Constructing Passenger Trips and Comparison of Regional Travel Patterns Using Crowdsourced Transit Data

Hyungsub JEE[1] and Jan-Dirk SCHMÖCKER[2]

[1]PhD Student, Department of Urban Management, Kyoto University
(Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: hjee@trans.kuciv.kyoto-u.ac.jp
[2]Associate Professor, Department of Urban Management, Kyoto University
(Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8540, Japan)
E-mail: schmoecker@trans.kuciv.kyoto-u.ac.jp

We discuss the possibility to obtain long-term travel behavior information through an open-source, crowdsourced, low-energy data collection effort. We utilize the "One Bus Away" travel planner that is used in several regions of the U.S. From users willing to participate, locational data are obtained via Google's "Android Activity Transition Application Programming Interface" that collects records as to when the user is changing the mode of movement among the possible states {still, walking, running, cycling, in-vehicle}. At a transition point the time and location are recorded. We discuss completeness problems of the data, and cleaning efforts needed to construct trips and tours. We then show differences in behavior over the course of two years for six regions in the U.S., where the application is used: Tampa, Puget Sound, San Diego, Washington, D.C., Rogue Valley and Spokane. Given limitations in our data we still obtain some interesting observations: We show that travelers in San Diego make more in-vehicle trips than travelers in Tampa and Puget Sound. Vehicle usage (versus non-motorized trips) is in Washington, D.C. are more seasonal dependent and has been possibly more impacted by COVID. By matching trip start and end points with bus stops we further illustrate that public transport serves the travel demand better in Puget Sound than in Tampa. People in Washington, D.C. appear to travel also to vary their destinations more over the course of the year. We conclude by discussing the possibilities of this data set, including replicability in other cities and further data analysis.

Key Words : *GPS Tracking, Transit Application, OneBusAway, Travel Activity, Regional Comparison*

1. INTRODUCTION

The Obtaining data on travel behavior remains a key question for the transport planning community. In particular for understanding the behavior of public transport users, new technologies have been leading to a range of new data collection methods. "Passive big data" through electronic fare collection have been at the forefront of researchers' interest; but also Wi-Fi sensors, Bluetooth Low Energy (BLE) tags and bus GPS traces have been used to understand vehicles loads, travel patterns, and the quality of the service provided[1)-3)].

Smart card data and other electronic fare collection data can be collected without additional contributions directly from passengers and can provide information on the trips taken by passengers. However, the data do not provide any information regarding his/her activities before or after the transit part of the journey from one place of activity to another. This means that, for example, wait times at bus stops and choices as to when, where, and how a user selected a specific transit route must be inferred. Analytical or simulation-based models on arrival time at bus stops, hence, often rely on assumptions or are based on re-call surveys[4)-6)].

Alternatively, many researchers have collected tracking data via dedicated applications. Commercial applications and other tools developed by researchers provide the feasibility to obtain detailed locational records. The data come, however, at a cost and are seldom collected on a large scale and for a large number of users. Battery drainage and privacy issues are at the forefront of concerns why users do not want their behavior to be recorded in detail over longer

time periods.

In this paper, we discuss the data analysis of an experiment with OneBusAway where, compared to other tracking exercises, users do not receive incentives, and data are collected relatively sparse but over a longer time period. We discuss the data cleaning, problems, and potential of the data. We illustrate the data by a comparative analysis between six US regions where data have been collected since late 2019 and we derive some initial results.

The remainder of this paper is organized as follows. After the literature review, we first explain OneBusAway (OBA) user tracking data in further detail and then define how passenger trips can be constructed using such data including the necessary data cleaning process. Section 4 presents the results comparing regional travel patterns using passenger trips constructed in Section 3. Finally, this paper is concluded in Section 5 with discussion of potential future studies.

2. LITERATURE REVIEW

Tracking individual riders provides a way to obtain detailed information as to where and when people conduct activities and how they travel to these locations. There is a rich set of literature on this topic based on a range of applications that allow such tracking. Maruyama et al[7] developed a tool applied in Kumamoto, Japan, that combines tracking with a smartphone-based survey. Using highly-effective incentive measures, they collected a sample of about 1000 travelers and used their tool to complement household survey data. Lue and Miller[8] estimated pedestrian route choice based on smartphone GPS data using an application that records the location every 50m and with a relatively small sample they could show that the observed route choice resembles other data collection procedures. Marra et al[9] instead describe the development of the application itself. They discuss algorithms to identify trips, activities and mode usage. A key aspect of their work showed that lower sample frequencies and collecting locational data when requested from other phone applications can reduce the problem of battery drain.

Rout et al[10] review a total of 96 papers that have used GPS tracking data with the objective to understand the interaction between pedestrians and their environment. They show that the topics and technology are still rapidly evolving. One of their conclusions is that it is possible that GPS data will allow longer-term observations of individuals, a topic that we pick-up with in the present study.

With respect to trip and tour tracking over larger areas, the focus of research has been on tourism activities. A main reason for this is that tourist data are generally not covered in detail in household surveys and must therefore be obtained in other ways. In some cases, the locational data are obtained from the mobile phone provider and in some cases via application records. Abkarian et al[11] characterize visitors of a large-scale event based on GPS traces, Sugimoto et al[12] more broadly provide behavioral patterns in a local urban tourist destination. Choe et al[13] discuss behavioral patterns in a national park based on locational records as well as surveys. We note that in general the fusion of GPS traces with other data sources appears to be important as also discussed in Gao and Schmöcker[14]. They merge their relatively sparse GPS records with Wi-Fi sensor readings to obtain travel pattern information of tourists in Kyoto. Kressner[15] merged electronic trip diaries with consumer marketing data.

