period of 8 months, only an amount comparable to a travel survey turned out to be useful further analysis due to the scattered nature of the data. Kernel density estimation performs best to detect activity locations; more activity locations are detected per user than reported in the travel survey. 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 as identified using Twitter data and as reported in a travel survey and captured by public transport smart card data are at large similarly distributed, but also show relevant differences among certain distance bands. This equally holds for the transitions between zones. Whether the differences between Twitter data and other data sources stem from differences in the population subsample, the clustering methodology or whether social networks are being used significantly more at certain locations is to be determined by further research. Despite these shortcomings, location-based social network data offers a promising data source for insights in activity locations and mobility patterns.

INTRODUCTION

The well-established four step transport model as well as state-of-the art agent-based models largely have largely relied on the same data sources over the last decades. Traditional data sources include, but are not limited to, household travel surveys, population censuses, business censuses, road networks and transit schedules.

Household travel diaries aim to give insight in questions surrounding in variables relevant as input for transport demand. These variables include, but are not limited to, mode choice, departure time choice, trip frequency choice and distances between different activity locations. Population and business censuses give insight in (aggregate) population statistics by home and work location. By combining these data sources it is possible to generate transport demand on different levels of detail.

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. Furthermore, both travel diaries and censuses are only conducted every 5 to 10 years and, dependent on the study area, limited in their availability to researchers and the general public. More recently, network data and public transport schedules have become available to the public in formats such as OpenStreetMap *(1)* and GTFS (General Transit Feed Standard *(2)*), are continuously updated or even available real-time.

Social network data offers the possibility to observe users over a larger time span for almost negligible costs. Previous research has for instance considered urban activity and mobility patterns *(3)* as well the recognition of mobility patterns in a range of cities *(4)*. These 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 investigate the possibilities of the usage of data obtained from the social networks for transport planning purposes. More specifically, we investigate the possibility to use Twitter data to complement or replace travel diaries with spatial and temporal information of locations of tweets. To these means, we collected data from the social networking and micro-blogging service Twitter for 8 months for Singapore and it's neighbouring countries Malaysia, Indonesia and Thailand. This data is complemented amongst others with Singapore's household interview travel survey and one week's of public transport smart card data. By merging these data sources, and applying several clustering methods, we address the following questions:

1. Is it possible to recognize activity location from social network data and, if so, do the detected activity locations correspond to activity locations reported in other data sources?
2. Is the spatial separation between detected activity locations comparable to distances between reported activity locations from travel diaries?
3. Is possible to derive origin destination matrices from social network data and how do these matrices correspond to observed trips?

LITERATURE REVIEW

Traditional data sources in transport planning and modeling

Many definitions and models in transport planning follow the inputs required for the classic four-step transport model *(e.g. 5)* and more recently, activity-based models *(e.g. 6)*. These models include trip generation, trip distribution, modal choice and traffic assignment models. In addition to these models, information is required about zonal productions (e.g. number of household) and zonal attraction (e.g. workplaces, leisure locations).

Household travel surveys follow a set-up allowing to estimate the models required for the four-step model. On one hand, data about households and persons is collected. This data includes household income, residential location, dwelling type, etc. On other hand, data about trips being performed by individual each household members is recorded. This includes a trip start time, trip end time, mode(s) used, number of transfers and trip purpose. Activity duration is derived from the difference between the trip end time and trip start time of subsequent trips. To obtain more detailed user data state-of-the-art travel surveys include or are supplemented by GPS data.

An increasing number of cities, regions and countries adopt public transport smart cards. While the main objective of these systems is to collect revenue; a side-result is a very detailed data of onb-oard transactions that can be used for numerous applications *(7)*. Dependent on the type of implementation of the smart card system, a trip start time and/or end time are available to the transport company. Several disadvantages of the usage of smart card data include the lack of trip purpose and the lack of knowing the exact origin and destination of a public transport user *(8)*. Despite these disadvantages, it still possible to extract trip duration (excluding waiting time) and an individual's approximate time at a location.

Social network data

Social network services build on the real-life social networks of people through online platforms to share ideas, activities and interests; the increasing availability of location-acquisition technology offers the extra possibility for people to add a location dimension to existing social networks in various ways *(9)*. Within the field of transport modeling, location-based social network data has been used, amongst others, to classify user's activity patterns *(10)*, detect traffic anomalies *(11)*, the reconstruction of popular traffic routes *(12)*, the recognition of mobility patterns in a range of cities *(4)* and the modeling of human location *(13)*.

While an increasing number of studies use geo-tagged social network data, less attention is being paid to the representativeness of social network data with regard to the general population *(14)*. One critique phrases it as following *(15)*:

In digiplace the wealthy, powerful, educated and mostly male elite is amplified through multiple digital representations. Moreover, the frequent decision of algorithm designers to highlight and emphasise those who submit more media, and the level of 'digital cacophony' that more active contributors create, means that a very small minority - arguably outliers in every analysis of normal distribution of human activities - are super empowered.

However, location-based social network data comes with a larger sample size for a longer period without any significant costs *(10)*. Several disadvantages limit the use of traditional econometric tools for these data sets *(10)*: (i) they do not possess detailed descriptions of activities, such as the start times and the end times, and activities can be either at static locations or en-route (ii) individuals are recognized by only an identifier without additional information on individual socio-economic characteristics; (iii) the data has missing activities, since only activities are observed that an individual shares in social media. In addition to this latter point, it should also be noted that only users active in social media are included.

DATA COLLECTION & PREPARATION

Social network data

The social networking and microblogging service Twitter was launched in 2006. As from March 31st there were 255 million average monthly active Users (MAUs), of which 198 million mobile MAUs *(16)*. Together, these users send 500 million tweets per day *(17)*.

