Transportation (2021) 48:2433–2460
https://doi.org/10.1007/s11116-020-10135-7

Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys

Chris Harding[1] · Ahmadreza Faghih Imani[2] · Siva Srikukenthiran[3] · Eric J. Miller[4] · Khandker Nurul Habib[4]

Published online: 20 August 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Given the limitations of traditional methods of data collection and the increased use of smartphones, there is growing attention given to using smartphone apps for activity-travel surveys. Smartphones, through their location-logging capability, allow for the collection of high-quality data on the travel patterns of individuals over multiple days while minimizing the burden on those being monitored. This paper presents the results of an investigation into the potential and limitations of smartphone apps as passenger travel survey instruments. It evaluates the accuracy and performance of various smartphone apps using properly recorded 'ground truth' data. Through an open and global invitation to travel survey app and trace processing suite developers, a total of 17 apps were recruited for testing. A controlled experiment was devised, and the accuracy of the apps evaluated based on their ability to reproduce ground truth trip information. Further, the performance of the apps in terms of battery drain was also quantified and evaluated. Results indicate that while accuracy in terms of the trip ends/starts is reasonably high in most cases, mode inference accuracy varied significantly, with a maximum 65–75% accuracy achieved. As such, until significant improvements in mode inference algorithms arise, purely passive location-logging smartphone apps cannot serve as full-fledged automated travel survey instruments. While this may seem problematic, with minor input from respondents regarding regularly visited locations and modes used, as well as specific test case tuning and use of external data such as General Transit Feed Specification, there is an excellent potential to significantly reduce overall response burden and allow for high quality multi-day travel diary data to be collected. Implications of our findings for app design are discussed.

Keywords Data collection · Smartphone apps · Travel survey

Ahmadreza Faghih Imani
s.faghih-imani@imperial.ac.uk

Khandker Nurul Habib
khandker.nurulhabib@utoronto.ca

Extended author information available on the last page of the article

Introduction

Smartphones, with their location-logging capability, enable the collection of high-quality data on the travel patterns of individuals over multiple days by minimizing the burden on those being monitored. Unlike with other methods, individuals do not have to recall and report every detail of their travel. Smartphone apps can harness background logging functionality and provide a user interface that makes it possible for participants to merely validate pre-recorded and processed trip information, either in real-time or at the end of the day. Using smartphone apps to generate travel diaries potentially solves some of the issues with traditional self-reporting survey methods; these include unreliable respondent memory, proxy bias, mode bias, and rounding of travel times. Further, smartphone apps provide the ability to collect travel diaries for multiple days or weeks which could potentially lead to better understanding of individuals' travel behaviour especially for intra-person variability in travel behaviour, the activity space of respondents and variability in travel over the course of the week, the link between time constraints and scheduling of discretionary travel, and trip chaining (Nahmias-Biran et al. 2018).

The household, person, and trip information typically collected in household travel surveys today can be obtained via smartphone, if the questions are included as part of an embedded questionnaire within the app. With apps, prompting can be employed to collect information in real time as well, making it less likely that memory affects the quality of the data collected. Using apps, maps can be generated on the fly for a given user's travel to help them answer questions, offering visual cues and reminders as to where and when they may have traveled. Further, the app can then ask the respondent to augment the already collected spatial and temporal information with contextual information: mode(s) of transportation, purpose(s) of the journey, and other attributes of interest. With smartphone traces, actual travel paths, including what road network links or transit lines were traveled on, can be inferred. This opens new avenues for analysts, as collecting such data was previously excessively burdensome for survey respondents. These potential benefits have led to several smartphone apps being developed around the world specifically for data collection in travel surveys.

This paper undertakes a rigorous assessment of the strengths and weaknesses of smartphone apps for addressing regional and national travel diary data collection needs. The assessment aims to address five major problems in the literature: (1) no universal metrics exist, (2) results in the literature are region and travel-pattern specific, (3) self-assessments are the norm; (4) small-sample and imperfectly generated ground truth data are typically used, and (5) that battery drain is reported on without context. Smartphone apps have historically been assessed by their designers, with no independent assessment comparing apps and processing suites using properly recorded ground truth data for all travelled links, legs, and overall trip information.[1]

The assessment in this study seeks to better understand the accuracy of diaries generated from the data collected using smartphone apps. Thus, along with recording ground truth, data are collected using 17 smartphone apps (Android and iOS combined) from 11

[1] In this paper, trips refer to movements from an origin to a destination. If a trip includes more than one mode, legs are the individual mode segments of a trip that get from origin to mode change point, between mode change points if more than one mode is used, and between the final mode change point and the trip destination. Finally, links refer to the road network links traveled on during the course of a movement episode (this in contrast with a stationary episode). These are drawn as polylines between intersections.

different development teams around the world. The accuracy of the apps and trace processing suites are evaluated based on the capability to accurately reproduce ground truth trip information, including leg ends, trip ends, trip starts, trip modes, and route. After processing the collected data, the performance of the apps is further quantified based on observed battery drain. The assessment was not designed to come up with a numeric score for the apps and processing suites, but rather to better understand the overall state of the art, range of performances among leaders and best practices in the field. As Wu et al. (2016) state, "results comparison has little meaning because different studies use a different quality of data, which has a great impact on the results" (Wu et al. 2016), and for this reason and others, a '90%' or '50%' accuracy value reported in this article is not deemed more 'accurate' an assessment of an app's performance than any prior published work. Rather, it is the range of performances and any broader lessons learned that are the focus.

The paper continues with an overview of smartphone apps for travel data collection. Section 3 discusses the data collection protocol. Section 4 presents the results of the assessment, and finally, Sect. 5 concludes the paper.

Apps for travel data collection

A traditional household travel survey requires respondents to participate, whether via telephone, web, or paper questionnaire, by reporting trips for one or more days. This is not without problems, including under-reporting of short and discretionary trips, low response rates and inaccurately recorded travel information (Bohte and Maat 2009; Harding et al. 2018; Schlich and Axhausen 2003; Wolf et al. 2004).

To mitigate these concerns, researchers first started collecting location information using standalone GPS loggers alongside user-provided input data. Such surveys are carried out by asking respondents about their travel information over the phone or via paper-based or web-survey, and then augmenting this information with traces passively collected using GPS devices (Auld et al. 2009; Li and Shalaby 2008; Schönfelder et al. 2002). The 2010–2012 California Household Travel Survey is an example of GPS being used at large-scale in this way (Kunzmann and Daigler 2013). However, the use of standalone GPS loggers leads to increased costs for data collection efforts (Wolf et al. 2014). Researchers not only have to manage recruitment and pre-interviews with respondents but also bear the burden of device distribution and retrieval. These issues, along with the growing ubiquity of personally owned smartphones that include location sensors with similar capabilities as standalone GPS loggers, have led to a move towards the use of apps.

Over the last decade, capabilities and market penetration of smartphones have increased substantially. Smartphones bring together location logging with a user interface that also makes it possible to carry out experience sampling, on-route or after-the-fact travel validation, as well as provide feedback to respondents. Many of the recent typical regional travel surveys around the world are conducted in conjunction with a smartphone app (Allström et al. 2017; Flake et al. 2017; Geurs et al. 2015; Joseph et al. 2019; Lynch et al. 2019; Nahmias-Biran et al. 2018). There remain, however, numerous technical challenge in the use of smartphones for travel data collection when it comes to the frequency and type of data that can realistically be collected (Prelipcean and Yamamoto 2018).

Battery drain is a crucial limitation of location-logging smartphone apps. The collection of location data using GPS places a significant drain on a smartphone's battery (Jariyasunant et al. 2014). As a result, considerable efforts have been expended researching hybrid localization techniques using GSM and Wi-Fi location data, and other built-in sensors, to augment or replace GPS under certain conditions or to reduce the frequency of location queries (Ellison et al. 2019). Some apps only request location data when a movement episode is detected via algorithms that use data from lower power smartphone sensors such as accelerometer readings or Wi-Fi signature (Patterson and Fitzsimmons 2016). Geofences are another way to limit battery drain: a geofence can be set up when a stop is detected, with conditions then applied such that querying and recording of location points do not occur until a respondent is determined to have left the geofence. These approaches, while effective in reducing battery drain, might also lead to inaccurate and patchy data, especially for subway trips (Krenn et al. 2011).

The reliability of smartphone location logging accuracy in different urban environments must also be considered. Aside from measurement error (Enge and Misra 1999), the "urban canyon" effect (signals bouncing off of high rises in densely built-up areas) that results in recording errors for standalone GPS loggers remains a problem with smartphone apps. It is possible, however, to use cellular networks (GSM) and Wi-Fi-derived location information to estimate the location and improve accuracy in dense urban environments (Ellison et al. 2019; Greaves et al. 2015; Jariyasunant et al. 2014; Zhao et al. 2015). In cases of GPS signal loss, such as underground locations, smartphone sensors can be employed to estimate present location by keeping track of the general direction and distance traveled from the last known point, known as dead reckoning (Randell et al. 2003).

While the broad technical limitations and potential of current generation smartphone apps is well understood, there is a lack of understanding among transportation professionals and researchers regarding the feasibility of rolling out smartphone apps for passenger travel data collection. To the best of our knowledge, this study provides the first large-scale, global, and multiplatform assessment of the accuracy of travel diary-type outputs, combining app and processing suites. The experiment devised involved the collection of data using a range of apps, followed by comparison against a rigorously recorded ground truth. Similar assessment frameworks were developed previously to compare GPS data to time use surveys or travel diaries (Lawson et al. 2010; Stopher and Shen 2011). In our assessment, guidelines were used to determine what trips were to be made, how long these trips were to take, and by what modes they should be carried out, all the while carrying all 21 project phones together and recording events using a separate device. Such an experimental design ensured that a mix of trip types was recorded and that no uncertainty existed in comparing processed traces to ground truth data on travel episodes.