Possibly because (academic) results are not yet convincing to practitioners and because apps that allow easy processing of tours from GPS traces are not yet available, Too et al[16] concluded that travel agencies were generally often not yet willing to invest in app development and data. Though the study was conducted four years ago and the field is rapidly developing, this seems to be still true today. Harding et al[17] also ask the question whether dedicated, commercial smartphone applications can replace travel surveys. They review 17 applications and conclude that the accuracy of recorded locations is in most cases fairly satisfactory, but that the mode inference is still often lacking. They suggest that in order for smartphone data to replace survey data, some additional input via data fusion with other data sources or with “pop-up” questions is needed. In line with this, Schmücker and Reif[18] conclude that GPS records from mobile network data are often preferable compared to app-based GPS data.

The problem that GPS data require sometimes significant cleaning before being useful is also highlighted in Abbruzo et al[19]. They propose data preprocessing algorithms to deal with outliers and to not miss main points of interests during a tour, a problem that also motivated the aforementioned studies on fusing GPS traces with other data sources.

Besides dedicated tracking applications, locational records are increasingly collected by all kind of applications. Some researchers have used “Google Location History” data obtained from users who allowed Google to track their locational records and are willing to share their data with the analyst[20)-22)]. Cools et al[21] conclude that the data is not (yet) of sufficient accuracy to replace travel surveys.

Nevertheless, “tracking as a side business” appears to be increasing despite security concerns[23]. Even

during app-installations of games and other applications where the need to provide locational records is not obvious, pop-up questions as to whether one is willing to provide locational records frequently appear. Such data are, however, rarely available to the transport analyst but are only exploited for commercial purposes. It is further not clear in how far the data are analyzed systematically for commercial interests to obtain travel patterns of the user. Commercial applications will be more interested in understanding whether specific locations are visited.

For journey planner applications the request to share locational records is understandable to users. We note that several journal planner applications ask the user if they are willing to share their information for data analysis purposes. In Kyoto, for example, the "Arukumachi" application provides (static) transit information and also asks users for their consent to be tracked and share the data for non-commercial analysis purposes. The data are, however, not very fine grained as often only relatively few data points for a user are available and such limited data do not allow a researcher to infer the above-mentioned decision making[24].

In our study we aim to further analyze the usefulness of locational records only collected in the background. We therefore consider a more challenging problem than obtaining tracking data that is already highly detailed. We explore data from an experiment with the OneBusAway (OBA) application[25]. The tracking data are based on Google's "Android Activity Transition Application Programming Interface" (AATAPI) which is triggered by OBA users who are willing to participate in the experiment.

The AATAPI provides a range of new possibilities. Barbeau and Cetin[25] report on the experiment and compare the records obtained from the app with manually collected ground truth data for a small sample of records. They find that generally there is scope to extract behavioral information from the data. A larger analysis of the collected data from the wider OBA users has not yet been conducted.

Further, to the best of our knowledge, the AATAPI application has not also yet been widely employed for travel behavior analysis within other applications. Researchers have used the API to detect sleep patterns and signs of depression[26] as well as to encourage general activity recognition if there are signs of a person being in-active for too long[27]. Aliannejadi et al[28] have further used the data from AATAPI in conjunction with other smartphone usage data to study people's smartphone usage patterns. The only transport application of AATAPI we find is an attempt to detect cycling trips with the API data, but also here no longer term data are obtained, nor is an attempt made to detect the whole daily trip patterns[29]. We hence believe that this is first study that analyzes travel pattern data collected from a large database, in several cities over a longer time period.

3. CONSTRUCTING PASSENGER TRIPS

(1) Datasets

Our "parent tool", OneBusAway (OBA) is an open-source mobile application for real-time public transportation information available in several cities mostly in the U.S. (see https://onebusaway.org[30] for details). Barbeau et al[31] describe the tool and its early usage expansion. **Fig. 1** shows screenshots from the application when viewing arrival information and a present location on the route. The OBA application accepts, stores, archives and interprets real-time bus location data in combination with transit schedules and other related data.



Fig. 1 OBA application usage such as viewing arrival information and a present location on the route

Starting in 2019 in the six regions where the OBA Android apps are used, a user who opens the OBA Android app for the first time will be asked by the app whether s/he is willing to participate in a study which aims to improve public transit. For those who agree, the raw multimodal travel behavior data will be collected in the background from the AATAPI. AATAPI generates point-type data based on the movement change of Android mobile devices. In the experiment with the OBA app[25], point-type data was converted to origin-destination link type data as described in **Fig. 2**. (**Fig. 2** is adapted from the previous study[25].) This study uses the data as raw data and an example of the data is shown in Fig. 3. Note that Google has a set of predefined activities including a "still activity". There is a very fine time granularity as the activity durations are often very short. This means, however, that to obtain travel statistics, often