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 and includes an indicator if it was sent from a mobile device or from a computer in real-time. Data has been collected for Singapore from September 10, 2013 until February 27, 2014.

When downloading data from Twitter, the possibility is offered to specify a geographic area. For this research we have specified the bounding box 'Singapore'. In total 4,121,433 tweets have been collected. While a geographic bounding box has been specified, not all tweets are geo-tagged with a longitude and latitude. Also, not all tweets are located in Singapore. Table 1 lists the number of users, tweets, geo-tagged tweets and tweets in Singapore. Additionally, an indicator has been included whether a user has tweeted 10 times or more within the earlier specified time-span. It can be seen that only 29% of the users Tweets 10 times or more within the collected time span. These users contribute over 90% of the tweets.

TABLE 1 Aggregates from different data sources

Data source / Indicator	All users	10 tweets or more	Percentage
Twitter			
Number of users	157,043	45,715	29.1
Number of tweets	4,121,433	3,800,904	92.2
Number of geo-tagged tweets	3,703,425	3,417,418	92.3
Number of tweets in Singapore	2,129,930	1,957,952	91.9
Number of tweets outside Singapore	1,573,495	1,459,466	92.8
Number of users tweeting only in Singapore	77,234	20,822	27.0
Number of users tweeting only outside Singapore	54,682	14,528	26.6
Number of users tweeting in Singapore and overseas	9,189	5,857	63.7
Household interview travel survey 2008			
Number of households	10,641		
Number of persons in household interview travel survey	36,978		
Smart card data			
Number of card identifiers in smart card data	3,475,574		
Number of journeys over 7 days	23,994,771		
Singapore statistics (2012 except were otherwise stated)			
Total population	5,319,000		
Total resident population	3,825,000		
Singaporeans	3,290,000		
Permanent resident	533,000		
Total non-resident population	1,494,000		
Land-area 2013 [km2]	716.1		
Population density 2013 [persons per km2]	7,540		
Per capita GDP 2013 [US$]	55226		

Smart card data

Singapore's public transport card was introduced in April 2002; smart cards can be used island wide for payment of all modes of public transport, regardless of operator. Though cash payment of single fares at higher rates is still possible, e-payments using smart cards account for 96% of all trips *(18)*. In this paper we use 7 days of smart card from trips made between April 6, 2013 to April 12, 2014.

Household interview travel survey

Trip information is given by the Household Interview Travel Survey (HITS) 2008. For this survey 1% of the population is questioned on their travel behavior on a single workday in person. The survey is conducted once every four years and is commissioned by the Singaporean Land Transport Authority (LTA). HITS contains data on three levels of aggregation. The highest level of aggregation contains household characteristics. Second, person characteristics are available such as age, income, profession and employment type. On the lowest level of aggregation information on trips is available such as mode, purpose, cost and time.

Other data sources

We enriched the aforementioned data sets with attributes from several other data sets. To each geo-tagged tweet, public transport trip and HITS-trip several layers of aggregation have been added. These include the 1,092 transport analysis zones (TAZ's), the 55 land-use planning zones but also land-use types. Also, Singapore's populations statistics *(19)* have been included as well as estimated work locations in Singapore by planning zone *(20, 21)*.

METHODOLOGY

Identification of clusters

To assess the suitability of Twitter data for transport demand analysis, we aim 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. For instance, it is possible to work while commuting or maintain social contacts. By observing an individual over longer span of time it would be possible to capture more activity locations than from a one or two day household travel survey. We assume that events (tweets) occurring at activity locations tend to be less geographically dispersed; en-route events would be more geographically dispersed. Partitioning geographically close activities into clusters should help identify those en-route activities, as their clusters should contain fewer events. In our current approach, we do not use the temporal attribute of a tweet directly in the clustering method. By following this approach we are aware that it is possible that sporadic activities, such visiting a concert or a new restaurant, and are accompanied by an event (tweet) are not detected.

K-means clustering

K-means clustering is one of the most popular clustering methods *(22)*. Since K-means clustering was proposed in 1955, a large number of studies has applied different variations of the method in a wide range of domains. Finding the optimal number of clusters k is a challenging but necessary task. A number of methods of obtaining the optimal k value is discussed in *(22)*. Those methods essentially try different values of k and select the best value based on predefined criteria, such as the minimum message length *(23)*, minimum description length *(24)*, gap statistics *(25)* and Dirichlet process *(26)*. A more general and easy-to-implement method for validating clustering results is the silhouette method *(27)*. The value of a silhouette measures (1) how well an observation is assigned to its cluster and (2) how dissimilar that observation to its neighboring clusters, and thus reflects the performance of the clustering analysis. This paper uses the value of silhouette to validate the clustering results of different values of k and selects the optimal value.

Clusters resulting from k-means clustering can be fairly large if measured by the convex hull of all the events (tweets) included in the cluster. For the goal of this research, we assert that a large cluster cannot necessarily constitute a single activity location. In this regard, we define a maximum threshold for the variance of 200 meter. Clusters which exceed this threshold are recursively broken down into more smaller clusters by recursive k-means clustering *(28)*. This process is also highlighted in Figure 1.

