IDENTIFYING HIGHLY DENSE AREAS FROM RAW LOCATION DATA

ELIAS JAKOBUS WILLEMSE[1], BIGE TUNCER[2] and
ROLAND BOUFFANAIS[3]
[1]*Singapore University of Technology and Design, and University of Pretoria, South Africa*
[1]*ejwillemse@gmail.com*
[2,3]*Singapore University of Technology and Design*
[2,3]*{bige_tuncer|bouffanais}@sutd.edu.sg*

Abstract. In this paper we show how very high-volumes of raw WiFi-based location data of individuals can be used to identify dense activity locations within a neighbourhood. Key to our methods is the inference of the size of the area directly from the data, without having to use additional geographical information. To extract the density information, data-mining and machine learning techniques form activity-based transportation modelling are applied. These techniques are demonstrated on data from a large-scale experiment conducted in Singapore in which tens of thousands of school children carried a multi-sensor device for five consecutive days. By applying the techniques we were able to identify expected high-density areas of school pupils, specifically their school locations, using only the raw data, demonstrating the general applicability of the methods.

Keywords. ; Machine Learning, Big-data, Location-analysis.

1. Introduction

The large-scale adoption of location-tracking capable devices, such as cell phones and some whereables, have given researchers access to large volumes of human mobility and activity data. In this paper, we demonstrate how such big data can be leveraged to identify high-density common-locations within urban areas. Such information provides key insight into locations that are close to saturation and may be in need of re-design and expansion to cater for forecasted population growth. Another potential use is in identifying formal communal locations with low densities which are likely in disuse. Such locations can then be directly studied and the cause of their disuse, such as an inconvenient location or poor design, can be investigated and corrected, or at the least avoided in future developments.

To extract the density information, we apply data-mining and machine learning techniques, commonly used in activity-based transportation modelling (Schuessler and Axhausen, 2009), on high volumes of time-stamped location data. The techniques are applied in four steps. First, velocity and proximity rules are applied to distinguish between travel point-sequences and activity point-sequences.

Intelligent & Informed, Proceedings of the 24th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2019, Volume 2, 805-814. © 2019 and published by the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

Next, a commonly used machine-learning clustering algorithm, is applied over the activity sequence-points to identify potential commonly used, thereby dense areas. The convex hull of the cluster points is then calculated and used to approximate the physical size of the areas. Thereafter, the number of unique visitors and total time spent in them is calculated, which in turn, is used to calculate the densities. An important benefit of this approach is that no external information on the location area is required to calculate densities. The density calculations can be repeated at different times of the day while keeping the size of the area constant. The calculated time-based densities can then be used to find and compare dense common areas, and analyse the typical density experienced during non-transport activities over the course of a day.

2. Related research

Within the urban and transportation planning communities there has been a variety of models to study human and city dynamics and mobility patterns (Gonzalez et al., 2008; Liao et al., 2007). Among them, and especially in transportation planning, activity-based models are considered as the state-of-the-art whereby travel demand is seen as a direct result of people wanting to participate in and conduct activities in certain places at given times (Jiang et al., 2017; Pinjari and Bhat, 2011). The uptake of activity-based modelling is a direct result of the large volumes of human activity location data now readily available to planners. Such big data can be indirectly obtained through mobile call data (Chen et al., 2017; Zheng, 2015) or more directly through GPS-enabled devices (Schuessler and Axhausen, 2009). More recently, the movement of people has been successfully traced by triangulating their location through the proximity of their mobile phones to known WiFi access points (Fakhreddine et al., 2018). The data generated by all the above methods give a detailed view of people's movement over time which can be processed to identify the timing, location and duration of their key activities.

Similar to the aim of this paper, Vieira et al. (2010) deal with the problem of dense areas detection where individuals concentrate within a specific geographical region and time period. The authors develop the Dense Area Discovery (DAD-MST) algorithm to automatically detect dense areas in cell phone antenna networks using Call Detail Records. A benefit of their approach is that the underlying shape of dense-areas is directly inferred from antenna networks. Their approach is demonstrated in a non-disclosed city and used to identify dense areas that are insufficiently covered by the public transport network.