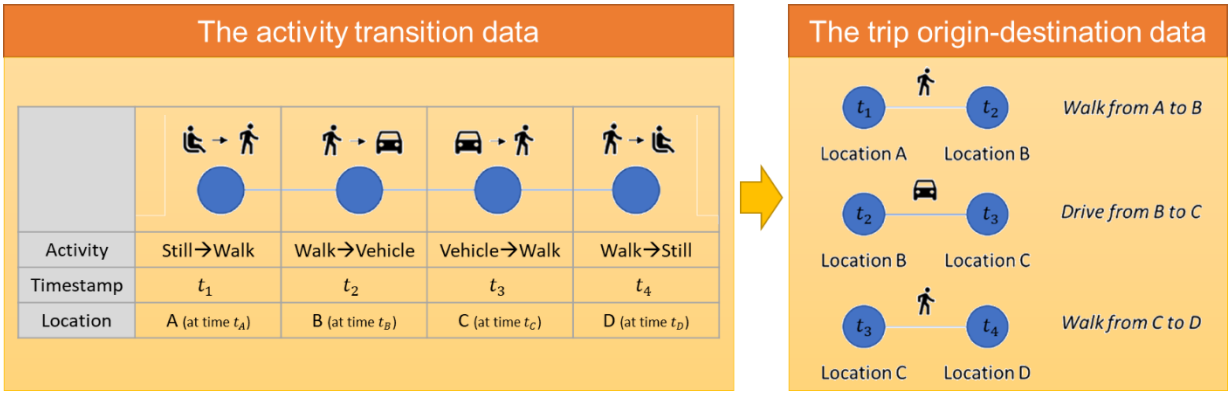


Fig. 2 Conversion from activity transition data to origin-destination data

User ID	Google Activity	Activity Start Date and Time* (UTC)	Origin location Date and Time (UTC)	Origin latitude	Origin longitude	Activity Destination Date and Time* (UTC)	Destination Location Date and Time (UTC)	Destination latitude	Destination longitude
JGxj3RcbvRTNA	STILL	2021-05-14T05:51:25Z				2021-05-14T19:18:38Z			
JGxj3RcbvRTNA	STILL	2021-05-14T19:22:04Z				2021-05-14T21:21:07Z	2021-05-14T21:21:15Z	47.62905	-117.248
JGxj3RcbvRTNA	WALKING	2021-05-14T21:21:07Z	2021-05-14T21:21:15Z	47.62905	-117.248	2021-05-14T21:27:44Z	2021-05-14T22:41:07Z	47.65693	-117.219
JGxj3RcbvRTNA	STILL	2021-05-14T21:27:44Z	2021-05-14T22:41:07Z	47.65693	-117.219	2021-05-14T21:31:36Z	2021-05-14T22:41:07Z	47.65693	-117.219
JGxj3RcbvRTNA	IN_VEHICLE	2021-05-14T21:52:41Z				2021-05-14T21:52:41Z	2021-05-14T22:41:07Z	47.65693	-117.219
JGxj3RcbvRTNA	IN_VEHICLE	2021-05-14T21:52:41Z	2021-05-14T22:41:07Z	47.65693	-117.219	2021-05-14T21:52:46Z			
JGxj3RcbvRTNA	WALKING	2021-05-14T21:52:46Z				2021-05-14T22:11:05Z			
JGxj3RcbvRTNA	STILL	2021-05-14T22:11:05Z				2021-05-14T22:11:25Z			
JGxj3RcbvRTNA	WALKING	2021-05-14T22:11:25Z				2021-05-14T22:21:13Z			
JGxj3RcbvRTNA	STILL	2021-05-14T22:21:13Z				2021-05-14T22:23:49Z			

Fig. 3 Example origin-destination data from a CSV file exported in a tabular format

records should be merged, as for example a walking trip might be recorded with several data entries as a user might have to stop at traffic lights. Furthermore, as can also be seen in **Fig. 3**, we observe a range of data gaps as well as data points where the location is missing.

We refer to this raw data as "Records". An elementary data record has an activity type $A_i \in$ Google activity type $G = \{$Still, Walking, Running, Walking/Running, In_vehicle, On_bicycle$\}$ with possibly incomplete information about starting and ending times and locations of a trips which we refer to as $(t_i^s, t_i^e, \boldsymbol{x}_i^s, \boldsymbol{x}_i^e)$. Activity transition data stores location information together with timestamps and activity transition information. Since location information is, however, handled separately in the Android device, the timestamp of location information differs from that of activity transition information.

(2) Data pre-processing

As a first step, the raw data are processed, including data cleaning and correction, to extract meaningful records. Typical data gaps include a time gap where activity transitions are temporarily omitted and a geospatial gap where only location information is temporarily omitted in records (but activity transition information still exists). Such omissions may have various causes, such as the location information being turned off by the user or a reboot of the device.

We first process the difference between the two timestamps mentioned in Section 3.1, then process the Records into "Stages" with an activity type $\tilde{A}_i \in G +$ {Unknown, Unknown-Still, Error, Mixed} and with completed information $(t_i^s, t_i^e, \boldsymbol{x}_i^s, \boldsymbol{x}_i^e)$. The additional categories are required to flag the uncertainty associated with the data cleaning. An example is if the timestamp of the location information has a difference of more than 1 minute to the time when the activity transition was recorded. (This is not seldom the case, presumably for preventing battery drainage by too often requesting GNSS information.) In that case, the location information is not trusted and treated as an incomplete record.

After that, the data set for each user is processed in five steps. First, if a "middle still" of less than 1 minute is stored between the same two activities, these three records are merged into one new record. Second, if incomplete records exist at the beginning or end before a long "still period", these records are removed so that the beginning and end of the data set can be complete records. Third, one or consecutive incomplete records are merged into one complete record using the location information of the preceding and following complete records. At this point, ge

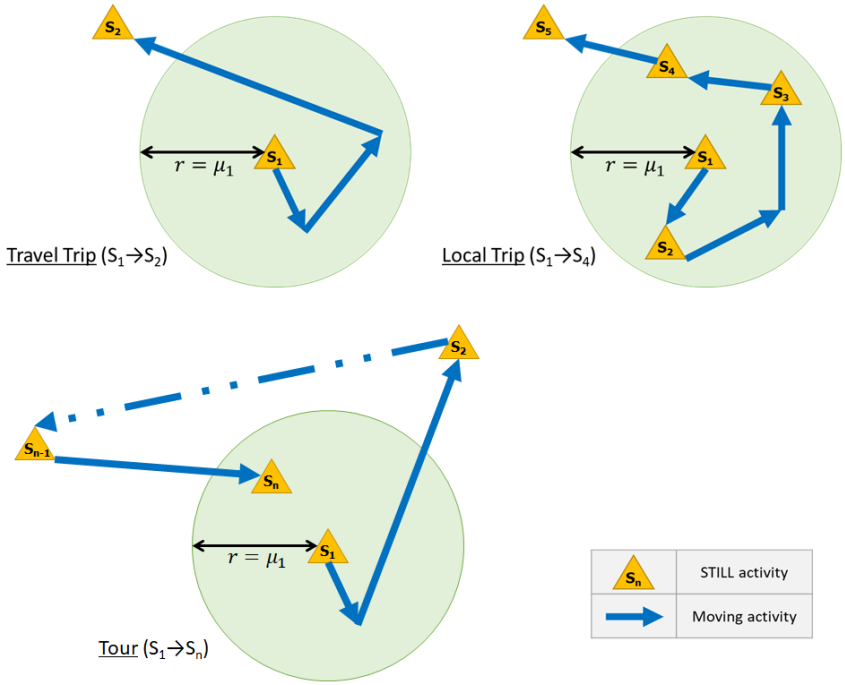


Fig. 4 Defining trips and tours

ospatial gaps have been eliminated, but time gaps remain in the data set. The remaining two steps deal with these: In the fourth step, a virtual record is created using the location information of the preceding and following complete records in place of the time gap to remove all time gaps. It is important to note that the analyses steps up to this point come at the "cost" that through these operations, the gaps are partially filled with Unknown, Unknown-Still, Error, and Mixed activity types. Finally, in the fifth step, if records of the same activity are consecutive, they are merged, resulting in stages. For example, two walk stages are only considered as separate trips if they are separated by a "still" record or another activity type.