Kernel density estimation and clustering

Kernel density estimation (KDE) provides us with another way of determination of individual's frequent visited locations. It is a non-parametric method for estimating a density function from a random sample of data *(29)*. A user-defined parameter called bandwidth h specifies the standard deviation of the Gaussian distribution function constructed around each data point to smooth the KDE result. A small value for h may under-smooth the KDE result, a large value for h can result in over-smoothing. The selection of the bandwidth is discussed in *(30, 31, 13)*. Basically there are two main methods for the selection of bandwidth: the fixed bandwidth and the adaptive

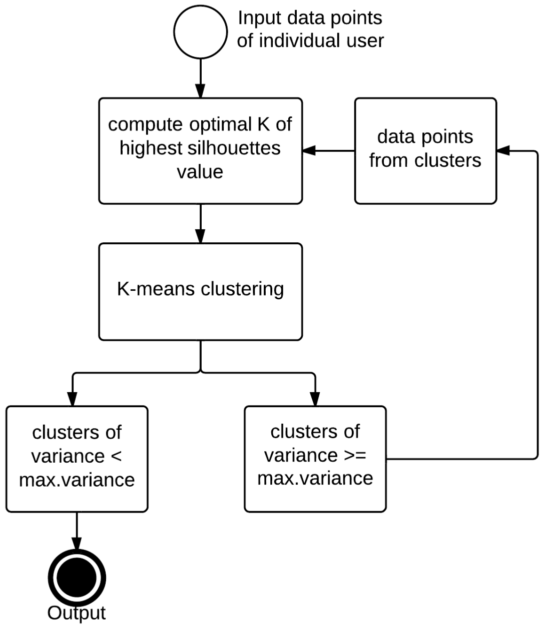

FIGURE 1 K-means clustering procedure

bandwidth method. Given that each location should be limited in area, this paper consider a universal fixed bandwidth for KDE, which is the same as the k-means clustering's maximum variance, namely $h = 200$ meter. To obtain clusters from the estimation procedure, contour lines are constructed based on the results of the KDE. All local peaks of the contour line are regarded as clusters and contour levels are assigned to corresponding kernels. The values of 'level' are relative, and they indicate the frequency of the location is visited. KDE by itself is not a clustering method. However, as the clusters (the peaks) are impacted by neighboring activities, neighboring activities within a certain distance (h) are grouped together. If an activity (tweet) belongs to more than one kernel, it is grouped to the closest one. This situation is very rare in our data set and occurs in less than 0.01% of the cases.

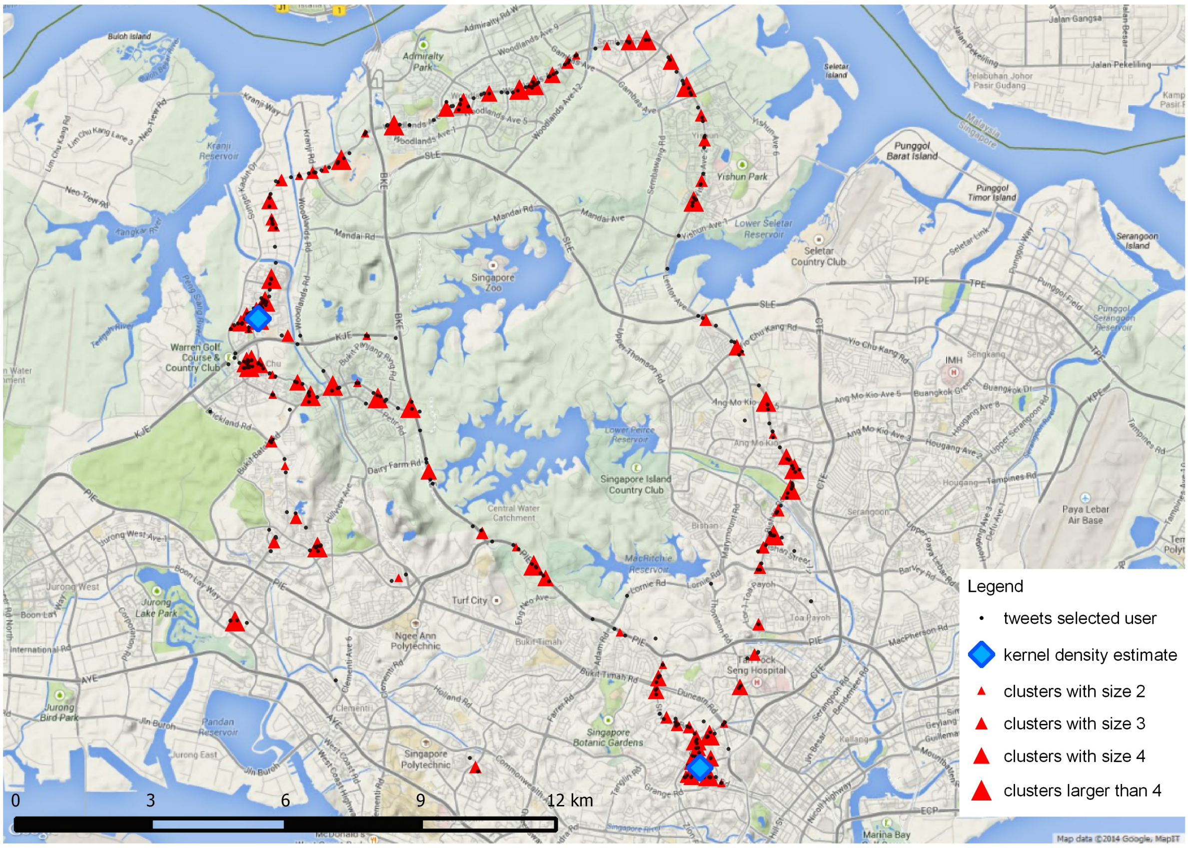

FIGURE 2 Depicted are 1,405 individual tweets of a randomly selected user and the detected clusters by means of k-means clustering and kernel density estimation

RESULTS

Detecting activity locations

Due to the nature of social network data, recognizing user's locations fundamentally differs from another frequently used location based data source used in transport research: GPS data collected through smart phones or dedicated GPS trackers. Whereas the latter data source would result provides location and speed information, making it possible to perform not only mode detection but also detect start and end times of trips activities; social network data only shows when a user is active on the social network and chooses to geo-tag his data. While it might be possible to detect individual locations by means of so-called location check-ins *(4, 10)*, where a user notifies his network that he is at an activity location, such as a shopping mall or restaurant the challenge with a data stream coming from Twitter is to determine whether a user is at an activity location or en-route.