Huang et al. (2017) follow an even more detailed approach to track visitors' movement through the Vanke Songhua Lake Resort, located in Jilin City, northeast of China. The authors make use of WiFi Indoor Position System and track the spatial distribution of visitors through time, and infer the peak-times of specific shops and streets within the resort. For their study, detailed information on the underlying geographical area of the resort and points-of-interest are known beforehand. There is thus a research gap to identify dense locations using detailed location information, but in cases with little information on the underlying geographical areas. For such a scenario, points of interest will have to be inferred directly and solely from detailed location data.

Identifying points of interest from raw GPS records is a common first step in activity-based transportation modelling. Schuessler and Axhausen (2009) develop rule-based methods that split locations of individuals into activity and transport based records. The authors demonstrate the effectiveness of their methods using location data from a sample of residents in Switzerland. In a second step, the transportation mode-choices between activities are analysed (Montini et al., 2014). Similar studies have been conducted in Singapore on large volumes of location data of students (Monnot et al., 2016; Tan et al., 2018; Wilhelm et al., 2017). It should be noted that in these studies, the research focus is on analysing the travel behaviour of individuals. Understandably, little attention is given to the geospatial overlap between activities of residents.

Joubert and Axhausen (2013) follow the activity-extraction approach, specifically to study the overlap of locations of commercial freight vehicles in South Africa. Their aim is to use commonly visited areas to identify formal links between different freight carriers. Commonly visited areas are identified using density-based clustering methods, and formal links are established between vehicles when they visit the same location. Joubert and Meintjes (2015) show that these locations coincide with transportation and delivery hubs, as well as high-density freight demand points, such as shopping centres. A benefit of their approach is that locations are inferred without using additional information.

With our aim to use raw location data of people to identify highly-dense areas, the reviewed approaches can be adapted and applied as follows. The methods of Schuessler and Axhausen (2009) can be used directly to extract activity locations. Thereafter, key statistics for the activities, such as the duration, can be extracted. Commonly visited activity locations can then be determined using density-based clustering methods, and the cumulative time spent at these locations can be calculated. Lastly, the physical shape of the commonly visited locations can be inferred based on the convex-hull of original activity location points associated with them. The physical size of the activity space and cumulative time spend within it then can be converted to the number of visitors per time-period per m^2.

3. Study area and data

The techniques of the previous section were adapted and tested on a large-scale real-world dataset consisting of 133 million records, collected as part of the National Science Experiment (NSE) project in Singapore. The data has been used by a large number of researchers and public agencies to improve their understanding of complex urban systems, mostly from transport, environmental and sensor technology perspectives (Happle et al., 2017; Monnot et al., 2016; Tan et al., 2018; Wilhelm et al., 2017). The dataset was generated through the deployment of 50,000 wearable sensors to Singaporean school children. Students would collect the devices on a Monday and wear it until Friday. On the following Monday, the devices were distributed to new school children. Thereby each week's data capture activities of different students. The data that we used was captured over 8 weeks between April and August of 2016.

When carried, the device continuously measured environmental data, such as temperature, and importantly for our study, its location. When the device was completely stationary it recorded at a 1-hour frequency. Otherwise, it recorded at a 14-second frequency. The location of the device was localised through surrounding WiFi access points and reported to be accurate to around 20 meters (Wilhelm et al., 2016). Weekends, as well as Mondays and Fridays, were not considered during the analysis since the devices were not in constant use on these days. Furthermore, for demonstration purposes and to limit the computational burden of the analysis, device records were considered for consecutive Wednesdays only, reducing the dataset from 133 to 33 million records. Since a new set of students was tracked in each week, the combined records represent a large sample of the activities of students on a typical weekday. Activities were extracted for active devices. For validation purposes, the clustering and density analysis were performed on three neighbourhoods, namely Toa Payoh, Jurong East and Punggol. The available records cover between 5% and 10% the school going population within each area, and between 0.5% and 1.8% of the total population.