(3) Local Trip, Travel Trip and Tour definitions

We then define the following categories into which the records are processed according to parameters ϑ_1 and μ_1 as defined below and shown in **Fig. 4**.

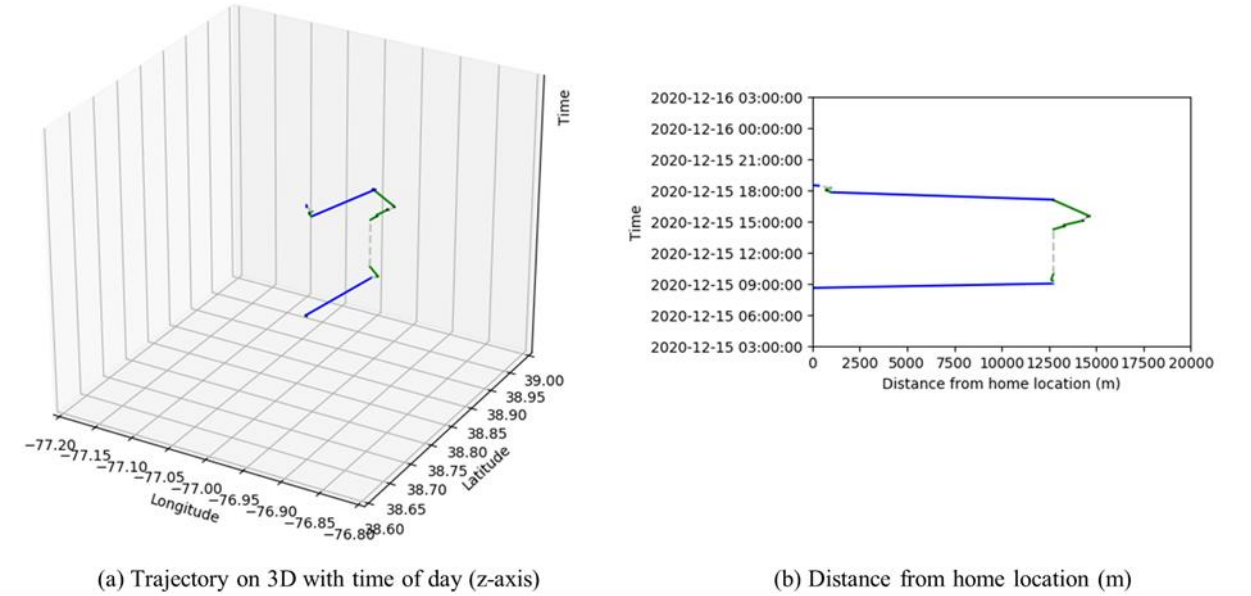
- Travel Trip: A set of stages from one Still activity > ϑ_1 to the next Still activity with > ϑ_1 and the distance between the location of these two still activities is > μ_1.
- Local Trip: A set of stages from one Still activity to another Still activity. The start location of the first and the location of each of the next Still activities are ≤ μ_1 and the stages do not belong to a travel trip.
- Tour: A set of 1⋯n trips with Haversine distance $d(\boldsymbol{x}_1^s, \boldsymbol{x}_n^e) \leq \mu_1$. Further, the distance between the first trip and at least one trip among the n trips must be > μ_1.

Our rationale to distinguish "Local trips" from "Travel trips" is that the former are trips within a complex, e.g., movements within a university campus, which would often not be reported in traditional surveys. Before defining trips and tours in each user data set consisting of only stages, the data set is divided once more into daily data sets starting at 3 am every day.

(4) Resulting user trajectories and creation of "potential bus trips"

Trips and tours are built for each user in each region using a ϑ_1 of 1 minute and a μ_1 of 200 meters. These values appear reasonable when comparing data to a few records for which we also obtained ground truth data (manually recorded location and time of activity transitions for comparison). The daily movement of passengers using OBA data can then be plotted as shown in **Fig. 5**. The user left the (presumably) home location in the morning, moved to a point about 1.25 km away in in-vehicle mode, then moved to a point about 2 km away on foot, and returned to the home location in the afternoon again in in-vehicle mode.

Given that the data are collected for a bus journey planner, we would like to distinguish bus trips from car trips within the set of "in-vehicle" records. To do so, we process the General Transit Feed Specification (GTFS) data of the cities for which we have obtained data and extract the location of bus stops. The trip



(a) Trajectory on 3D with time of day (z-axis)

(b) Distance from home location (m)

Fig. 5 One day illustrations of the data processing

Table 1 An example of the data processing and the resulting number of trips and tours

User id	Records	Stages	Travel trips	Local trips	Tours	Observed days	In vehicle trips	Potential bus trips
1	388	378	25	21	3	41	81	54
2	288	252	6	2	1	3	4	0
3	20802	19360	1000	292	150	312	1970	946
4	239	143	4	4	1	2	2	1
5	374	352	27	23	0	53	33	26
6	603	577	71	15	16	44	118	96
7	29	29	0	2	0	3	2	0
8	719	713	52	13	9	59	81	0
9	4207	4149	111	112	1	184	187	130
10	65	65	5	3	1	2	6	0

starting and ending points are compared to the bus stop locations. If both ends of an ”in-vehicle” trip are within 100 m of a bus stop, we refer to such an in-vehicle trip as a “Potential bus trip”. To confirm whether the trip might be indeed a bus trip would require us to compare the schedule and the boarding and alighting time of the user. We suggest, however, that “potential bus trip” itself is also a meaningful measure. A high ratio of potential bus trips indicates that users conducted many trips that could be conducted with the local public transport service. It is hence a measure of bus service accessibility.