The methods employed for smartphone location logging and trace processing assessment have so far been ad-hoc, with recording and processing algorithms and outputs being project-specific, making it impossible to truly compare between apps or processing suites (Dalumpines 2014). In this study, all apps and processing suites are simultaneously compared to allow for a more objective evaluation of their potential to accurately record start and end times, and infer trip locations, modes of transportation and the links traveled.

The existing literature has established the features and functions, as well as the general quality of smartphone data. Before this study, however, no effort had yet been carried out to compare (1) multiple apps and processing algorithms, (2) in a particular region, (3) where data were collected following a strict protocol. Generating a perfect ground truth and carrying 21 phones together running 17 apps allowed us to quantify performances in a manner never done before, but it also meant we could not make use of the data to better assess the learning potential of apps, purpose identification, protocol compliance or any other notion of user interaction. Therefore, assessment results in this paper provide an objective evaluation of the expected performance of apps. The results obtained provide sufficient information for decision-makers to better understand the potential role of smartphone apps in regional travel data collection efforts.[2]

Data collection protocol

App selection

Personal invitations to participate were sent to app developers and research groups for whose products the team were able to find information, as well as broader invitations were made around the world through professional networks, e.g., IATBR listserv, TRB Committee mailing lists and personal networks. Responses were received from 19 companies and university research groups, with apps or processing suites being proposed for assessment. Accounting for both Android and iOS versions of the apps, this resulted in a total of 29 apps and data processing software suites. Of these, ten were commercial (consultant or other private firm-developed applications that require paid contracts to make use of) and 17 from academia (developed by research institutions and either not currently or exclusively commercial in use), along with two apps that were open source (with code available on GitHub) or free on the App/Play store with traces available for download.

After a pre-assessment to ensure the apps functioned properly and were worthy of further evaluation, eight app candidates were selected for Android and nine for iOS from 11 different development teams. Given knowledge of both the apps and processing suites available and those made by institutions who declined participation, a representative cross-section is believed to have been recruited that allows for reporting on the at-the-time state of the art. For privacy reasons, in this article each app is assigned a number, instead of being referred to by name.

Different phone models were employed. All phone models except one had all potentially relevant sensors, namely gyroscope, accelerometer and compass. The testbed was comprised of 11 iOS devices and 10 Android devices. To run the apps, active SIM cards with data plans were used.

Setup and data collection

Data collection was conducted starting in June 2016. Travel data were collected for 32 non-consecutive days. Additional days of battery drain-only recordings were also collected to see how the phones would perform when (1) only running the location logging apps without movement, and when (2) neither running the apps nor moving. This allowed us to isolate the effect the apps themselves have on battery drain.

All phones were configured the same way to ensure that the settings would lead to as uniform a drain as possible when testing each app, irrespective of device. No other apps were installed, while screen brightness was minimized and notifications were turned off. Non-location-logging app drain was kept to a minimum, such that whatever drain was

[2] While the success of a travel diary data collection project depends just as much on recruitment and design of user interface as it does on the more technical accurate recording and processing of traces, the design aspect of the work would have meant an entirely different project being undertaken. For more information on more human elements of app design dealing with user experience design, the readers are referred to one of the appendices in Harding (2019).

Table 1 Trips and hours of travel by validation status and main transportation mode

Main mode of transportation	Disaggregate	Trips recorded	Validated	Non-validated	Hours of travel
Walk		52	28	24	9.4
Bike		77	29	48	21.8
Local bus/streetcar	Bus	59	29	30	27.0
	Streetcar				
Subway/SRT		41	22	19	22.2
Regional transit (GO)		33	18	15	21.6
Coach		4	2	2	28.0
Car		51	33	18	15.9
Total		317	161	156	77.8

observed could be attributed to location sensor use, app-specific upload or download, or local trace processing. During collection, the location logging apps were brought up on screen every few hours to check whether data were logging correctly and to respond to prompts from certain apps when validating travel in real-time. These were the only times the screen on the devices were turned on. This approach was believed to lead to an assessment of apps with results relevant farther into the future, independent of what was deemed a realistic or ‘normal’ use of a smartphone.

Each device ran one app at a time, with apps cycled through devices over the collection period. A log was also kept of all trips made. To minimize the potential for random errors when reporting performance, all devices were carried together at all times, with the same trips recorded for all phones.

To produce a robust assessment, as many trips as possible were logged while maintaining realistic dwell times for stop detection. The trip-making was set to be as random as possible by varying length, dwell time, travel modes, and built environment (urban or suburban). Car, transit, bike, and walk trips were made. For transit trips, bus, subway, and regional trains were included. The trip scheduling design ensured all modes would include a minimum number of trips—approximately 30. The number of coach trips, however, were kept to a minimum because of scheduling logistics, as well as the small benefit from including such a seldom-used mode of transportation. Table 1 presents a summary of number of trips made by main mode of transportation. All local trips are made within the Greater Toronto Area (Fig. 1).

Since data were collected simultaneously using 21 smartphones, the total number of trips potentially recorded is much greater than the number of trips physically made. Ground truth information was generated reflecting the start and end locations and times, mode change points, and all associated modes of the travel episodes. Paths were also constructed from the OpenStreetMap (OSM) network dataset such that every link traveled represented the exact link taken for a particular trip leg.

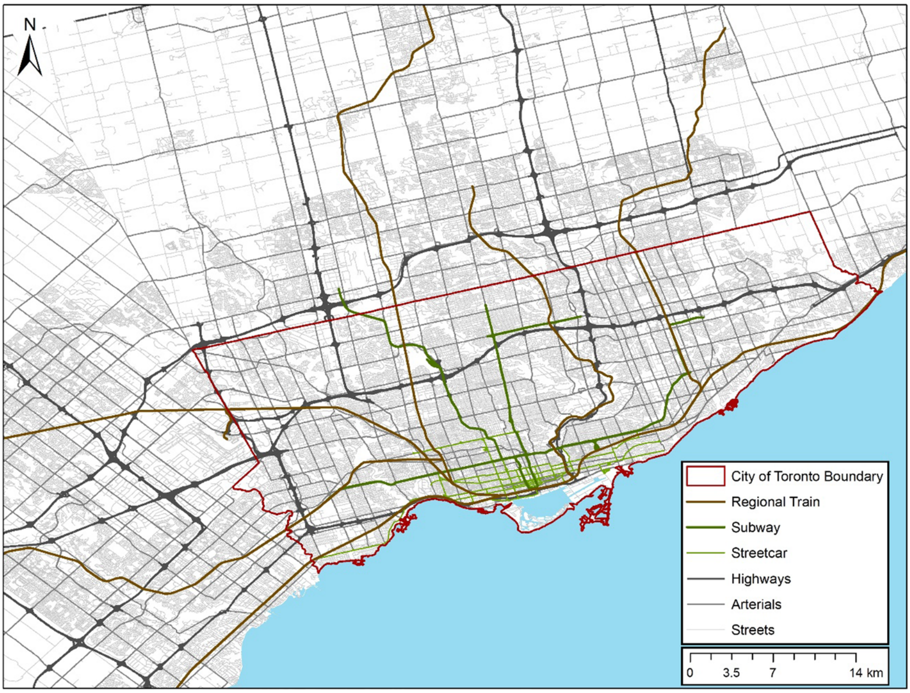

Fig. 1 Greater Toronto Area transportation network

Analysis and results

In total, 317 trips were made while carrying all 21 phones, consisting of 553 legs (including short access and egress legs). Over roughly 130 h of travel time, 3655 km were traveled, and 1063 battery recordings were taken. Travel recorded included trips made on foot, by bicycle, bus, streetcar, subway, commuter rail, car, and finally multi-modal trips.

A total of 11,304 events (i.e., recorded information about trip legs and stops from all 21 phones) were collected or extracted for analysis and comparison with ground truth (GT) information. Reporting of app performances with respect to reproducing ground truth are reported both for 'Native' app-processed data (i.e., X app producing trip tables using data recorded by the same app), as well as app data processed using a common Travel Diary eXtractor (TDx). TDx is a generic trace processing suite that was developed to convert any app's raw location traces to a universal format and then process the traces to remove noise, identify trip ends and infer main mode of transportation (Harding 2019). This allowed for all apps, irrespective of whether they are bundled with a trace processing suite or not, to be included in the assessment.

Certain apps merely produce traces, while others produce event or trip tables, some containing inferred and validated modes. Of the 17 apps tested, only two provide shapefiles that could be compared with link-matched ground truth. For the remainder, routed and point-to-line paths were generated using the TDx and ArcGIS. This allowed for the measurement of overlap between actual routes and routed smartphone traces.

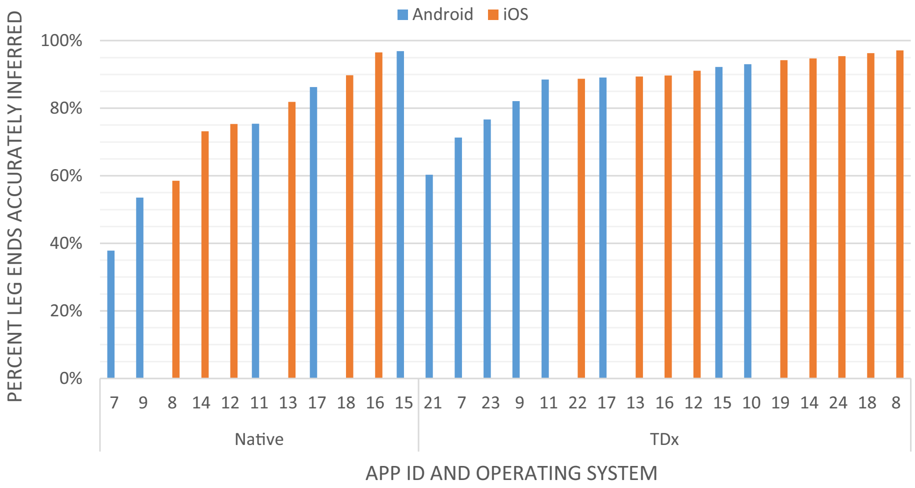

Fig. 2 Percent app-data inferred leg ends that match with GT

Assessment of travel episode detection and attribute inference

Any observations that potentially contained errors because of procedural mishaps were first removed from the final datasets. Such mishaps were identified when verifying app status during data collection, when uploading to server at the end of the data collection days or when manually reviewing the trace and battery reading files before running any scripts on them. These accounted for approximately 3.5% of all observations. This section provides detail from the assessment of identified trip leg ends, trip start, end and modes, and route overlap.