There are some important limitations to the data which we acknowledge. First, the data only represents school children, therefore their activities mainly consist of home and school activity anchor-points, with some minor activities occurring elsewhere. Dense locations are expected to be dominated by areas in and around the schools through which the devices were distributed. Second, the proportion of students covered is small, especially when estimating absolute totals, such as density. The totals can be inflated based on the sample proportion, but care must be taken due to bias in the data. Third, the data is limited to weekdays, and we expect dense locations to be different on weekends. Lastly, activities may take place over different floors within buildings, as investigated by Tuncer et al. (2017). Any calculations of density must be interpreted with caution as the participants may be spread over multiple floor-levels. Despite these limitations, the data still allows for the demonstration of the techniques and the results give some insight to the density experienced by school children over the course of the day.

4. Methodology

In this section, we describe in detail the methods used to estimate the density of high-use common areas. The input required for the methods is time-stamped location coordinates for a large sample of people, with a unique identifier for each person's records. From here on, a record refers to a single time-stamped location-reading of a person participating wearing the device. The extraction and analysis proceeds in three main steps described in the rest of the section. All the methods were implemented in Python 3.7 with the scikit-learn package (Pedregosa et al., 2011) used for the clustering step.

The methods that we applied to identify activities is based on the approach of Schuessler and Axhausen (2009), which was designed for raw data from global positioning systems, without additional information. The methods takes as input each participant's time-ordered set of location records, $R = \{r_1, \ldots, r_n\}$, where n is the total number of records of the participant. Each record, $r_i = \{(x_i, y_i), t_i\}$, consists of the coordinate pair, (x_i, y_i), which we assume is in meters, of the participant's location at time t_i. Depending on the format of the input data, the location may need to be converted from geographic coordinate system to a projected coordinate system, making it possible to directly calculate Euclidian distances in meters between points, without having to use the Haversine-formula. As a preliminary filtering step, points falling outside the geographical borders of the area being considered can already be removed.

To account for location inaccuracies, which are unavoidable with location tracking devices, a Gauss kernel smoothing function with a kernel bandwidth of 10 seconds is applied to all the location positions. Two rules are then used to identify activity sequences, consisting of consecutive points i to j. In rule 1, a sequence of points is flagged as stationary activity points when the speed between the points is lower than 0.01 m/s and their total duration is at least 120 seconds. For rule 2, a sequence of points is flagged as dense activity points when they occur spatially close to each other.

Once the activities have been identified, the next step is to cluster activities that occur relatively close to each other, while discarding those the occur in isolation. The Density-Based Spatial Clustering with Applications of Noise (DBSCAN) (Ester et al., 1996) was ultimately chosen for our application. The algorithm has two input parameters, namely, ϵ, which is the maximum distance between points to be considered as part of the same cluster, and p_{min}, the number of points required to form a cluster. For our application, both were set equal to ten.

Once the activities are clustered, the final step is to calculate the number of unique visitors, the number of activities, and the total time spent by participants on their activities within the clusters. Doing so is straight-forward using the calculated attributes of the activity clusters. Using the start-and-end times of activities, the number of activities occurring in specific time-interval can be also calculated, which translates to the number of people within the cluster area for the time interval under consideration. Proportional allocations can be made for the time intervals in which activities start and end. The only metric that still needs to be calculated to estimate the density of the activity-cluster is its physical size in m^2. To do so, we revert back to the original activity location points and calculate the convex hull of the points. The area of the convex hull of the points represent a rough physical boundary in which the activities took place, and the area of the convex hull is used as a proxy for the cluster's area. The size of the common activity location is then used to measure its density in people per m^2.

5. Results

In this section, we present full results for activity extraction over all participants records, as well as clustering and density calculations for the three study areas. As mentioned, the data used to test the methods consisted of 33 million records, recorded over consecutive Wednesdays. During the activity extraction phase, a total of 441,438 thousand activities were identified from 53,251 active participants. Key distributions for the activities are shown in Figure 1. The median number of activities per person was seventeen, whereas the median duration of activities was 9 minutes. Both distributions are right skewed, with durations being more so. This reflects the difference between short bursts of activities occurring during the day while the participants are more active, and long home activities when the participants are inactive during the night.