Table 1 shows an example of the data processing for a few users from Washington, D.C., U.S. and the resulting number of trips and tours. The data illustrate that we observe a large variation in data amount for different users. In data analysis omitted here for brevity, we observe long tail distributions with respect to records per user. What is further obvious from Table 1 is that we observe, for many persons, only a fraction of the total trips made by the users. For example, for User 1, we observe 25 travel trips over 41 days for which the user provided records. This is likely an undercount and trips that were not recorded have been made in addition to the travel data recorded.

We refer to days for which we have any record for the user as “User-days”. Aggregating the results for Washington, D.C., we find that the average numbers per user-day are: 70 records, 15.9 stages, 1.5 travel trips, 0.8 local trips, 0.2 tours, 2.4 in-vehicle trips, and 0.9 potential bus trips, respectively. Further, 65.2% of passengers are not ending their day trip within 500m of their starting point. This is not in line

with our expectations and results from other travel surveys. We suggest a reason is the incompleteness of the data. This leads to few home-to-home tours per user-day. Increasing the threshold (ϑ_1, μ_1) for distinguishing travel trips from local trips clearly decreases travel trips but has an unclear effect on tours. Also for other regions we observe that on average, only less than 2 travel trips and 0.3 tours per user-day are identified in six regions and we clearly do not observe all trips. Nevertheless, sufficient trips and tours of transit users are constructed for further behavioral analysis.

4. REGIONAL COMPARISON

(1) Regions

OBA is available in multiple US regions but our experiment is limited to users from six regions: Puget Sound (greater Seattle), San Diego, Tampa, Rogue Valley (Southern Oregon), Washington, D.C. and Spokane (eastern Washington), U.S. The location of six regions is illustrated in **Fig. 6**. We use data from July 2019 to January 2022 in these six regions.



Fig. 6 Map of the U.S., showing the location of six regions

Table 2 The number of users and records

Region	Number of records	Number of users
Puget Sound	56,240,592	8,183
San Diego	4,916,284	1,672
Tampa	4,112,311	2,452
Washington, D.C.	740,032	140
Rogue Valley	343,999	183
Spokane	74,365	72

OBA data is still being collected in these regions and data after February 2022 will be applied to future studies. The number of registered users in each region and the total number of records collected are shown in **Table 2**. Further, the number of users observed for each quarter and the number of observed records per user-day for each quarter are shown in **Fig. 7**. Many locational records have been provided in the Puget Sound area due to the larger OBA user base in the region. More difficult to explain are the differences in user records per day. In particular in 2021, we obtain more information from users in Puget Sound and Washington, D.C. One explanation is that users in these cities use public transport and hence OBA more, which might also trigger more records to be collected.

(2) Results and discussion

The data constructed in Section 3 are processed to compare travel patterns between regions. To illustrate that we can obtain some comparative information as to how much users travel with motorized vehicles in the cities, we extract the percentage of user-days for which we obtain two or more in-vehicle trips. The curve for the percentage of user-days, aggregated by a quarter in Washington, D.C., exceeding the number of vehicle trips plotted on the x-axis is shown in **Fig. 8**. This is calculated using the average number of trips per user-day obtained, e.g., if a person provides some data for 10 days of the quarter, but at only 6 of these, we can identify two or more in-vehicle trips from the record, then the ratio for this person is 0.6 for the quarter. The implicit assumption is that on the other days of the quarter, the person did not have her/his mobile phone on a setting that allowed us to obtain data. **Fig. 8** illustrates that, with the exception of Quarter 3 in 2019, we only observe for around 50% of user days at least 2 in-vehicle trips. There is variation over time with the lowest number of in-vehicle trips per user day being observed in early 2020, likely due to Covid. The larger number of in-vehicle trips per user day in the early days of the experiment might be due to the first few users participating in the study being very active travelers.

Fig. 9 then shows the average for all users by city and quarter. To reduce the effect of "random, short-term users", we also included a figure for "long-term users" only. These are users who provide data for at least 200 user days. We note that this means the sample size is significantly reduced: Puget Sound has 247, San Diego has 55, Tampa has 38, Washington, D.C. has three, Rogue Valley has six and Spokane has only one long-term user. The reason why the long-term user rate in Washington D.C. is lower, given its generally large user number is not clear. One explanation might be that the OBA app is used more by short-term visitors in D.C.

For the fairly reliable data in **Fig. 9** (top) from Washington, Puget Sound and Tampa, we observe consistently higher rates of persons taking 2+ in-vehicle rides in San Diego than in Puget Sound and Tampa. For Washington we observe larger variation over time in the top figure with all users, which partly seem to be seasonal variations. We also note that we