Leg ends

Figure 2 shows the percent of leg ends inferred from app traces that have a match in time (± 7 min) and space (800 m) when compared to the ground truth leg end data.[3]

As seen in Fig. 2, the performance of apps and their processing suites varied considerably when it came to producing accurate travel episodes from traces. The difference between the performances of native processing suites (native meaning that a given app's traces were processed by an algorithm or suite designed by the same team that designed that particular app) was considerably higher than that of multiple app traces being run through the same TDx. Leg end accuracy ranged from 38 to 97% on the native side, and 60 to 97% on the TDx side. Figure 2 also shows that it is rather straightforward to correctly identify leg ends from traces. All that is required for this task is a clear understanding of dwell times that are

[3] The cut-off values in this paper were decided upon after trial and error testing in a previous project by the same team, taking Toronto subway spacing, network attributes such as block size and minimum dwell times into consideration (Miller et al. 2016). The data collection protocol was compatible with the values previously determined using trial and error, namely staying in a given location for 3 min or more before beginning another travel episode.

Table 2 Legs included in analysis, by transportation mode and OS

Mode (for leg)	Android	iOS	Total
Walk	848	946	1794
Bike	733	897	1630
Car	344	461	805
Bus	183	231	414
Streetcar	83	91	174
Subway	143	186	329
Regional rail	157	193	350
All modes	2491	3005	5496

Table 3 Trips in analysis data, by main transportation mode and OS

Main mode (for trip)	Android	iOS	Total
Walk	227	255	482
Bike	612	737	1349
Car	279	374	653
Bus	185	202	387
Tram	113	135	248
Subway	166	202	368
Regional rail	182	225	407
All modes	1764	2130	3894

locally relevant (not too short to accidentally include traffic signal delays or boarding and alighting times at transit stops) and reasonably complete and accurate location traces.

Ignoring the validation data recorded, the data that make up the analysis dataset for legs includes 5496 observations, shown by OS and ground truth mode of transportation in Table 2.

Looking only at the best performing algorithm for any app, 13 of 17 apps produced high enough quality data that 88% or more of all inferred leg ends matched the GT leg ends. The four worst performers were all Android apps; this was the case with many of the accuracies looked at, while for battery drain the reverse was found. Three out of the four had larger than average recording intervals (more infrequent location point recording), at 23, 60, and 108 s post-filtering, the likely reason for the poorer end detection accuracy.

Of all the processed datasets, those from 2 apps that had native processing suites outperformed TDx, Apps 15 and 16. It is interesting to see that, at least in this respect, it seems possible to build a flexible trace processing algorithm that performs reasonably well for detection of leg ends when applied to datasets produced by multiple apps.

Trip ends

Instead of simply labelling trips as 'correct' or 'incorrect', events were placed into six different categories, based on whether the time, the location or both were correct, as well as whether a travel episode was underway at the moment a leg or trip was detected. This

Table 4 Percent of app-inferred trip ends by status when compared to GT trip ends—significant errors highlighted

OS	ID	Native status end code						TDx status end code						Trace description			
		A (%)	B (%)	C (%)	D (%)	E (%)	F (%)	A (%)	B (%)	C (%)	D (%)	E (%)	F (%)	Constant logging	Uniform rec. freq.	Median rec. int. (s)	Acc (m) post-filter
And	9	63	1	1	21	11	1	82	0	1	3	3	12	Y	Y	1	7
And	11	86	0	2	4	0	8	89	0	3	1	2	6	Y	Y	1	8
And	21							67	2	12	7	9	4	Y	Y	60	145
And	17	90	1	3	3	4	0	93	0	2	2	1	2	N	N	5	22
And	10							94	0	1	1	1	5	N	N	7	9
And	23							81	2	5	3	4	5	N	N	20	116
And	15	98	0	2	1	0	0	95	0	2	1	1	1	N	N	61	13
And	7	38	16	7	21	16	2	74	1	3	16	4	1	N	N	108	*
iOS	12	87	0	3	4	1	6	94	0	2	0	1	4	Y	Y	1	11
iOS	22							89	1	5	3	0	2	Y	Y	2	292
iOS	13	82	0	0	16	1	0	88	0	5	6	0	1	N	Y	1	8
iOS	14	89	0	5	5	1	0	95	1	0	2	1	1	N	Y	1	8
iOS	19							96	0	2	1	0	2	N	Y	5	*
iOS	18	95	1	1	2	0	1	98	0	1	2	0	0	N	N	4	14
iOS	24							98	0	0	0	0	2	N	N	5	8
iOS	16	97	0	2	1	0	0	93	1	2	2	0	2	N	N	17	14
iOS	8	57	12	1	26	3	0	99	1	0	1	0	0	N	N	23	*

Code A: trip ends matched in time and space, where an app-inferred trip end is both within 800 m and 7 min of an actual trip end

Code B: a trip end detected to between 800 m and 1.5 kilometres of the actual end of a trip, as well as within 5 min of the true trip end

Code C: a trip end detected within 800 m of the actual trip end, but where the time is greater than 7 min (trip end detected late)

Code D: a trip end detected while the user is in motion, within 800 m of a valid trip end, but not with a trip end greater than 7 min from a true trip end (truncated trip)

Code E: noise, where a trip was detected outside a travel episode

Code F: any trip previously identified as a type A, but where it was found that more than one app-inferred trip end matched a particular GT trip end—a particular form of noise, where a correction was applied to ensure that a given true trip end could not be used more than once when matching app-generated trip ends

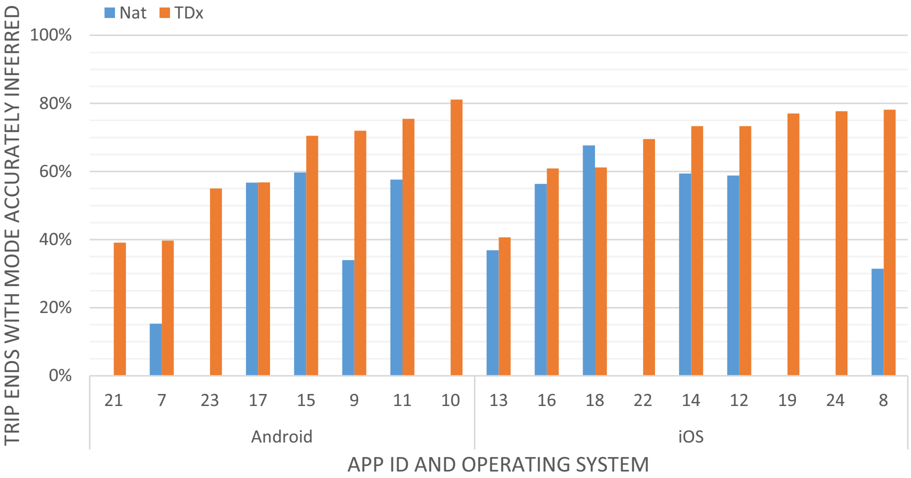

Fig. 3 Percent correctly identified trip ends

allowed for differentiation of legs or trips accidentally detected on-route to events, truncated events, and finally, those detected when the phones were not in movement—noise.

Ignoring validation data recorded, the data that make up the analysis dataset for trips includes 3894 observations, shown by OS and ground truth mode of transportation in Table 3.

Error categorization is shown for trips in Table 4. To help better understand the errors present, certain attributes of the apps are included on the right side of the table. These are 'Constant logging' (whether an app logs points continuously, irrespective of movement), 'Uniform recording frequency' (whether traces are recorded with the same frequency irrespective of travel speed), 'Median recording interval' (in seconds) and 'Post-filtering accuracy' (in metres).

The apps most likely to generate 'duplicate' trip ends (code F) are those which log continuously, with high frequency and also with high reported accuracy. TDx was initially developed with the explicit goal of being able to identify short, active transportation trips, and as such, it may be that small clusters of high reported-accuracy points at times are labeled as short 'trips' when in reality those short series of points are actually a continuation of prior or subsequent movement episodes. Native processing for certain apps that share these same attributes (Apps 11 and 12) appear to share this propensity.

Concerning trips inferred while not in movement (noise, or code E), there is a more even distribution among the apps. Most of the apps with higher than average noise trip detection rates record coordinates more infrequently (median recording frequencies higher than 20 s). Only App 9 recorded at high frequency and is also associated with increased noise trips. Truncated trips (code D) occurred in Apps 7 and 21 (low recording frequency and post-filtering accuracy), and App 22 (low spatial accuracy). App 13 makes use of an aggressive battery saving technique that resulted in both truncation (code D) and delays in detecting a significant movement in space (code C).