Figure 2 highlights this difference for a randomly selected user. The selected user has tweeted 1,405 times. While the data might look similar in terms of detected trajectories, these tweets are not necessarily ordered by time. The user's main locations have been identified by both k-means clustering and kernel density estimation (KDE). The optimal number of clusters per user has been calculated. Due to the nature of clustering, each data point (tweet) needs to be assigned to a cluster. This is also shown in Figure 2.

To determine the merits of both the k-means clustering and KDE are evaluated by the number of clusters recognized per user and the strength of each cluster. Currently, the strength of each cluster is evaluated as following:

- 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. A distinction is made between clusters having 1, 2,3, 4 and 5 or more tweets.
- For clusters recognized by *kernel density estimation* the strength is calculated as the contribution (the level) of a single cluster to the sum of the level of each cluster of a single user. Clusters contributing less than 5%, 10% and 20% respectively to the sum of the levels are filtered out.

The results of the evaluation are presented in Figure 3; results only include users tweeting in Singapore or Singapore and overseas and tweeting 10 times or more. An intuitive result is found: if the threshold levels for a cluster's strength are set low, the number of clusters found by both methods is high; when setting thresholds value high a lower number of clusters is detected. If a minimum of cluster size of 4 is set for the k-means clustering, 44% of the users has only 1 cluster and 22% percent has 2 clusters. If a minimum contribution level of 20% is set for KDE, 67% of the users has only 1 cluster; 80% of the users has more than 1 cluster if a minimum contribution level of 5% is set. From this the relationship between the threshold to set and the number of clusters becomes apparent. If the goal is to determine the number of frequently visited locations a thresholds will need to be set. However, if the goal is determine a users activity space it is possible not to set thresholds and by doing so, not deleting user information.

Figure 3 also allows for a comparison with travel survey data. Respondents with only 1 cluster included retirees, homemakers and domestic workers. Over 50% of the respondents has 2 clusters. The applied cluster methodologies detect more activity locations than are reported.

Clusters detected by means of KDE and using a threshold of 10% are compared against Singapore's population statistics *(19)* and estimated work locations *(20, 21)*. The results of the comparison are presented in Figure 4. Compared are the number of users with one or more clusters against the population (top) and the number of work locations (bottom). It can be seen that the percentage of detected clusters in several zones matches the population percentage in several planning zones, most notably in the planning zones Bukit Timah, Novena, Marine Parade, Kallang and Queenstown. The first three planning zones are known for the high percentage of private property and correspondingly higher income. A further distinction by age and income is necessary to further analyse potential Twitter users. The Downtown Core has the highest number of work locations; however not the highest percentage of Twitter users. Both the shopping district Orchard and the airport Changi have a high number of Twitter users clusters as compared to the number of workplaces.

One of the advantages of social network data is that the costs of collecting records for a longer time span are virtually free. Figure 5 shows the number of tweets collected, as a proxy for time, versus the number of number of clusters recognized with different thresholds for the number of clusters when using KDE. Only users tweeting 10 times or more have been included. The left-most plot shows a counter-intuitive result: despite the high number of tweets not a high number of clusters is recognized. The three other plots show the number clusters detected with different thresholds for the level of contribution. While a high number of tweets is required to detect 1 or more clusters, the effect of a high number of tweets per user on detecting the number of clusters per user is limited.

A second advantage of social network data is that the collection of data is not limited by geographical boundaries. Earlier the number of users and tweets in Singapore and outside of Singapore has been presented (Table 1). In Table 2 a breakdown is presented of the number