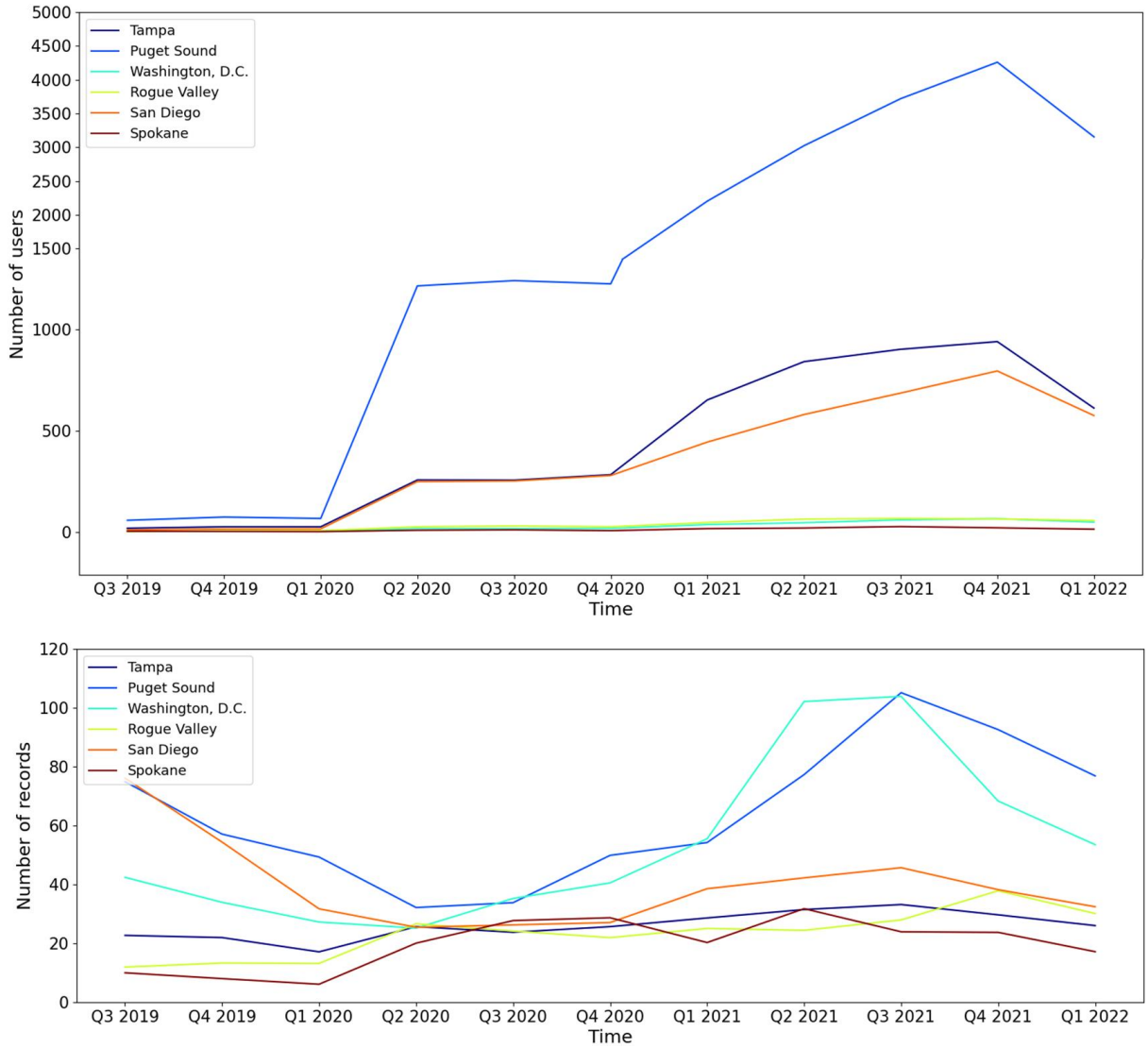


Fig. 7 The number of observed users (top) and records per user-day (bottom) for each quarter

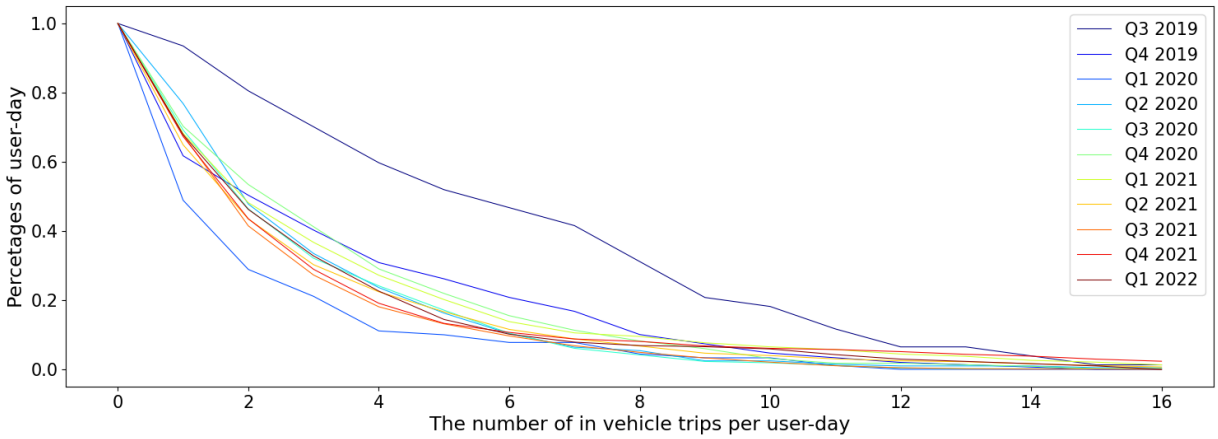


Fig. 8 Percentage of user-days with at least n vehicle trips in Washington, D.C.