Figure 3 shows the overall percentage of the correctly identified trip ends; apps without a native processing suite have no blue bar. This metric is calculated as follows:

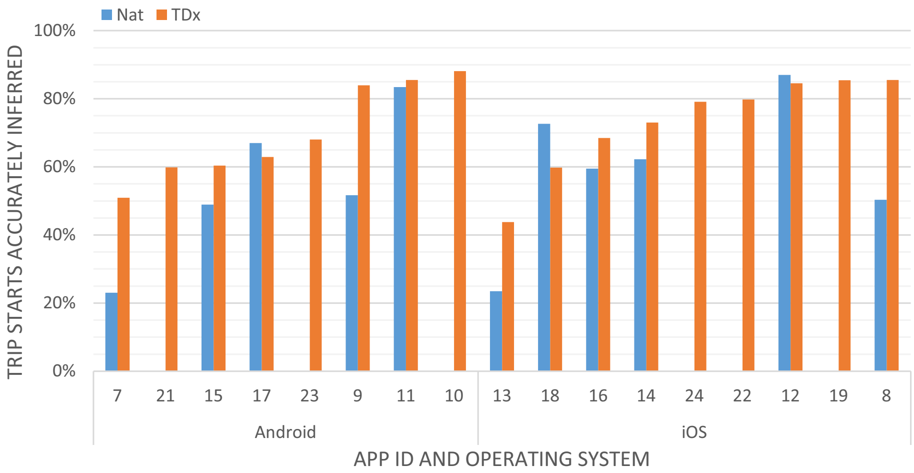

Fig. 4 Overall trip starts correctly identified

$$\text{Percent Trip Ends Correct} = \frac{\text{Trip Ends Detected Correctly}}{\text{Total Trip Ends in GT} + \text{Incorrectly Identified Trip Ends}}$$

As can be seen, the trip-end identification performance of apps varied much more than the measures mentioned to-date, with performances as low as 15% and as high as 81%. This would indicate that certain apps and processing suite combinations lead to considerable errors in terms of missed trips and incorrectly inferred trips; these may be related to truncated trips or noise points. On the other hand, there were both Android and iOS apps that produced results near or above 70%.

Two of the three apps with the lowest reported accuracies were specifically designed for very low battery consumption. To save on battery drain, these apps have median recording intervals over 60 s with average accuracy (post-filter) of 145-m. They were designed to run without the user being burdened and, consequently, are better suited to providing information on trips made over long distances, where activity durations are significant and where the path and mode inference are not needed. These apps demonstrate the trade-offs inherent in focusing on high frequency and high accuracy logging.

While the result may not be generalizable to the processing of real-world movement episode traces, TDx is shown to provide equal or greater performance to native processing approaches in all but one case within our test. This would appear to indicate that for detecting trip ends, there is no considerable difference between a generic processing suite and that which is developed for a particular app. Since a diverse set of modes were selected in making trips in this assessment, app-specific processing suites not configured to reliably capture short and active transportation trips were at a disadvantage. This also demonstrates tailoring and calibrating algorithms to a specific desired output and locally relevant information can improve the accuracy of collected data. Concerning the operating system, both iOS and Android contained high performers, but iOS apps produced data that more consistently led to correct detection of trip ends.

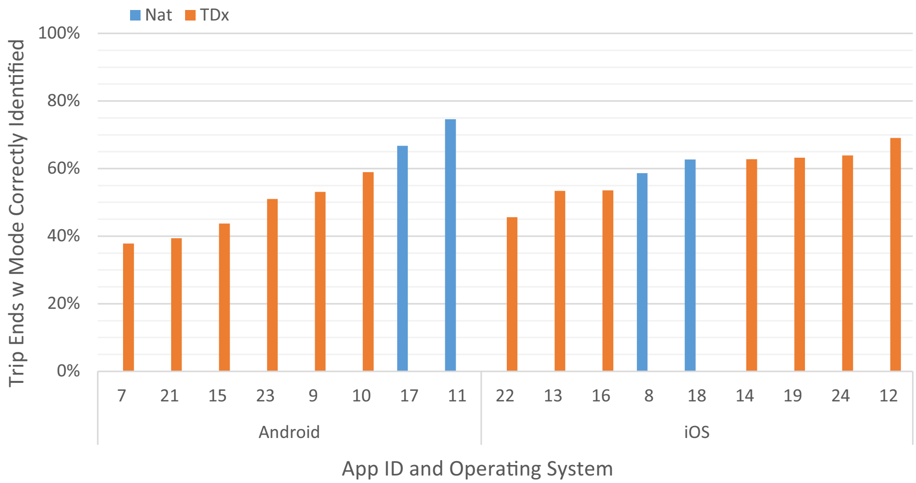

Fig. 5 Overall mode inference accuracy based on mode at trip end

Trip start

The trip start is another important dimension to capture. Figure 4 demonstrates how trip starts can often be missed before corrections are applied. This is a result of the cold-start issue. Cold-starts refer to situations where an app may take a long time to acquire location information, leading to a difference in the detected and actual location of the beginning of a trip. Large gaps in time and space lead to an inability to pinpoint when a respondent left a given location and when they arrived at another. If a respondent lives, works, or shops at a location connected to the subway, this can be especially problematic. It is possible to miss the departure and find a user far away from their previously detected location hours later; in such a situation, it would be impossible to differentiate between a missed trip start and several missed trips.

The best way to ensure that significant gaps do not occur is to periodically record locations, even at lower spatial accuracies, such that when a significant movement does occur, some information can be made available on where and when location information was last available. This is a very important aspect that is not properly handled by most apps. Apps that record data continually (9, 11, and 12) perform better in this respect.

Trip modes

In this assessment, trips were made using arbitrarily selected modes of transportation, such that there would be no correlation between particular OD pairs, time of day or broad location within the study region and mode of transportation used. Results from our comparison of trace processing algorithms (shown in Fig. 5) indicate that it is possible to infer mode with a maximum 65–75% certainty (Apps 17, 11 and 12), where all modes (Walk, Bike, Transit, and Car) are evenly weighted. Equally weighting each mode allows for a generalized measure of performance, independent of the mode split of a region.

Mode inference in TDx is carried out using a point system, where attributes of trips and the points that make them up are used to infer whether the main mode of transportation used for a given travel episode is walk, bike, car, surface transit (bus/tram), subway or regional rail. Information on the road and off-road network (one-way or both, as well as road type—highway, access ramp or other), transit line shapefiles, ferry route and railway corridor, gaps created near subway or train station entrances, average, 85th percentile and maximum speeds, share of trip (time and distance) in different speed bounds, and distance covered are all used in various ways to assign points and determine mode of transportation (for a longer description of TDx' mode inference approach, read appendix B.4 of Harding 2019).

Native mode inference for other apps varied from team to team. To the best of our knowledge, no participant in the assessment modified their app to account for real-time transit information from Toronto's regional transit agency. Any app that makes use of globally available routing, such as Open Streets Map, would have been able to work, but we were not made aware of any tailoring to apps or background processes.

In assessing the accuracy of mode inferences, we only looked at main mode of transportation identified, not access and egress modes. TDx, for all intents and purposes, only assigns a mode of transportation to trips as a whole, and for multimodal trip cases makes use of the following hierarchy: walk, bike, car, transit (broad category that defaults to surface transit), subway, regional rail. The same hierarchy was applied to modes inferred by Native processing, such that a trip with a regional rail component would be labeled as having regional rail as main mode no matter what other modes its legs were detected as, and so on down the hierarchy.[4] Use of the same hierarchy for comparing TDx and Native mode inference to ground truth to which only a main mode was assigned ensured consistency in results.

It should be stated that apps with intelligent pattern-matching algorithms that depend on learning from 'respondents' would have been unable to do so in our experiment. Therefore, accuracies estimated are lower than what those apps might achieve a few days into a data collection effort where respondents report their actual mode of transportation for previous trips.

Further, the types of public transit used within the Toronto area may also exhibit very different patterns in terms of speed, acceleration, and vibration compared with other cities where apps may have been calibrated. Accuracy can be increased by fine tuning the apps to a specific city transport network and use of General Transit Feed Specification (GTFS). Finally, no information on mobility tool ownership, preferences, or patterns could be utilized to improve predictions. Taking these into account, the percentages correctly identified should be interpreted as very conservative estimates of the mode inference accuracies to be found during a full data collection effort.

Another way to present this information is to show inference accuracy by the *main* mode. Figure 6 shows the percent of modes inferred correctly for each app by the main mode of travel for the recorded trip. These figures provide insight as to where certain apps' algorithms may perform particularly well, or less so for the main trip mode. The significant performance differences per app for each mode of transportation indicates that predictions

[4] People 'access' modes higher up in the hierarchy using modes below. While there can be cases where a trip's access mode might be longer than the 'main' mode assigned (ex: if a suburbanite drives 30 km to a subway station, parks there and then takes the subway into town the last few kilometers), in more cases we would be dealing with walk or bike access to transit, as well as drive access to commuter rail.

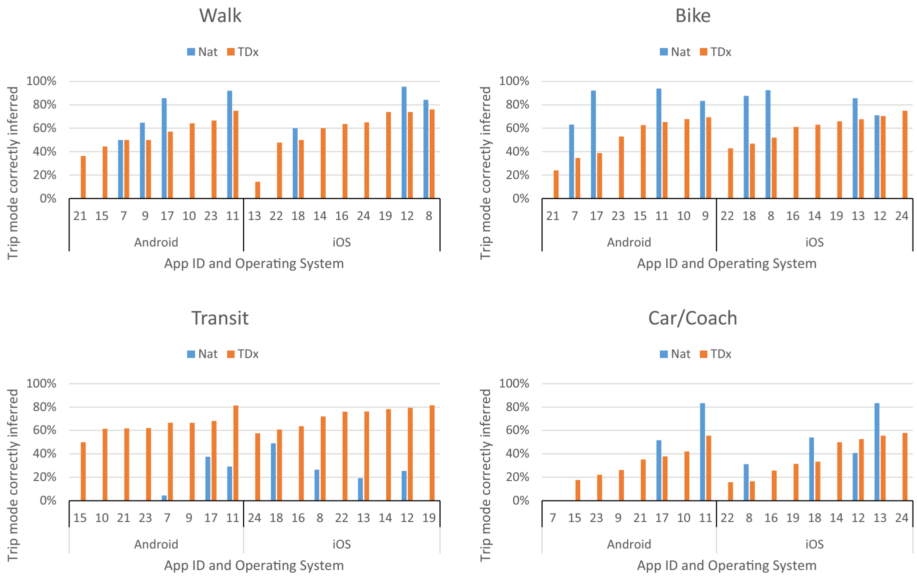

Fig. 6 Mode inference accuracy based on mode at ground truth trip end

made by mode inference algorithms are greatly impacted by design, expected use cases, and geographical context. For example, TDx was originally designed for a very urban area with little data on car trips. As such, it performed most poorly when inferring car trips. Also, some of the native app processing suites aimed at better capturing active travel did not differentiate between motorized surface modes (car, bus and streetcar), significantly reducing the percent ‘correctly’ identified. Transit inference was very poor for apps’ native processing suites more broadly. While part of the blame for the poor performance can be attributed to differences in transit travel patterns between Toronto and the cities where the apps had previously been rolled out, the result remains underwhelming.