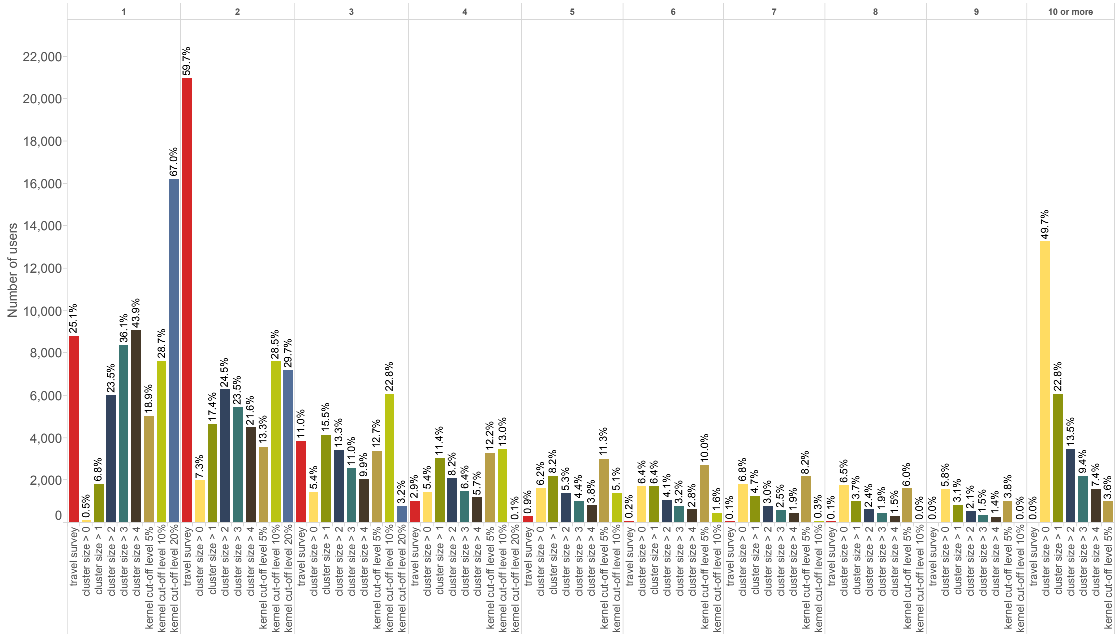

FIGURE 3 Highlighted is the number of clusters detected per user when different criteria are set for the strength of a cluster; as a reference the number of locations in the travel survey is taken. The top row indicates the number of clusters detected (1,.., 10 or more). The bars indicate the number of users having this number of clusters. Each pane contains the same criteria for cluster strength. For k-means clustering the number of tweets belonging to a cluster can vary between 1 and more. For kernel density estimation a contribution level of 5%, 10% and 20% per cluster to the sum of all level per user is used as a threshold.

TABLE 2 Breakdown of the number of user with a cluster applying a kernel density estimation with a threshold of 10%. Indicated is if a user only has clusters in Singapore or both in Singapore and overseas. Johor Bahru is across the border from Singapore and accessible by foot, car, frequent bus services; Batam is accessible by ferry.

Country	Region	Only Singapore	Singapore and overseas
Singapore	Singapore - all	14,628	3,944
Malaysia	Johor Baharu		1,517
	Other Malaysia		39
Indonesia	Batam		426
	Other Indonesia		27
Thailand	Thailand - all		67

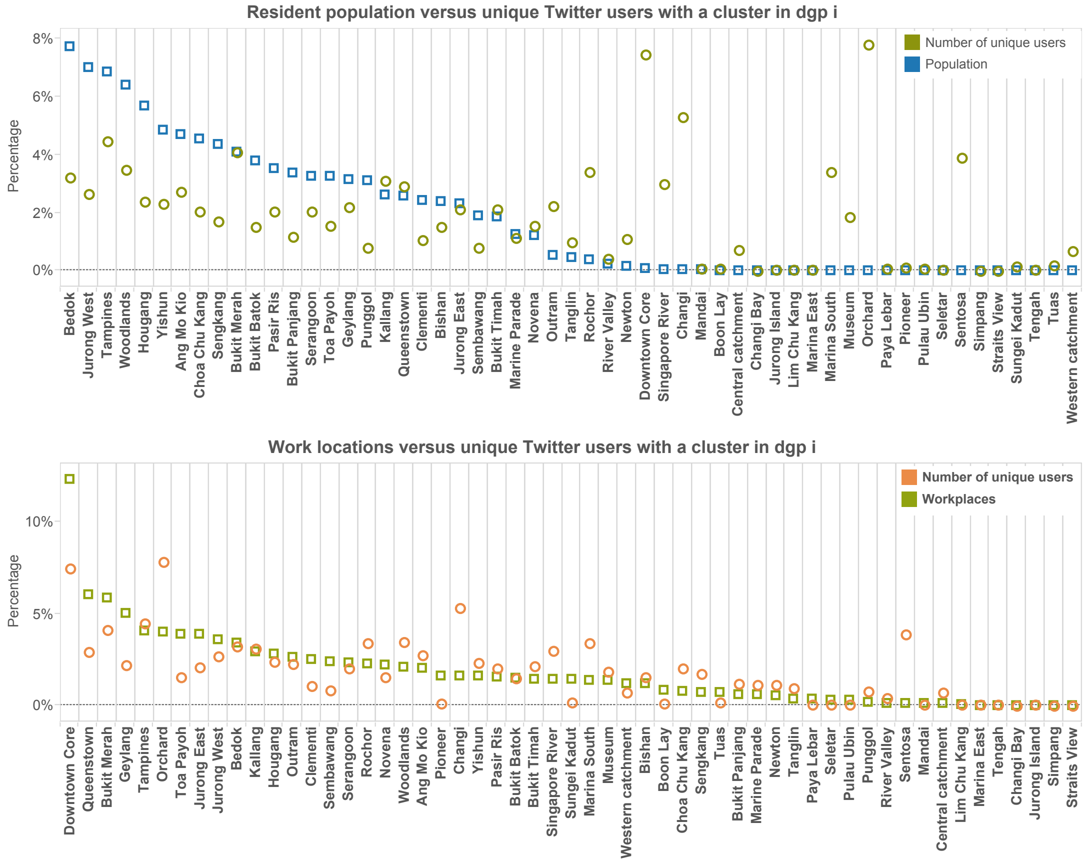

FIGURE 4 **Percentage of unique users with a cluster in planning zone *i* plotted against the population (top) and work locations (bottom). To give an indication of the absolute figures: Bedok has a resident population of 589,038 according to the 2010 Singapore census; the number of detected work locations in the downtown core amounts to 185,000.**

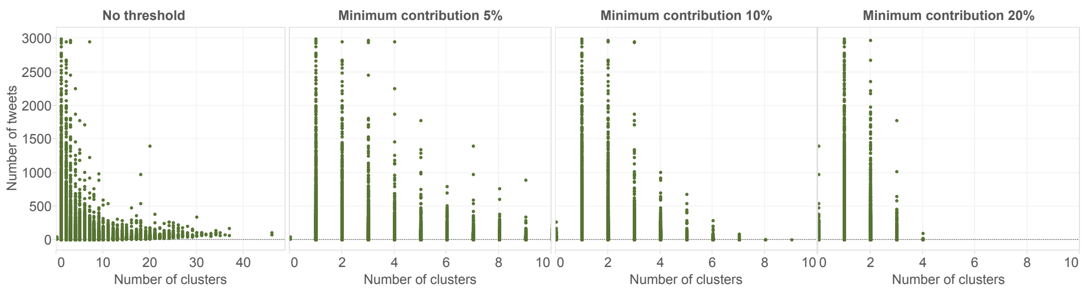

FIGURE 5 **Depicted is the number of tweets versus the number of clusters detected; each point represents a user. The left-most scatter plot shows the case where no threshold is set for the contribution of a single cluster to the total level of user; the three other scatter plots present the results for a 5%, 10% and 20% threshold respectively.**