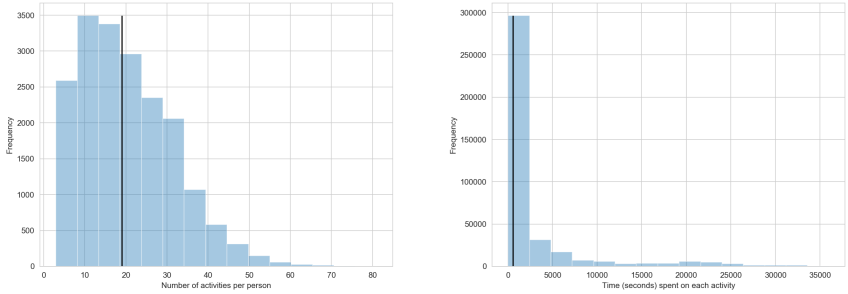

Figure 1. Distribution and median for the total number of activities extracted per participant and the total duration of each activity.

The next step was to cluster the activities using the DBSCAN algorithm on the three study towns. We first evaluated the impact of the input parameters of the algorithm, namely ϵ, the maximum distance in meters between points to be considered as part of the same cluster, and p_{min}, the number of points required to form a cluster. Both were ranged from 5-50 at intervals of 5 units and the resulting number of clusters produced and the fraction of points clustered, with the rest considered as outliers, were captured for the 25 parameter combinations. Results for the experiments are shown in Figure 2.

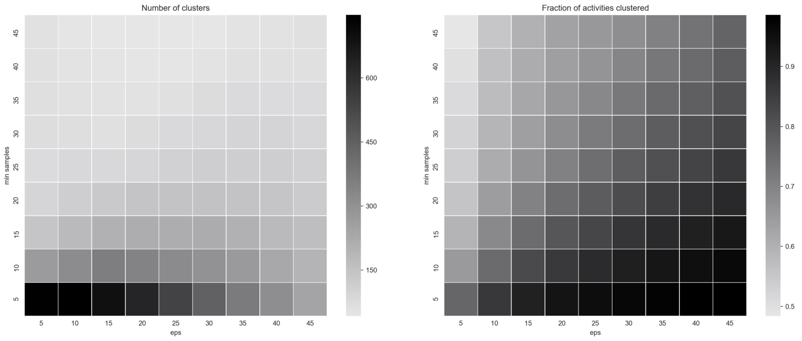

Figure 2. Number of clusters and fraction of points clustered at different levels of epsilon and p min.

At $p_{\text{min}} \geq 15$, the algorithm becomes insensitive to the parameter values, while still finding the same limited number of clusters. The only change is in the fraction of activities that are included in the clusters. In conjunction with lower values of ϵ, more activities are considered as outliers, which were expected. The experiment shows that the clustering outcome is more sensitive towards $p_{\min}$. The minimum number of activities to justify a common-location depends on the density of the area being considered, as well the proportion of participants included in the study. It, therefore, has to be critically evaluated when applied in semi-similar settings.

Since the data only considers a small fraction of participants, relatively small values of $\epsilon = 10$ and $p_{\min} = 10$ were used to extract clusters for the final analysis. The clusters still totalled well over 300. Prior to identifying high-density locations, all clusters with less than ten unique visitors were removed. Our reasoning was that locations with less than 10 visitors cannot be considered as commonly visited, and in a preceding round of experiments it was observed that they have very low densities. Of the 300 clusters, only 33 met this criterion and where further analysed. Another round of clustering with $\epsilon = 10$ and $p_{\min} = 10$ was applied to original activity points of the clusters to remove outliers, thereafter the area of the convex-hull of remaining points was calculated. The total duration, in days, of the activity durations was calculated. Since the analysis focussed on a single day, the total duration in person-days can interpreted as the number of people throughout the day within the cluster. This value together with the convex hull of the area was used to calculate density. The first four graphs in Figure 3 shows distributions for the number of unique visitors for the clusters, the total duration, in days, of all activities within them, their final area size in m^2, and finally their density in people per m^2.