Finally, at least one of the apps (13) is designed to ignore walk trips entirely, not recording location points unless a given speed threshold is passed; it is not surprising that it did not perform well when it came to detecting walk trips. This recording quirk makes it useless for recording walk trips, but also makes it extremely battery efficient, presenting a clear trade-off.

There exist many different methods by which the mode inference can be tackled, from rule-based methods to Bayesian Models, Fuzzy Logic, machine learning with Neural Networks or Support Vector Machine, etc., some of which exhibit very high reported accuracies (Stenneth et al. 2011; Xiao et al. 2015) (for a review of methodologies, see Wu et al. 2016). The work of Stenneth et al. (2011) will be addressed directly, as an example that demonstrates why we are reticent to detailing the exact mechanics of any one app or processing suite, and instead opt to report on the <u>spread</u> of performances obtained. In their paper, the authors detail a method of matching app-recorded trips with real-time transit feed data to obtain high accuracy transit mode inference. Their method is not only reliant on the availability of high-quality real-time transit location data from all operators in a particular geography of interest (more and more common, but not simple or straightforward to clean and integrate), but also on high frequency and high accuracy location data generated on the app side.

With these two components, Stenneth et al.'s method allows for travel episodes to be labeled correctly, but is dependent on significant local tuning, as well as transfer points being correctly identified such that legs can be matched with transit data. While these are attributes that would be very desirable in any given data collection location, these can not be guaranteed moving forward operating systems are increasingly federating requests for location information (Marra et al. 2019), nor are real-time transit feed available everywhere. Another example of high accuracy mode inference is the work done by Xiao et al. (2015). The app developed for their analysis recorded GPS traces for every second and thus resulted in heavy battery drain and thus an additional battery package was provided to participants which is not applicable for large-scale surveys.

The mode inference methods described in literature, can all be put to use in increasing accuracies for mode inference, but the existence of an approach does not mean that it can be realized in any and every context, or easily integrated to any app and processing suite—were it so simple and straightforward, the apps tested by our team in 2016 would have performed significantly better at identifying surface transit modes. And to drive home the point that any specific format of location trace recording should not be relied upon when building a processing suite.

While a mode inference algorithm that works perfectly in all types of setting would be ideal, results indicate this is unrealistic. As a result, any app used for collecting data from the general public should feature some form of validation, whether in real-time or in the form of a travel diary to be verified at the end of the day.

Route overlap

Much like the issue with 'correct' ends or modes being quantifiable in multiple ways, route accuracy can also be measured in different ways. The difference between app-inferred links and ground truth links needs to be quantified in a way that does not disregard either truncation or incorrectly identified links. Thus, a hybrid quantification approach was chosen, with non-matched trips ignored. As such, if an app only recorded 50% of all trips, but that 50% of the whole was perfectly matched to the GT shapefile, then the result produced is 100%, not 50%. This distinguishes the link-matching from overall recording accuracy issues.

A buffer of 20 m was used when comparing GT legs to app-inferred legs assigned to the road network, while a 50-metre buffer was used when comparing GT subway and regional rail to routed, point-to-line and app-provided shapefile polylines. The values of 20 and 50 m were decided upon based on prior work in the same region by the authors (Miller et al. 2016), as well as tests run in GIS using recorded traces before and after filtering to remove noise, as well as verifying the accuracy when assigning points to the road network using the link matching algorithm built into TDx. There is no hard and fast rule about what distance is appropriate universally.

More specifically, if points are associated to the right road network link, 20 m works even in cases where roads with multiple lanes are coded as 2 parallel links, one for travel in each direction. For regional rail and highway travel, 50 m is sufficient as trains follow a track, whereas travel along highways wider than roads also justifies an increase in buffer distance. This is very region specific however, and we make no claims that these values are to be used in a context other than Toronto.

Routed polylines refer to app data assigned to the road network using a link-matching algorithm within TDx and then connected using ArcGIS with the shortest path assignment to fill in gaps within a given trip. Point-to-line polylines are produced by sequentially

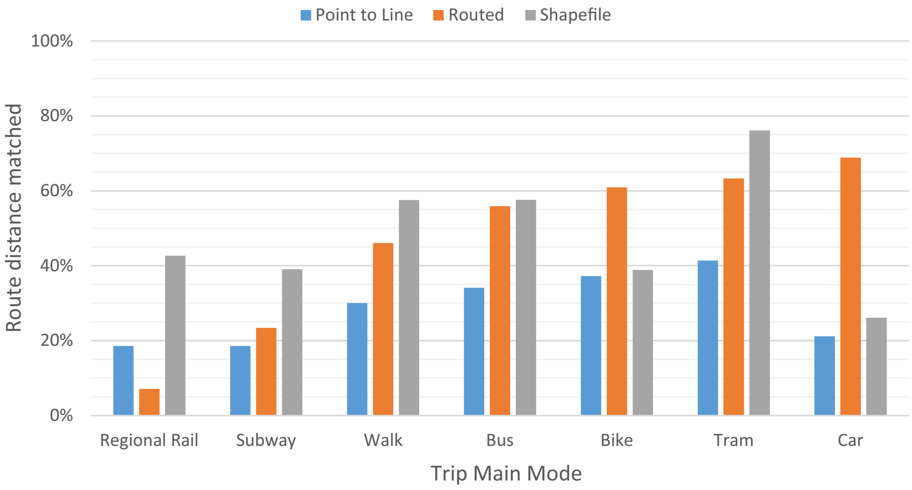

Fig. 7 Percent overlap in route identification, all apps, and operating systems combined

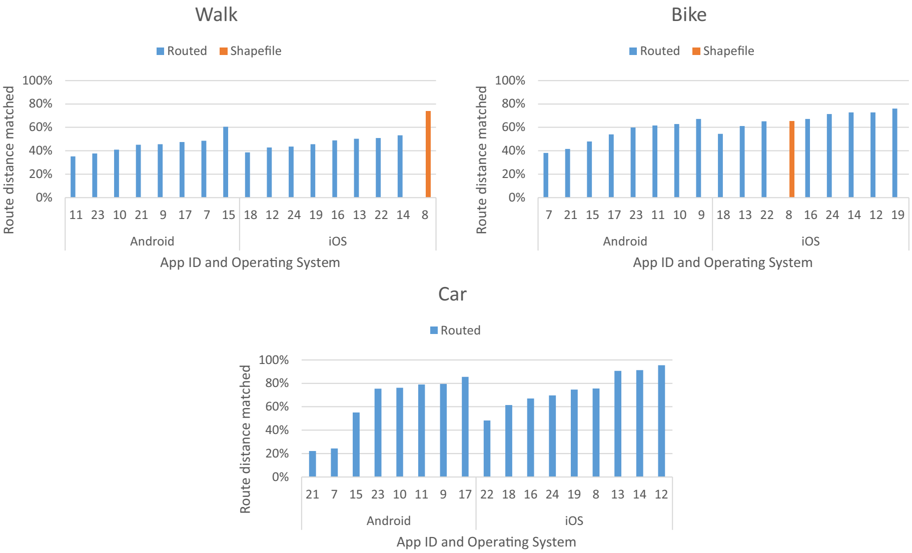

Fig. 8 Percent overlap walk, bike and car modes, best performance shown for each app

linking filtered points (as per TDx). Finally, the app-provided "Shapefile" bars present results comparing the GT shapefile with the shapefile for the two apps which made such an output available. Each approach has its strengths and drawbacks. The routed link matching can be used to identify exact links for input into path-choice models while permitting gaps in recording and sparser data. In contrast, the point-to-line approach is very simple and straightforward, but exact links are not identified and paths may be drawn along areas where no streets are present. The filtering applied in TDx removes much of the scatter that

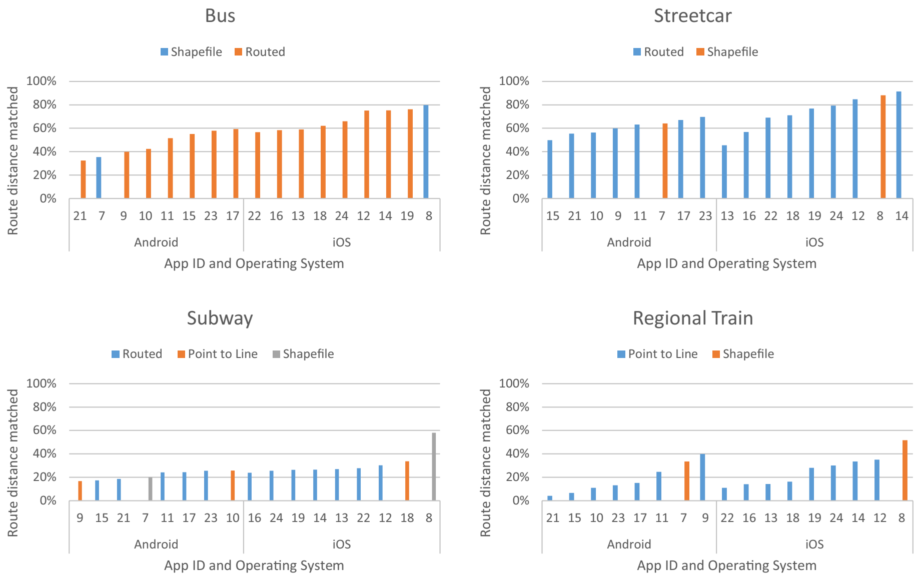

Fig. 9 Percent overlap transit modes, best performance shown for each app

would generate jagged lines and leads to a smoother path being drawn with less confusion as to overall distance traveled.