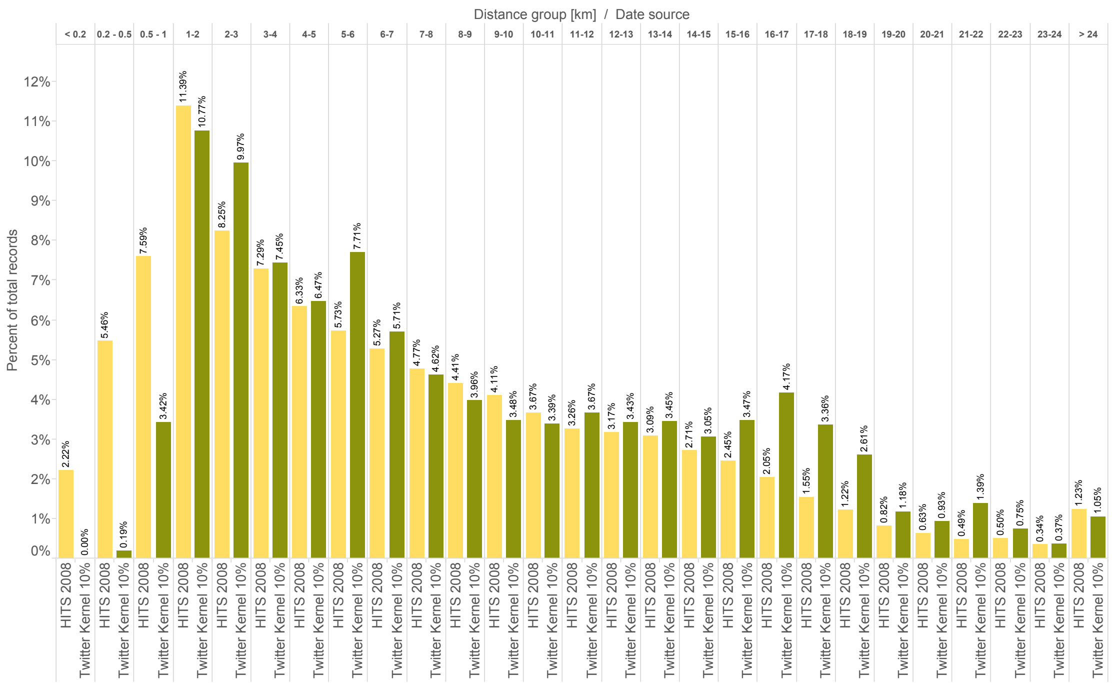

FIGURE 6 Comparison of distances between reported activity locations in the household interview travel survey 2008 (26,422 persons) and activity locations detected in Twitter by means of kernel density estimation (17,930 users).

of users with clusters only in Singapore, and in Singapore, Malaysia, Indonesia and Thailand. Clusters are detected with KDE and a threshold of 10%. Almost 4,000 users have a cluster in Singapore and outside of Singapore. The majority of these users have one or more clusters in the province adjacent to Singapore, Johor Bahru.

Comparing distances

In addition to the visual inspection of clusters and assessing the total number of cluster per user, we compare the the distances between clusters detected in Twitter and distances between reported activity locations in the Singapore household interview travel survey (HITS) 2008. To assess whether the distances between different data sources correspond for both data sources all the Euclidean distances between all unique reciprocal locations per user are calculated. For example, if user reports trips to three distinct locations (e.g. home, work, leisure) we calculate the distances between home-work, home-leisure and leisure-work. A similar procedure is followed for clusters detected in Twitter by means of KDE with a threshold value of 10% as a reference case.

In Figure 6 the results of the distance comparison are presented. It can be seen that the distances between activity locations in both data sources correspond very well for most distance categories. However, 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%).

Detecting transitions

The third comparison involves comparing origin-destination matrices derived from public transport smart card data with transitions observed in Twitter data. Origin-destination matrices from smart card data are derived from journey start and end transit stops; no attention is being paid to transfers. For instance, consider a user traveling travel from zone A to zone C with a transfer in zone B. The travel from A to C is considered a journey. However, the user could have transferred in zone B and is thus required to tap his card if the transfer involves a bus trip. This transfer is left out of the analysis. In order to analyze Twitter data according to similar definitions, we take as a basis clusters detected with kernel density estimation and apply a threshold of 10%. Tweets located within the contour of the kernel are considered to be part of the cluster. Subsequently, all tweets of each user are ordered by time to determine common transitions between locations. By doing so, we assume that transitions will occur from time to time between these activity locations. For instance, consider a user having two clusters with each 2 tweets. Tweet 1 is created on July 5, 9am and belongs to cluster 1 in zone X, tweet 2 is created on July 6, 10am and belongs to cluster 2 in zone Y. The user's movement from zone X to zone Y is counted as a single transition. A limitations of this approach is that other possible transitions of this user, that occur outside of the measured location-based social network (tweets), are not measured.

In Figure 7 the transitions as calculated from detected locations with KDE and a threshold of 10% versus public transport smart card data (top) and household interview travel survey data (HITS, bottom) per planning zone. Intra-zonal and weekend trips have been excluded. To compare the results from both data sources, 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 trend similar to both smart card data and HITS. The correlation coefficient between HITS and smart card data is 0.88 and the p-value associated with the fit is less than 10^{-3}, the correlation coefficient between HITS and Twitter is 0.71 and the p-value associated with the fit is less than 10^{-3} and the correlation coefficient between smart card data and Twitter is 0.76 and the p-value associated with the fit is again less than 10^{-3}.