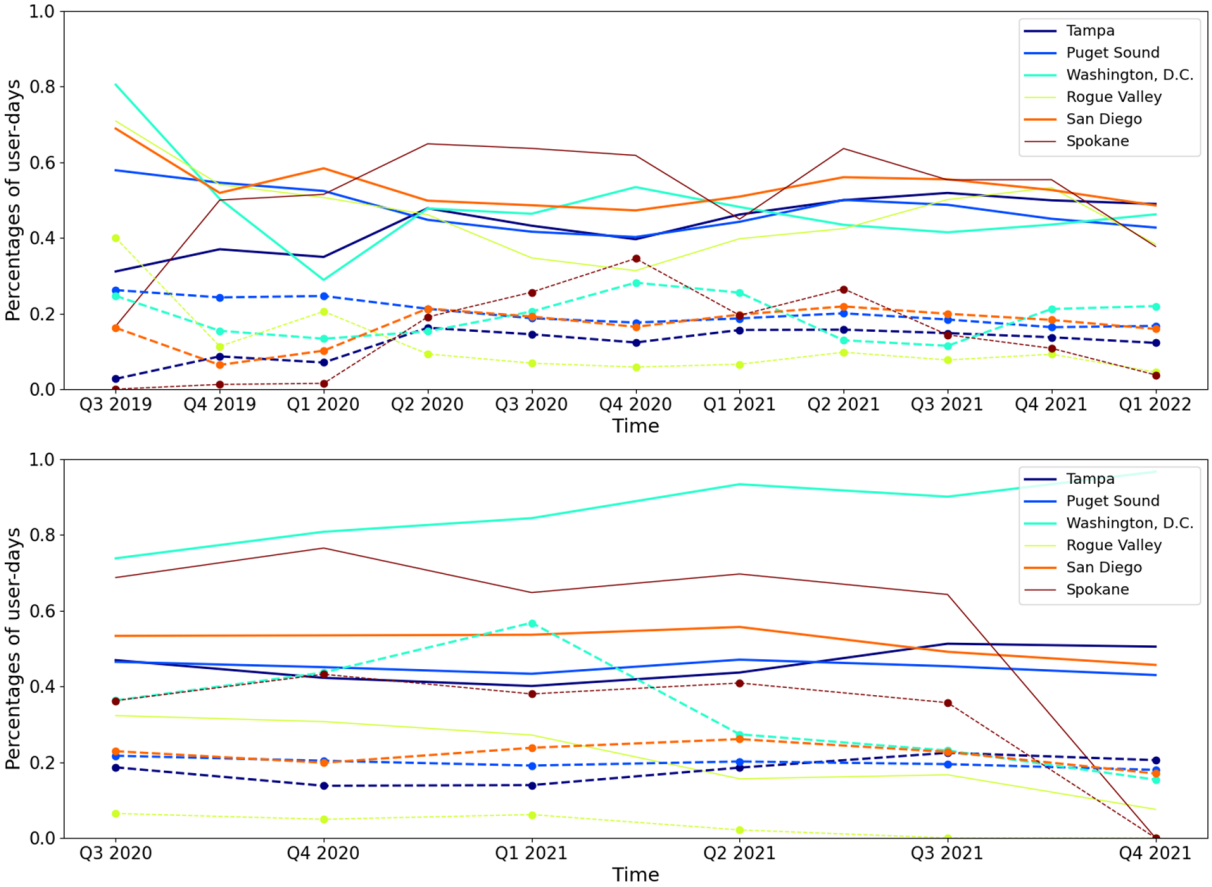


Fig. 9 Percentages of user-days with 2 or more in vehicle trips (solid) and potential bus trips (dashed). Top figure: all users; bottom figure: only long-term users

observe for D.C. a larger variation in the potential bus trips. This suggests that trip patterns in D.C. change more than in other cities. In the summer months, trips seem to be proportionally less to destinations close to bus stops. This might indicate higher usage of alternative modes as well as that more variety of destinations are visited. Among our three cities with largest sample sizes, we observe that the proportion of potential bus trips is consistently lower in Tampa compared to Puget Sound and San Diego. Interpretation of Spokane and Rogue Valley, the smaller cities in our sample, it need to be conducted with reservations due to the lower sample size. Notwithstanding, we observe high in-vehicle rates for Spokane and more seasonal variations, possibly because of it geographic location.

5. CONCLUSIONS

This study discusses a new method to obtain long-term tracking data from a fairly large user base. The data are obtained through Google's Android Activity Transition API which is embedded in our experiment in the OneBusAway application. We obtain data from users without providing any incentives and since the data are relatively sparse do not appear to negatively impact battery life of the users. This comes at the cost that the data quality is reduced. The low-priority data collection means that the GNSS records themselves are sometimes not reliable due to the activity transition record and the time-stamp and location record being obtained with at least a slight variation in time. We discuss the extensive data cleaning which allows us to create whole day trajectories but with some time periods marked as "unknown". Our approach fills in gaps where possible by utilizing the information of records before and after the incomplete one and merging records of very short time duration. Alternative machine learning approaches are also conceivable, where patterns might be inferred using records from other days to learn the missing information of a target day. Given the noise in our data, this is, however, a challenging problem.

Given the limited data, even with our approach, a significant number of trips are likely missing from the full trip pattern. We have to conclude hence that the data are not of sufficient quality to reliably obtain trip estimates for a population. Nevertheless, we suggest that for our cities with large databases of partial trips, some trends and changes over time are observ-

able if a stable database of users is maintained. In future work, we will conduct more detailed spatial-temporal analysis of usage patterns in the city and graphically illustrate these over the now four-year long data collection period. In addition, sufficient data are obtained to study change in longer term behavior in more detail for a small number of participants. Finally, and most importantly, this analysis is considered also the initial step to obtain detailed information as to how people use journey planner information. In addition to the location data of the users presented here, the information that users obtained from the OBA application has also been retained. This allows us to analyze the pre-trip decision making including when the user obtained information compared to when they began their journey and which bus stop and route the person chose among various options for a potential trip.

ACKNOWLEDGMENTS: This work was supported by JST SPRING, Grant Number JPMJSP2110. Thank you to the Open Transit Software Foundation Board, the maintainers of OneBusAway, for their financial and intellectual support of this project.