The results obtained for route overlap vary significantly from app to app. Substantially higher matches, relative to TDx, were obtained for apps where the developers provided a shapefile for the travel recorded. Figure 7 shows the percent of route length correctly identified by mode as opposed to one holistic number. In it, no attempt is made to present only the best performance. Instead, all apps' outputs are averaged.

To better understand the performance of the apps, Figs. 8 and 9 present the percent accuracy of each app for different modes. The accuracy in path inference was calculated based on the percent of the trip distance accurately inferred from the total length of the trip. As such, it is expected that walk trips, for which cold-starts account for a more significant proportion of overall travel distance, have lower reported accuracies. Further, travel along non-represented links such as through parks and alleyways, along uncharted footpaths or any other non-network locations are of more substantial concern for walk trips. In the case of car trips, it was promising to see such a high share of overall distance properly link-matched: around 80% of all travel distance was accurately link-matched across the board. This may have been a result of the low percent of access and egress mode as a share of travel; however, for most applications, this bodes well. Subway trips are not expected to be adequately matched by any method given that the trips take place underground. Looking for route overlap for the subway is not a particularly informative task; instead, what is essential is ensuring high accuracy in assigning subway mode to subway trips.

Battery drain

One of the greatest challenges in designing an app that produces decent quality data is to not excessively deplete a phone's battery when running. The battery drain imposed will

Table 5 Battery drain per hour, linear model with mAh drain as dependent variable

Variable	Coef.	SE	t-stat	Variable	Coef.	SE	t-stat
App iOS				*App-TravelTime iOS*			
App8	41.13	7	5.88	travMinApp8	0.59	0.34	1.76
App12	59.58	8.01	7.44	travMinApp12	1.84	0.35	5.27
App13	21.45	6.59	3.25	travMinApp13	0.7	0.38	1.84
App14	72.44	7.12	10.18	travMinApp14	0.15	0.38	0.39
App16	48.98	6.42	7.63	travMinApp16	1.49	0.36	4.15
App18	35.96	6.37	5.65	travMinApp18	0.96	0.37	2.61
App19	78.45	7.46	10.52	travMinApp19	0.13	0.37	0.35
App22	22.67	7.59	2.99	travMinApp22	0.47	0.35	1.34
App24	42.89	6.75	6.35	travMinApp24	1.08	0.35	3.11
App android				*App-travel time android*			
App7	8.26	5.77	1.43	travMinApp7	0.19	0.37	0.51
App9	24.37	6.58	3.71	travMinApp9	0.85	0.34	2.46
App10	20.45	7.51	2.72	travMinApp10	1.01	0.37	2.73
App11	9.47	6.87	1.38	travMinApp11	1.5	0.34	4.39
App15	13.01	6.45	2.02	travMinApp15	0.01	0.37	0.03
App17	20.86	5.92	3.52	travMinApp17	0.32	0.37	0.87
App21	13.68	6.67	2.05	travMinApp21		Omitted	
App23	7.23	7.29	0.99	travMinApp23	0.73	0.35	2.11
Device				*OS*			
iPhone 6	7.51	3.24	2.32	iOS	− 13.4	5.8	− 2.31
iPhone 5		Omitted		Android		Omitted	
Nexus	17.04	4.22	4.04	Travel time (min)	0.22	0.26	0.86
Life One	− 11.46	4.4	− 2.6	60 Hz Freq Error	22.71	14.77	1.54
Dash	5.3	3.21	1.65	Constant	16.74	4.08	4.11
Studio G LTE		Omitted		*Number of observations*	1040		
Bad Handset	64.86	4.26	15.24	R^2	0.61	*Adjusted* R^2	0.60

Coefficients highlighted in green indicate lower drain associated with a particular app, while coefficients in red indicate high drain relative to the other apps in the assessment

play a significant role in whether users run the app for long periods, but what level of battery drain might be 'acceptable' to respondents is not clear. The acceptable drain is significantly affected by the day's technology, as well as what respondents expect of their phone in terms of battery life. In our analysis, the battery drain expected from running the current generation of apps is examined.

While it is true that smartphone design—form factor (size, shape, and style), battery capacity and associated performance—is a rapidly evolving space, we would like to preface this section by stating that it would be incorrect to see the rapidly evolving situation with devices as an indication that the battery drain issue is solved only because time has elapsed since the data were collected. This is primarily because GPS remains a battery-intensive source of location information, while at the same time being the most reliable means by which frequently updatable and accurate location information can be obtained in any context (as long as there is a relatively clear line of sight to the sky). Wi-Fi and cell tower triangulation can provide some information on location, but do not

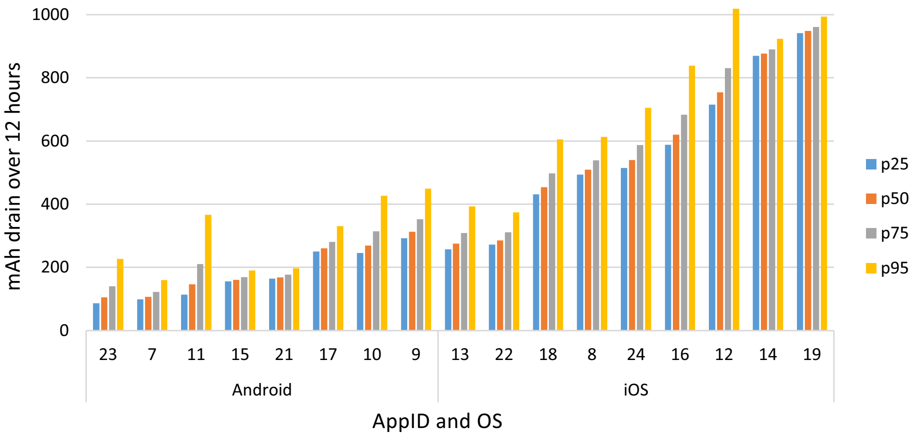

Fig. 10 Drain in mAh per app, using TTS 2011 respondents' travel as input

allow for precise information to be recorded about the movement occurring from one point to another (required for mode inference), nor do they provide reasonable accuracy away from dense urban centres.

To perform this analysis, battery recordings were taken on all phones throughout the data collection process. Of the 1063 battery drain recordings taken, 750 included episodes with travel. The remainder were recordings with an app running but no movement (218), or episodes with no apps running (95). Since the phones have different size batteries, the apps were assessed by the battery consumption in mAh (milliamp hours) instead of percentage drain.

Given the many parameters involved in battery consumption, in place of simple summary statistics, a linear regression model was constructed. Attributes considered include both general attributes, such as operating system, device type and travel time, as well as app and travel time-specific attributes. Key app attributes, such as continuous logging, frequency, and accuracy, are explored in Table 5. One of the variables included as a control was 'Bad Handset', as one of the refurbished iPhones exhibited poor battery performance irrespective of the app installed. Binary variables were tested for other devices, but this was the only one that proved statistically significant and to have a substantial effect.

Certain apps experience a much higher battery drain per hour than others. This drain is further broken down into that from simply running an app for an hour (having it active in the background—'App##' in the table) from that which is associated with the number of minutes spent traveling in a given hour while having the app turned on ("travMinApp##").

There was considerable variation in collected and modelled values for the apps' battery drain. Other than device-specific coefficient estimates, app-specific travel time coefficients proved to be significant in many cases, with a statistically non-significant generic travel time coefficient found once the app-specific ones were included. "60 Hz Freq Error" was also included as a control variable, as one of the apps was accidentally run at 60 Hz instead of 12 Hz on three separate days. The traces were modified pre-TDx by removing 4 of 5 points, while the effect on battery drain was captured by labeling the drain observations as being recorded on days where this error occurred, with the binary variable for this error included in the model.

To have a clear understanding of the impact on battery drain of the actual travel patterns of individuals in a given environment, reported travel data in the Toronto region were extracted from the regional household travel survey (TTS 2011) (Data Management Group 2016). Then, the mAh drain over 12 h in the region is calculated for each app and shown in Fig. 10 (with the 25th percentile, 50th, 75th, and 95th shown). It should be noted that individuals who did not report travel on a given day were counted as 0 km and 0 min-traveling individuals, not removed from calculations. This pulled the average drain down but provided a more representative depiction of the actual travel patterns in the region.

As can be seen in Fig. 10, there are a few Android and iOS apps with very similar drains associated—17, 10 and 9 for Android, and 13 and 22 for iOS. This being the case, the drain associated with the most battery-intensive location logging apps running iOS is still considerably higher than for Android apps.

For a variety of reasons, we believe that it is important to build an app for the 75th or even 95th percentile. First, because the actual travel times of individuals are likely to be somewhat longer than regional household travel survey-reported travel times as a result of trip under-reporting. Second, because periods of ‘motion’ where the device is not perfectly immobile, but also not legitimately engaging in a travel episode may lead to increased drain. Third, because travel time varies both between respondents, but also between days for a given individual. This last point is important, as any attempt to solve the battery drain issue by adding a pause button or accepting that respondents will at times manually shut down the location logging function means that there will not only be a bias in who runs the app for extended periods (fewer travelers with significant commutes experiencing more significant drain burden), but also what complete days of data are reported for the individuals who do report travel using the app (fewer days with large cumulative amounts of travel).

Looking, for example, to App 11 on Android, or 12 on iOS in Fig. 10, a significant difference is seen between the drain in terms of mAh for a long travel day from that of a short travel day over 12 h. The data used (the 2011 TTS, movement episodes only) have a median in-movement travel time of 19 min, a 75th percentile time of 56 min and a 95th percentile total travel time of 2 h and 26 min. If for those reported in-movement travel times additional time is added for unreported errands, dropping off children, going for a walk or even simply being away from known Wi-Fi hotspots, it is not difficult to imagine 3 h or more of movement recording time being required at least once a week. It is necessary to design apps for this potential drain.