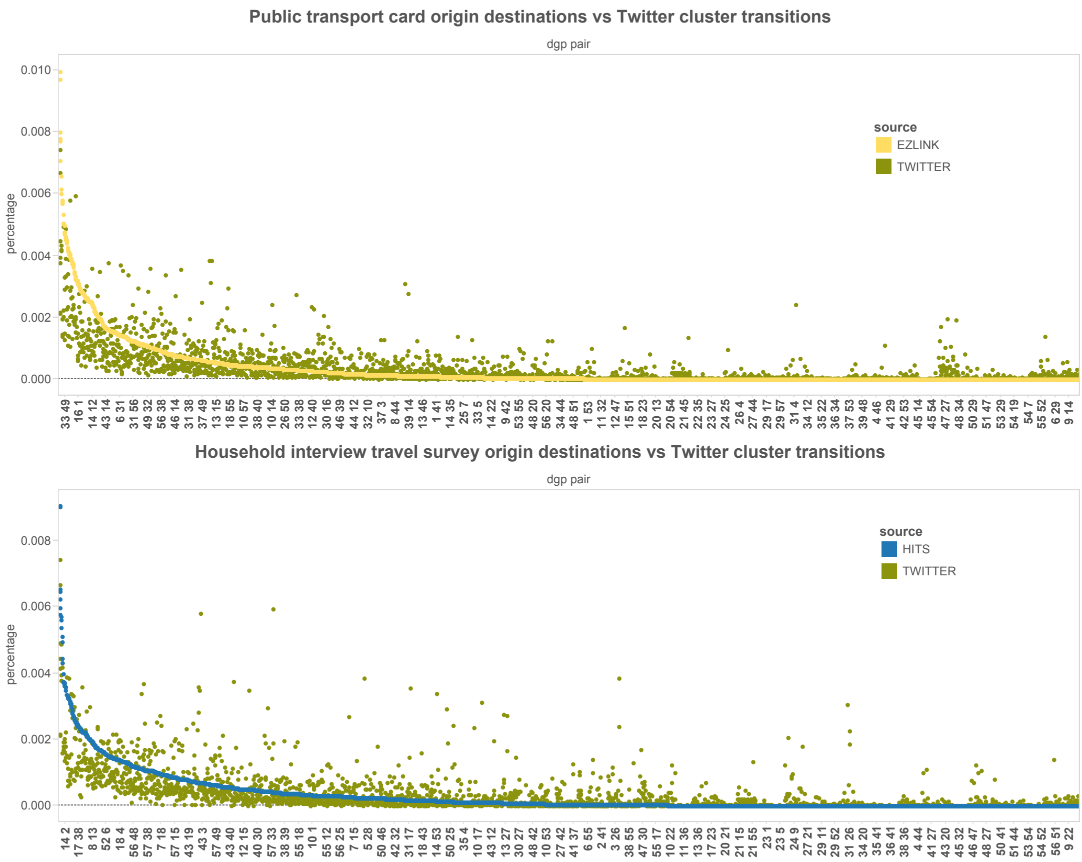

FIGURE 7 Transitions as calculated from Twitter versus weekday public transport smart card data (top) and household interview travel survey data (bottom) journeys per planning zone pair. Intra-zonal trips have been excluded. The relative flow per origin-destination pair is shown. Records are sorted by the percentage per od-pair from public transport smart card data and household interview travel survey data respectively.

DISCUSSION & OUTLOOK

This paper has addressed the detection of an individual's activity locations from data from the social network service Twitter, the spatial separation between these locations and the transitions between these clusters. Whereas previous work *(3, 4, 10)* has only considered a subset of this data, namely location check-ins, we include all available data. While Twitter is sometimes considered 'Big Data' this can be considered relative to other data sources such as GPS. For Singapore we observe around 27,000 unique users tweeting 10 times or more in a time span of 8 months and correspond to less than 0.5% of Singapore's population. These users tweet 2 million times in total.

A first challenge lies in the distinction between en-route Twitter events and Twitter events at activity locations. The application of kernel density estimation for the detection of clusters, as proposed by *(13)*, yields more promising results than k-means clustering. The kernel density approach requires a bandwidth h. Setting a high value for h can result in over-smoothing. Translated to the detection of activity locations, this can result in a lower amount of detected locations in each others proximity. A second parameter setting concerns the goodness-of-fit of a cluster. Due to the lack of speed information, as is the case with GPS data, to filter en-route events a threshold is required. This threshold not only filters en-route events but also less frequently visited locations. From the comparison between the detected cluster and population statistics it can either be deduced that Twitter events occur less frequent at home locations and/or that Twitter users form only sub-sample of the population; in several homogeneous planning zones however a match between detected clusters and population statistics can be observed. A further distinction of population statistics by age and income remains for further work. A similar, but less pronounced trend could be observed when comparing detected locations with work locations. Two planning zones, the main shopping area Orchard and the airport Changi, show a higher amount of detected locations. Also open for further work remains the inclusion of the temporal component in the clustering algorithm *(e.g. 10)*.

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. Whether this is due to over-smoothing or the fact that primary school students are not active on Twitter is open for discussion. As not only clusters in Singapore are detected but also clusters in neighbouring countries insight is gained in transborder traffic. The transitions between planning zones correspond to public transport smart card and travel survey data. As a next step a cut-off time will be introduced: tweets being more than n hours apart will be discarded.