As expected, the number if unique visitors in the clusters were quite high, with a median of just under eighty. The ten clusters with more than 100 visitors all match schools that participated in the study. Although not explicitly showed, the total duration of the activities per cluster is directly correlated with the number of unique visitors. The cluster areas were also quite small with a median value of less than 1000 m^2, representing a 50-by-20 meter areas. The large clusters again coincide with areas within participating schools. The density of the locations was also quite small, with the largest being 0.08 people per m^2. The median density is only about 0.005. Where the values coincide with school locations we deem them to be accurate, which includes the highest observed density. For other locations, adjustments can be made to account for the less than 2% sample size included in the analysis. At the very least, the analysis results show potentially dense areas which can be further investigated through direct observation or using more representative data.

The last analysis conducted was to measure the density per hour-of-day. Results for the 33 clusters are shown in the bottom of Figure 3. As expected, there is a rise in the density of the locations from 07:00 as participants gather in either respective schools, thereafter the density stays fairly consistent until 14:00 at which point it decreases as students move to other less student-dense location. There is another minor spike later in the evening, indicating another batch of activities. Coinciding with the thick-right tail of the original density distribution, there are two locations within schools with high-density levels throughout the course of the day, including the very early morning hours. These points coincided with small areas with a few continuously active devices, indicating that the ten-unique visitor limit was too lax. Despite these limitations, the results show to the expected density trends, allowing us to recommend with caution, and subject to further parameter analysis, that they can applied to larger and more representative datasets.