One way to categorize apps broadly in terms of their accuracy and drain is to represent these attributes on a scatterplot, with a drain on the x-axis and accuracy on the y-axis.

Using the hypothetical trade-off matrix in Fig. 11, apps can be situated in one of the four quadrants. As can be seen in the Android and iOS portions of the figure, some apps with a lower drain outperform others.

- Apps (and processing suite combination) in quadrant 1 are the best, in that they lead to high rates of correct trip attribute inference with a low drain;
- Apps in quadrant 2 have high accuracies in terms of the trace and travel episode data they produce but lead to important battery drain. This indicates a large burden to respondents and potentially reduced number of days respondents would be willing to participate. With the advances in battery technology, however, these apps might become more appealing;
- Apps in quadrant three subject respondents to a low level of drain, but also collect data of lower quality—such apps could be used to identify longer distance trips, but would not be suited for the analysis of urban movement or mode identification;

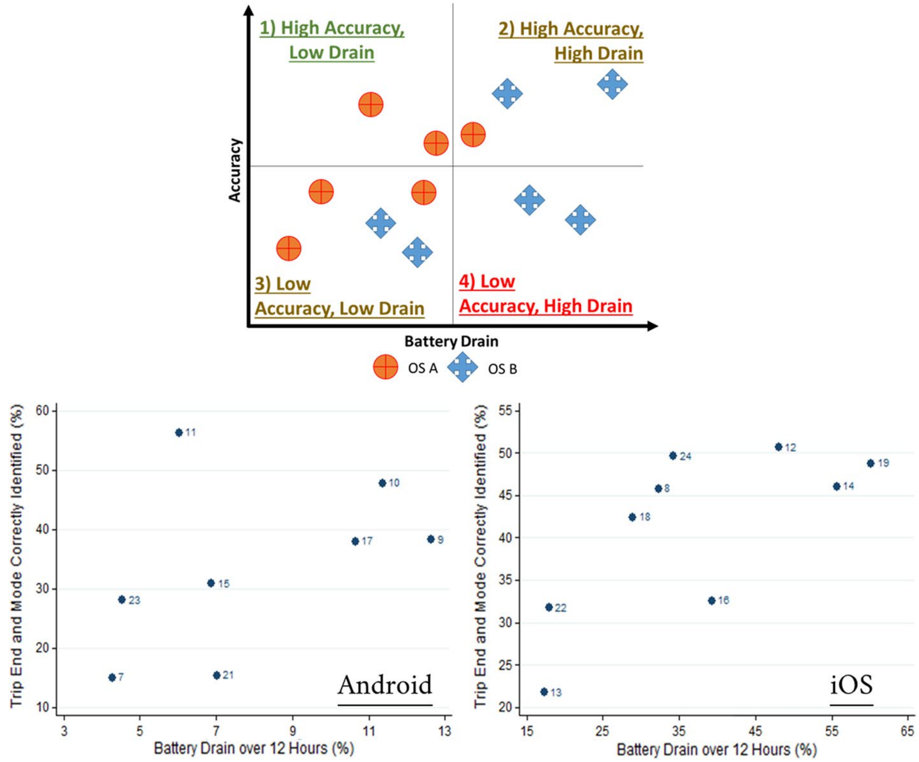

Fig. 11 Hypothetical trade-off matrix (above) and actual drain/accuracy matrices (below)—battery drain as a percent of device battery (x-axis) and its relation to combined correct % trip end and mode inference (y-axis), Android on left and iOS on right

- Apps in quadrant four should never be rolled out, as they provide lower quality data with significant battery drain.

Conclusion

The paper investigates the potential and limitations of smartphone apps as travel diary data collection instruments. The traces were collected using 17 different apps (Android and iOS combined) installed on 21 phones carried together at all times. At the same time, the ground truth information was recorded.

Results show that accuracy in terms of the trip ends detected is high in most cases, but performances are more highly varied for mode inference. This indicates that until significant improvements can be made to mode inference algorithms, the collection of user-submitted validation data is required. This can take the form of real-time prompts when trips are detected or travel diaries to be completed at the end of the day. This is especially true in a setting where the mix of travel modes is diverse, as well as where multi-modal trips account for a significant share of overall travel.

While it may seem problematic to hear that purely passive apps are not ready for rollout, what we would instead conclude from the work presented and our experiences launching

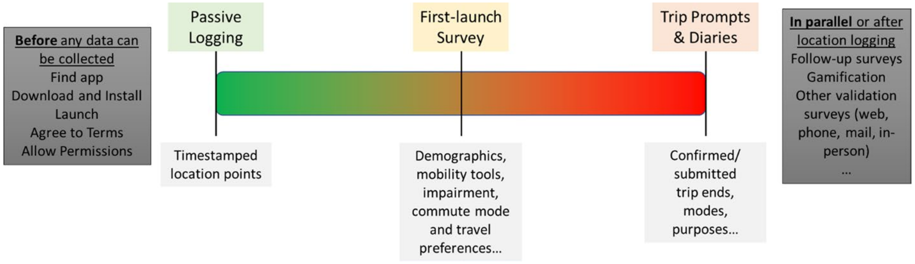
Before any data can be collected
Find app
Download and Install
Launch
Agree to Terms
Allow Permissions
Passive Logging
Timestamped location points
First-launch Survey
Demographics, mobility tools, impairment, commute mode and travel preferences…
Trip Prompts & Diaries
Confirmed/ submitted trip ends, modes, purposes…
In parallel or after location logging
Follow-up surveys
Gamification
Other validation surveys (web, phone, mail, in-person)
…

Fig. 12 Travel diary app burden spectrum

multiple apps, is that a few key pieces of information should be collected from respondents at first launch in any app-based travel data collection effort, in addition to location traces. These include (1) home and anchor locations, as well as (2) habitual and other commute modes. To these essential questions, we also highly recommend adding an initial launch question on (3) frequency of use of different modes and (4) trip prompts or diary validation for a handful of trips. The initial launch questions (1, 2 and 3), presented towards the middle of the "burden spectrum" in Fig. 12, add only a minute or two to the overall burden placed on users, comparable to download, install and setup time for sensor permissions, while massively improving the utility of the data recorded. A handful of validated trips in turn greatly improve the potential for accurate inference of mode used for travel episodes with passive location trace data alone, with an even greater improvement in the realm of discretionary travel as non-commuting trips will generate data which vary more greatly in attributes, while also being impossible to compare with reported commute modes through isolating comparable OD pairs.

As described in Harding (2019), users are not annoyed with answering a few simple and well-phrased questions. What they are annoyed with is being asked poorly worded questions, being prompted when not actually traveling and apps that drain their battery. And as the much greater challenge is recruiting respondents, not convincing them to answer 5 questions instead of 1, we must reiterate that some information should always be collected.

Significant differences in battery drain were observed between Android and iOS apps overall, with no pronounced difference in trip and mode detection accuracy. Battery drain differences were much more significant among the iOS apps (modeled spread of 42.9%) than among the Android apps (modeled spread of 8.4%). Regarding accuracy, while both iOS and Android contained high performing apps, iOS apps were more consistent. The trend toward higher accuracy with iOS is something likely to remain for years to come, and should be kept in mind when preparing datasets; the same way a response may be labeled as having been given over the web or the phone, self or proxy reported, the OS on which an app was run should become a standard field for smartphone app user tables.

Further, results indicate that higher recording frequency and location accuracy lead to improved trip end and mode inference accuracies, as expected. The relationship between frequency, accuracy, and overall performance, however, is not linear in form, which is what one might have expected. More accurately detected travel is somewhat associated with higher battery drain. There is, however, a point beyond which more frequent logging at high accuracy does not improve performance, but merely leads to higher respondent burden in the form of battery drain. This is indicative of the potential for a high performing app to be designed and tailored for regional travel survey needs that would perform well without causing excessive battery drain.

Also, if the point is simply to get a better understanding of the prevalence of trip under-reporting or of distances traveled and not reported, where very short trips or mode use is not of interest, it should be stated that apps have a great potential to be of use. Many of the problems exhibited by apps in our experiment result from short dwell times (minimum dwell times of 3 min instead of 10, 20 or more minutes), short distances traveled and difficult to distinguish modes in congested urban areas. If more information on these phenomena are not of interest, then the TDx results for trip identification with 80–90% accuracy are a clear indication that apps can be a tool to provide the answer to these questions if the data are processed properly. It becomes a question of what the point of the study is.

We can get total distances traveled by personal vehicle by asking for odometer readings, but perhaps it is of interest to better understand the timing of under-reported trips, in addition to their aggregate sum of travel distance.

If the goal of the smartphone data collection effort is to serve as a satellite for a regional and national household travel surveys, where the core is collected via telephone or web survey,– as is the case in most regions of the world at this time-, the smartphone app should aim at being a multi-modal travel diary app to be run for 5 to 7 days, collecting information in a mostly passive manner, allowing for mode of transportation to later be inferred (at least in the aggregate), and limiting the amount of short trips missed. Given the results obtained in our work, we would recommend the ideal app configuration would employ:

1. Constant recording if possible with low drain OR periodic logging of lower accuracy points to fill in gaps if high accuracy location recording is discontinuous;
2. Uniform or non-uniform recording frequency (neither appears to have a clear benefit);
3. Recording interval of 5 s while in motion (appears to be sufficient, with no clearly observable benefits of higher frequency);
4. Median recording accuracy of 15 or so metres (appears to be sufficient).

Such a survey then would allow for a greater understanding of trip under-reporting, active transportation demand, weekend and multiday travel patterns, trip chaining, travel time reliability, route choice and discretionary travel more broadly, to name the most obvious applications.