Despite these remaining questions, 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. As another potential data source for capturing transport data we see mobile phone applications such as Strava and Moves. As such, the results are similar to a GPS-based travel survey. However, the user base which can be touched upon is many times larger.

ACKNOWLEDGEMENTS

The research conducted at the Future Cities Laboratory is funded by the Singaporean National Research Fund and the ETH Zurich. The authors would like to thank Professor Kay W. Axhausen for his suggestions and comments. We wish to express our gratitude to the Land Transport Authority for providing us invaluable data sets on transport in Singapore. Also we are very thankful to the Urban Redevelopment Authority for providing us with a wide range of data sets.

REFERENCES

1. OpenStreetMap (2011) The Free Wiki World Map, webpage, `http://www.openstreetmap.org`. Accessed 25/05/2011.

2. Google (2014) What is GTFS?, `https://developers.google.com/transit/gtfs/`.

3. Hasan, S., W. Lafayette and S. V. Ukkusuri (2013) Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media, paper presented at the *UrbComp*, ISBN 9781450323314.

4. Noulas, A., S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo (2012) A tale of many cities: universal patterns in human urban mobility, *PloS one*, **7** (5) e37027, ISSN 1932-6203.

5. Ortúzar, J. d. D. and L. G. Willumsen (2011) *Modelling Transport*, 4th edn., John Wiley & Sons, Chichester.

6. Axhausen, K. W. and T. Gärling (1992) Activity based approaches to travel analysis: Conceptual frameworks, models and research problems, *Transport Reviews*, **12** (4) 323–341.

7. Pelletier, M.-P., M. Trépanier and C. Morency (2011) Smart card data use in public transit: A literature review, *Transportation Research Part C: Emerging Technologies*, **19** (4) 557–568, ISSN 0968090X.

8. Bagchi, M. and P. White (2005) The potential of public transport smart card data, *Transport Policy*, **12** (5) 464–474, ISSN 0967070X.

9. Zheng, Y. (2011) Location-based social networks: Users, in Y. Zheng and X. Zhou (eds.) *Computing with Spatial Trajectories*, chap. 8, 243–276, Springer New York.

10. Hasan, S. and S. V. Ukkusuri (2014) Urban activity pattern classification using topic models from online geo-location data, *Transportation Research Part C: Emerging Technologies*, **44**, 363–381, ISSN 0968090X.

11. Pan, B., Y. Zheng, D. Wilkie and C. Shahabi (2013) Crowd Sensing of Traffic Anomalies based on Human Mobility and Social Media Categories and Subject Descriptors, paper presented at the *SIGSPATIAL GIS '13*, Orlando, FL, USA, ISBN 9781450325219.

12. Wei, L., Y. Zheng and W. Peng (2012) Constructing popular routes from uncertain trajectories, paper presented at the *KDD '12*, 195, Beijing, China, ISBN 9781450314626.

13. Lichman, M. and P. Smyth (2014) Modeling Human Location Data with Mixtures of Kernel Densities, paper presented at the *KDD '14*, New York, NY, USA, ISBN 9781450329569.

14. Grossenbacher, T. (2014) Studying Human Mobility Through Geotagged Social Media Content, Master thesis, University of Zurich.

15. Haklay, M. (2012) 'Nobody wants to do council estates' - digital divide, spatial justice and outliers, `http://povesham.wordpress.com/2012/03/05/nobody-wants-to-do-council-estates-digital-divide-spatial-justice-and-outliers-aag-2012/`.

16. Twitter (2014) Twitter Reports First Quarter 2014 Results, `https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=843245`.

17. Twitter (2014) About Twitter, `https://about.twitter.com/company`.

18. Prakasam, S. (2008) The Evolution of e-payments in Public Transport - Singapore's Experience, *Japan Railway & Transport Review*, (50) 36–39.

19. Department of Statistics Singapore (2010) Census of Population 2010, *Technical Report*, Singapore.

20. Chakirov, A. and A. Erath (2012) Activity identification and primary location modelling based on smart card payment data for public transport, paper presented at the *13th International Conference on Travel Behaviour Research (IATBR)*, Toronto, July 2012.

21. Ordóñez Medina, S. A. and A. Erath (2013) Estimating dynamic workplace capacities by means of public transport smart card data and household travel survey in Singapore, *Transportation Research Record*, **2344**, 20–30.

22. Jain, A. K. (2010) Data clustering: 50 years beyond K-means, *Pattern Recognition Letters*, **31** (8) 651–666.

23. Figueiredo, M. A. and A. K. Jain (2002) Unsupervised Learning of Finite Mixture Models., *IEEE transactions on pattern analysis and machine intelligence*, **24** (5) 381–396, ISSN 0162-8828.

24. Hansen, M. and B. Yu (2001) Model selection and the principle of minimum description length, *Journal of the American Statistical Association*, **96** (454) 746–774.

25. Tibshirani, R., G. Walther and T. Hastie (2001) Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63** (2) 411–423, ISSN 1369-7412.

26. Rasmussen, C. E. (2000) The Infinite Gaussian Mixture Model, in S. Solla, T. Leen and K.-R. Muller (eds.) *Advances in Neural Information Processing Systems 12*, 554–560, MIT Press.

27. Rousseeuw, P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics*, **20**, 53–65, ISSN 03770427.

28. Ashbrook, D. and T. Starner (2003) Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing*, **7** (5) 275–286, ISSN 1617-4909.

29. Silverman, B. (1986) *Density estimation for statistics and data analysis*, no. 1951, Chapman and Hall, London.

30. Hall, P., S. J. Sheather, M. Jones and J. S. Marron (1991) On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation, *Biometrika*, **78** (2) 263, ISSN 00063444.

31. Sheather, S. and M. Jones (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society. Series B*, **53** (3) 683–690.