REFERENCES

1) Welch, T. F., & Widita, A. (2019). Big data in public transportation: a review of sources and methods. *Transport reviews*, 39(6), 795-818.
2) Sun, W., Schmöcker, J.-D. and Fukuda, K. (2021) Estimating the route-level passenger demand profile from bus dwell times. *Transportation Research Part C*, 130, 103273.
3) Schmöcker, J-D. (2021) Big Data for Public Transport Planning. In: Vickerman, Roger (eds.) *International Encyclopedia of Transportation*. vol. 5, pp. 134-139. UK: Elsevier Ltd.
4) Caulfield, B., & O'Mahony, M. (2007). An examination of the public transport information requirements of users. *IEEE transactions on intelligent transportation systems*, 8(1), (pp. 21-30).
5) Kisaka, M., Ahn, Y., & Tsukaguchi, H. (2014). Route Choice Associated with the Provision of Public Transportation Information. *Proceedings of International Symposium on City Planning 2014*.
6) Fonzone, A., Schmöcker, J. D., & Liu, R. (2015). A model of bus bunching under reliability-based passenger arrival patterns. *Transportation Research Part C: Emerging Technologies*, 59, (pp. 164-182).
7) Maruyama, T., Sato, Y., Nohara, K., & Imura, S. (2015). Increasing smartphone-based travel survey participants. *Transportation Research Procedia*, 11, (pp. 280-288).
8) Lue, G., & Miller, E. J. (2019). Estimating a Toronto pedestrian route choice model using smartphone GPS data. *Travel Behaviour and Society*, 14, (pp. 34-42).
9) Marra, A. D., Becker, H., Axhausen, K. W., & Corman, F. (2019). Developing a passive GPS tracking system to study long-term travel behavior. *Transportation research part C: Emerging Technologies*, 104, (pp. 348-368).
10) Rout, A., Nitoslawski, S., Ladle, A., & Galpern, P. (2021). Using smartphone-GPS data to understand pedestrian-scale behavior in urban settings: A review of themes and approaches. *Computers, Environment and Urban Systems*, 90, 101705.
11) Abkarian, H., Tahlyan, D., Mahmassani, H., & Smilowitz, K. (2022). Characterizing visitor engagement behavior at large-scale events: Activity sequence clustering and ranking using GPS tracking data. *Tourism Management*, 88, 104421.
12) Sugimoto, K., Ota, K., & Suzuki, S. (2019). Visitor mobility and spatial structure in a local urban tourism destination: GPS tracking and network analysis. *Sustainability*, 11(3), 919.
13) Choe, Y., Lee, C. K., Choi, J., Kim, M., & Sim, K. W. (2023). Identifying tourist spatial and temporal patterns using GPS and sequence alignment method. *Journal of Travel Research*, 62(6), 1181-1201.
14) Gao, Y. and Schmöcker, J.-D. (2023). Inferring travel patterns and the attractiveness of touristic areas based on Fusing Wi-Fi sensing data and GPS traces with a Kyoto Case Study. *102nd Annual Meeting of the Transportation Research Board*. Washington D.C., U.S.
15) Kressner, J. D. (2017). Synthetic household travel data using consumer and mobile phone data (No. NCHRP IDEA Project 184).
16) Too, W. G., Thong, C. L., Chit, S. M., Chaw, L. Y., & Lee, C. Y. (2020). Features of mobile tracking apps: a review of literature and analysis of current apps compared against travel agency requirements. In *Design, Operation and Evaluation of Mobile Communications: First International Conference, MOBILE 2020, July 19–24, Proceedings 22* (pp. 107-120). Springer International Publishing.
17) Harding, C., Faghih Imani, A., Srikukenthiran, S., Miller, E. J., & Nurul Habib, K. (2021). Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation*, 48, (pp. 2433-2460).
18) Schmücker, D., & Reif, J. (2022). Measuring tourism with big data? Empirical insights from comparing passive GPS data and passive mobile data. *Annals of Tourism Research Empirical Insights*, 3(2), 100061.
19) Abbruzzo, A., Ferrante, M., & De Cantis, S. (2021). A preprocessing and network analysis of GPS tracking data. *Spatial Economic Analysis*, 16(2), (pp. 217-240).
20) Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using Google Location History data to quantify fine-scale human mobility. *International journal of health geographics*, 17, (pp. 1-13).
21) Cools, D., McCallum, S. C., Rainham, D., Taylor, N., & Patterson, Z. (2021). Understanding Google location history as a tool for travel diary data acquisition. *Transportation research record*, 2675(5), (pp. 238-251).
22) Parady, G., Suzuki, K., Oyama, Y., & Chikaraishi, M. (2023). Activity detection with google maps location history data: Factors affecting joint activity detection probability and its potential application on real social networks. *Travel Behaviour and Society*, 30, (pp. 344-357).
23) Li, R., Diao, W., Li, Z., Yang, S., Li, S., & Guo, S. (2021). Android custom permissions demystified: A comprehensive security evaluation. *IEEE Transactions on Software Engineering*, 48(11), (pp. 4465-4484).
24) Kobayashi, H., Zhang, C., Schmöcker, J.-D., Nakao, S. and Yamada, T. (2021). Markovian Analysis of tourist tours based on Travel App Data from Kyoto, Japan. Presented at *25th International Conference of the Hong Kong Society for Transportation Studies (HKSTS)*, December 12-14.
25) Barbeau, S. J., & Cetin, C. (2021). Improving the Quality and Cost Effectiveness of Multimodal Travel Behavior Data Collection: A Case Study. *Transportation Research Board*

100th Annual Meeting, 32.
26) Hong, J., Kim, J., Kim, S., Oh, J., Lee, D., Lee, S., ... & Choi, Y. (2022, June). Depressive symptoms feature-based machine learning approach to predicting depression using smartphone. In *Healthcare* (Vol. 10, No. 7, p. 1189). MDPI.
27) Sun, J., Park, S., Jung, G., Jeong, Y., Lee, U., Chung, K. M., ... & Hadjileontiadis, L. (2020). BeActive: Encouraging physical activities with just-in-time health intervention and micro financial incentives. In *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures* (pp. 17-20).
28) Aliannejadi, M., Harvey, M., Costa, L., Pointon, M., & Crestani, F. (2019, March). Understanding mobile search task relevance and user behaviour in context. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (pp. 143-151).
29) Janoti, W., & Ferreira, P. (2021). detectBiklio: detecting bicycle usage with a smartphone app. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (pp. 900-907).
30) OneBusAway. https://onebusaway.org.
31) Barbeau, S. J., Borning, A., & Watkins, K. (2014). OneBusAway multi-region–rapidly expanding mobile transit apps to new cities. *Journal of Public Transportation*, 17(4), (pp. 14-34).