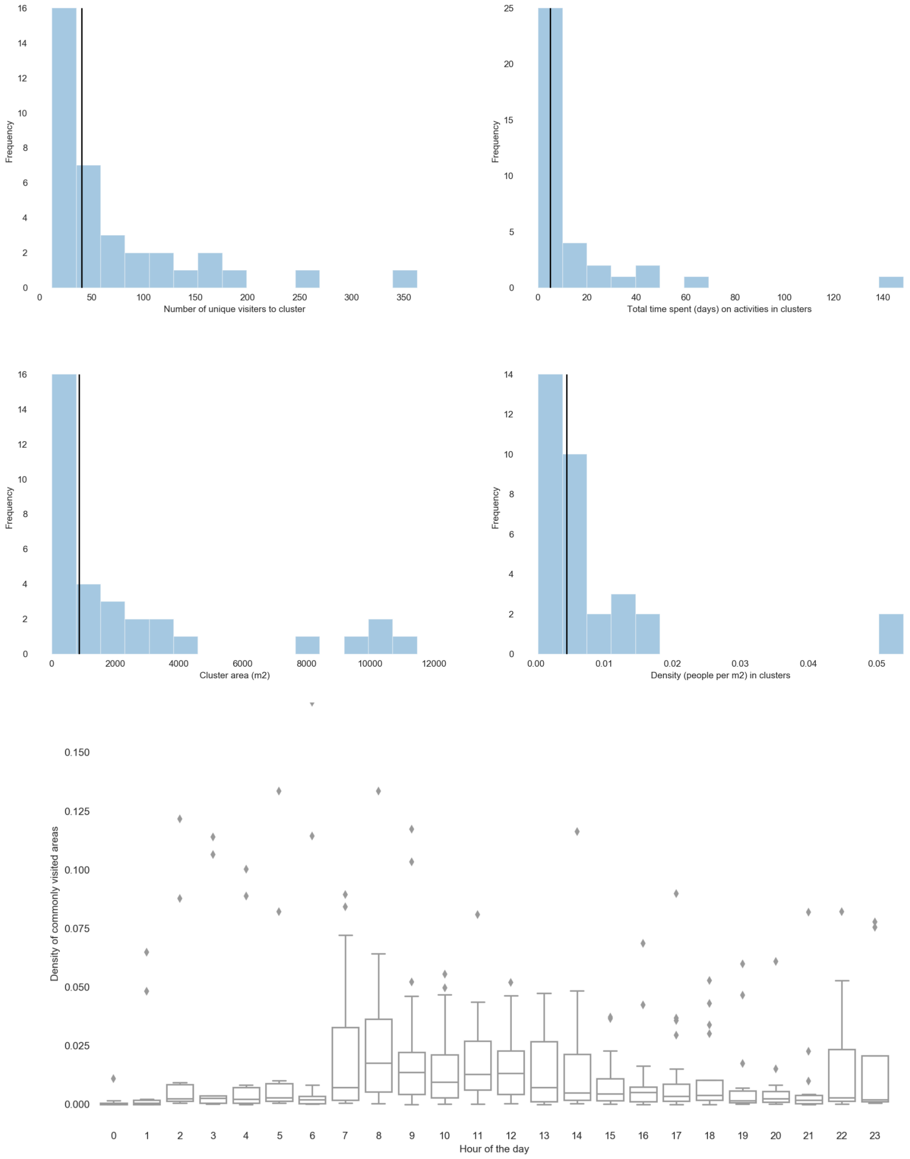

Figure 3. Distributions for the number of unique visitors per cluster, the duration of activities within it, the size of the clusters, and the density of the clusters (top four figures), and distribution of the density of clusters over the course of a day (bottom figure).

6. Conclusion

Our aim with the study was to investigate whether commonly applied techniques from activity-based transportation modelling could be used to identify high-density activity locations, and further to quantify the density levels. The methods were applied to raw data from school pupils in Singapore, and the initial results showed that high-density areas can be extracted and analysed over the course of a day. The methods consisted of a rule-based activity extraction step, followed by clustering using the DBSCAN algorithm. Thereafter the convex-hulls of the activity points associated with commonly visited clusters were calculated to estimate the size of the areas. A benefit of the methods is that they are relatively straight-forward to implement with intuitive parameters that can be tuned for specific application areas. In terms of generality, the scale of the data obtained through study does limit the generality of the techniques. Still, the techniques may be of use in other settings, specifically where similarly high-volumes are obtained through cellphone devices. As is usually the case with an initial investigation of this type, there exist many opportunities for improvement. First, the activity extraction parameters were taken directly from literature and better values and more advanced methods can be investigated. Second, the methods were tested on a small subset of the available data, and the observed results may be indicative of random fluctuations in the travel behaviour of participants. A full application on the whole dataset and all areas in Singapore will produce more robust results. Lastly, the methods can be applied to more comprehensive datasets with ground-truth data to validate that they are effective. The actual floor-space of known popular areas can also be compared against the calculated convex-hull area to analyse its accuracy.

Acknowledgement

This research is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister's Office under the Land and Liveability National Innovation Challenge (L2 NIC) Research Programme (L2 NIC Award No. L2NICTDF1-2017-4). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Singapore Ministry of National Development and National Research Foundation, Prime Minister's Office, Singapore.

References

Chen, N., Xie, J., Tinn, P., Nagakura, T. and Larson, K.: 2017, Data Mining Tourism Patterns - Call Detail Records asComplementary Tools for Urban Decision Making, *Protocols, Flows, and Glitches: Proceedings of the22nd International Conference on Computer-AidedArchitectural Design Research in Asia (CAADRIA 2017)*, 685-694.

Ester, M., Kriegel, H.P.a. and Xu, X.: 1996, A density-based algorithm for discovering clusters inlarge spatial databases with noise, *Proceedings of the Second International Conference onKnowledge Discovery and Data Mining (KDD-96)*, 226-231.

Fakhreddine, A. and Tippenhauer, N.O.a.: 2018, Design and Large-Scale Evaluation of WiFi ProximityMetrics, *European Wireless 2018; 24th European WirelessConference*, 1-6.

Gonz'alez, M.C. and Hidalgo, C.A.a.: 2008, Understanding individual human mobility patterns, *Nature*, **453**(7196), 779.

Happle, G., Wilhelm, E. and Schlueter, A.: 2017, Determining air-conditioning usage patterns in Singapore from distributed, portable sensors, *Energy Procedia*, **122**, 313-318.

Huang, W., Lin, Y. and Wu, M.: 2017, Spatial-Temporal Behavior Analysis Using Big DataAcquired by Wi-Fi Indoor Positioning System, *Protocols, Flows, and Glitches: Proceedings of the22nd International Conference on Computer-AidedArchitectural Design Research in Asia (CAADRIA 2017)*, 745-754.

Jiang, S. and Ferreira, J.a.: 2017, Activity-based human mobility patterns inferred frommobile phone data: A case study of Singapore, *IEEE Transactions on Big Data*, **3**(2), 208-219.

Joubert, J.W. and Meintjes, S.: 2015, Repeatability \& reproducibility: Implications ofusing GPS data for freight activity chains, *Transportation Research Part B: Methodological*, **76**, 81-92.

Liao, L., Patterson, D.J. and Fox, D.a.: 2007, Learning and inferring transportation routines, *Artificial Intelligence*, **171**(5-6), 311-331.

Monnot, B., Wilhelm, E.a., Zhou, Y.a., Lu, H.Y. and Jin, W.: 2016, Inferring Activities and Optimal Trips: Lessons FromSingapore's National Science Experiment, *Complex Systems Design \& Management Asia*, Cham, 247-264.

Montini, L., Rieser-Sch"ussler, N.a. and Axhausen, K.W.: 2014, Trip purpose identification from GPS tracks, *Transportation Research Record*, **2405**(1), 16-23.

Pedregosa, F., Varoquaux, G., Gramfort, A.a., Thirion, B., Grisel, O.a., Prettenhofer, P., Weiss, R.a., Vanderplas, J., Passos, A.a., Brucher, M. and Perrot, M.a.: 2011, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825-2830.

Pinjari, A.R. and Bhat, C.R. 2011, Activity-based travel demand analysis, *in* R.L. Andr'e de Palma (ed.), *A Handbook of Transport Economics*, Edward Elgar Publishing, 1-36.

Schuessler, N. and Axhausen, K.: 2009, Processing raw data from global positioning systemswithout additional information, *Transportation Research Record: Journal of theTransportation Research Board*, **2105**, 28-36.

Tan, S.B., Zegras, P.C.a. and Arcaya, M.C.: 2018, Evaluating the effects of active morning commutes onstudents' overall daily walking activity inSingapore: Do walkers walk more?, *Journal of Transport \& Health*, **8**, 220-243.

Tuncer, B., Benita, F. and Tay, H.: 2017, Identification of building floors in a 3D citymodel, *Smart Cities: Improving Quality of Life Using ICT \&IoT (HONET-ICT), 2017 14th International Conferenceon*, 16-20.

Vieira, M.R., Frias-Martinez, V.a. and Frias-Martinez, E.: 2010, Characterizing dense urban areas from mobilephone-call data: Discovery and social dynamics, *2010 IEEE Second International Conference on SocialComputing (SocialCom)*, 241-248.

Wilhelm, E., MacKenzie, D., Zhou, Y.a. and Tippenhauer, N.O.: 2017, Evaluation of transport mode using wearable sensordata from thousands of students, *Proceedings of Annual Meeting of the TransportationResearch Board (TRB)*, 1-18.

Wilhelm, E., Siby, S., Zhou, Y.a., Jayasuriya, M.a., Kee, J.a. and Tippenhauer, N.O.: 2016, Wearable environmental sensors and infrastructure formobile large-scale urban deployment, *IEEE Sensors Journal*, **16**(22), 8111-8123.

Zheng, Y.: 2015, Trajectory data mining: an overview, *ACM Transactions on Intelligent Systems andTechnology (TIST)*, **6**(3), 29.