Finally, there is a problem with the lack of a standard data format or processing suite. This could mean that additional time and resources need to be allocated to merge datasets. More significantly still, differences in performance observed for the apps would be very difficult to control for over time.

One aspect of the data collection protocol that had a significant effect on the reported mode accuracy is that no within-respondent patterns could be observed in the travel data. As the modes used each day were random, there were no patterns to be observed, and it was not possible for processing suites which can learn and improve over time to apply these methods. Further, because of the design of the trip recording protocol, very short trips were only made a few times, with almost every single trip being at least 5 min long. As a result, the data produced can not be used to answer the question of whether purely passive traces can be used to differentiate, for example, an individual stepping into a store for 1 min before waiting for transit, from an individual who simply walked to a transit stop. There are ways to make use of sensor information like loss of GPS signal to better understand these grey areas with short stops, but our work does not shine a light on this. Given many surveys, the Transportation Tomorrow Survey in Toronto notably, do not currently include such 'incidental' stops (stepping into the dry cleaner or convenience store on the way to work, for example) (Malatest 2018), any improvement to the capture of short trips is an increase in the accuracy of reported travel.

The assessment should not be received as an indictment of smartphone apps, but a call for their roll-out to be carried out in a strategic manner that maximizes the likelihood of high-quality data being produced. The analysis in this paper does show that while smartphone apps are a promising tool for travel data collection, given their current performance, caution must be exercised in using the collected data. Some apps and approaches led to highly reliable travel diaries being reproduced from less-than-ideal traces (with shorter dwell times, higher randomness, absence of respondent preferences, limited learning from travel patterns, or using external dataset such as GTFS), while other apps and approaches are not fairing well in this specific testing context. Further, it is important to remember that the study has evaluated the 2016's state of the art. As smartphones evolve and new ways to interact with them emerge, the app performances are expected to improve. That being said, this paper should remain a valuable resource, as it guides as to what to look for, and look out for, in the apps that are employed for travel data collection.

Acknowledgements This study was funded by the Transportation Information Steering Committee for the Greater Toronto-Hamilton Area. The authors are solely responsible for all comments and interpretations presented in the paper. An abridged version of the paper was presented at the 2020 Annual Transportation Research Board Conference.

Author contributions The authors confirm contribution to the paper as follows: study conception and design: C. Harding, A. F. Imani, S. Srikukenthiran, K. M. N. Habib, E. J. Miller; data collection: C. Harding; analysis and interpretation of results: C. Harding, A. F. Imani; draft manuscript preparation: C. Harding, A. F. Imani; overall project supervision: K. M. N. Habib. All authors reviewed the results and approved the final version of the manuscript.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Allström, A., Kristoffersson, I., Susilo, Y.: Smartphone based travel diary collection: experiences from a field trial in Stockholm. In: Transportation Research Procedia, pp. 32–38. Elsevier B.V. (2017)

Auld, J., Williams, C., Mohammadian, A., Nelson, P.: An automated GPS-based prompted recall survey with learning algorithms. Transp. Lett. **1**, 59–79 (2009). https://doi.org/10.3328/TL.2009.01.01.59-79

Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp. Res. Part C Emerg. Technol. **17**, 285–297 (2009). https://doi.org/10.1016/J.TRC.2008.11.004

Dalumpines, R.: GIS-based episode reconstruction using GPS data for activity analysis and route choice modeling. https://macsphere.mcmaster.ca/handle/11375/15956 (2014)

Data Management Group: Transportation Tomorrow Survey. Ontario, Canada, Toronto (2016)

Ellison, A.B., Ellison, R.B., Ahmed, A., Rance, D., Greaves, S.P.: Spatiotemporal identification of trip stops from smartphone data. Appl. Spat. Anal. Policy. **12**, 27–43 (2019). https://doi.org/10.1007/s12061-016-9188-0

Enge, P., Misra, P.: Special issue on global positioning system. Proc. IEEE **87**, 3–15 (1999). https://doi.org/10.1109/JPROC.1999.736338

Flake, L., Lee, M., Hathaway, K., Greene, E.: Use of smartphone panels for viable and cost-effective GPS data collection for small and medium planning agencies. Transp. Res. Rec. J. Transp. Res. Board. **2643**, 160–165 (2017). https://doi.org/10.3141/2643-17

Geurs, K.T., Thomas, T., Bijlsma, M., Douhou, S.: Automatic trip and mode detection with move smarter: first results from the Dutch Mobile Mobility Panel. Transp. Res. Proc. **11**, 247–262 (2015). https://doi.org/10.1016/j.trpro.2015.12.022

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., Crane, M.: A web-based diary and companion smartphone app for travel/activity surveys. Transp. Res. Proc. **11**, 297–310 (2015). https://doi.org/10.1016/j.trpro.2015.12.026

Harding, C.: From smartphone apps to in-person data collection: modern and cost-effective multimodal travel data collection for evidence-based planning (2019)

Harding, C., Nasterska, M., Dianat, L., Miller, E.J.: Effect of land use and survey design on trip underreporting in Montreal and Toronto's regional surveys. Eur. J. Transp. Infrastruct. Res. (2018). https://doi.org/10.18757/ejtir.2018.18.1.3218

Jariyasunant, J., Sengupta, R., Walker, J.: Overcoming battery life problems of smartphones when creating automated travel diaries (2014)

Joseph, L., Neven, A., Martens, K., Kweka, O., Wets, G., Janssens, D.: Measuring individuals' travel behaviour by use of a GPS-based smartphone application in Dar es Salaam, Tanzania. J. Transp. Geogr. (2019). https://doi.org/10.1016/j.jtrangeo.2019.102477

Krenn, P.J., Titze, S., Oja, P., Jones, A., Ogilvie, D.: Use of global positioning systems to study physical activity and the environment: a systematic review. Am. J. Prev. Med. **41**, 508–515 (2011). https://doi.org/10.1016/J.AMEPRE.2011.06.046

Kunzmann, M., Daigler, V.: 2010–2012 California household travel survey final report (2013)

Lawson, C.T., Chen, C., Gong, H.: Advanced applications of person-based GPS in an urban environment (2010)

Li, Z.J., Shalaby, A.S.: Web-based GIS system for prompted recall of GPS-assisted personal travel surveys: system development and experimental study. In: Transportation Research Board 87th Annual Meeting, Washington DC (2008)

Lynch, J., Dumont, J., Greene, E., Ehrlich, J.: Use of a smartphone GPS application for recurrent travel behavior data collection. Transp. Res. Rec. J. Transp. Res. Board. **2673**, 89–98 (2019). https://doi.org/10.1177/0361198119848708

Malatest: Transportation Tomorrow Survey 2016: Design and conduct of the survey (2018)

Marra, A.D., Becker, H., Axhausen, K.W., Corman, F.: Developing a passive GPS tracking system to study long-term travel behavior. Transp. Res. Part C Emerg. Technol. **104**, 348–368 (2019). https://doi.org/10.1016/j.trc.2019.05.006

Miller, E.J., Harding, C., Zhang, Y.: Waterfront Toronto transportation carbon model system update—final project report (2016)

Nahmias-Biran, B., Han, Y., Bekhor, S., Zhao, F., Zegras, C., Ben-Akiva, M.: Enriching activity-based models using smartphone-based travel surveys. Transp. Res. Rec. J. Transp. Res. Board. **2672**, 280–291 (2018). https://doi.org/10.1177/0361198118798475

Patterson, Z., Fitzsimmons, K.: DataMobile smartphone travel survey experiment. Transp. Res. Rec. **15**, 35–43 (2016). https://doi.org/10.3141/2594-07

Prelipcean, A.C., Yamamoto, T.: Workshop Synthesis: New developments in travel diary collection systems based on smartphones and GPS receivers. In: Transportation Research Procedia, pp. 119–125. Elsevier B.V. (2018)

Randell, C., Djiallis, C., Muller, H.: Personal position measurement using dead reckoning. In: Seventh IEEE International Symposium on Wearable Computers, pp. 166–173. IEEE (2003)

Schlich, R., Axhausen, K.W.: Habitual travel behaviour: evidence from a six-week travel diary. Transportation (Amst) **30**, 13–36 (2003). https://doi.org/10.1023/A:1021230507071

Schönfelder, S., Axhausen, K., Antille, N., Bierlaire, M.: Exploring the potentials of automatically collected GPS data for travel behaviour analysis a Swedish data source. Arbeitsberichte Verkehrs-und Raumplanung (2002). https://doi.org/10.3929/ethz-a-004403386

Stenneth, L., Wolfson, O., Yu, P.S., Xu, B.: Transportation mode detection using mobile phones and GIS information. In: GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, pp. 54–63. ACM Press, New York (2011)

Stopher, P., Shen, L.: In-depth comparison of global positioning system and diary records. Transp. Res. Rec. J. Transp. Res. Board. **2246**, 32–37 (2011). https://doi.org/10.3141/2246-05

Wolf, J., Bachman, W., Oliveira, M.S., Auld, J., Mohammadian, A., Vovsha, P.: Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines. Transportation Research Board, Washington, D.C. (2014)

Wolf, J., Bricka, S., Ashby, T., Gorugantua, C.: Advances in the application of GPS to household travel surveys. In: National Household Travel Survey Conference., Washington DC (2004)

Wu, L., Yang, B., Jing, P.: Travel mode detection based on GPS raw data collected by smartphones: a systematic review of the existing methodologies. Information **7**, 67 (2016). https://doi.org/10.3390/info7040067

Xiao, G., Juan, Z., Zhang, C.: Travel mode detection based on GPS track data and Bayesian networks. Comput. Environ. Urban Syst. **54**, 14–22 (2015). https://doi.org/10.1016/j.compenvurbsys.2015.05.005

Zhao, F., Ghorpade, A., Pereira, F.C., Zegras, C., Ben-Akiva, M.: Stop detection in smartphone-based travel surveys. Transp. Res. Proc. **11**, 218–226 (2015). https://doi.org/10.1016/j.trpro.2015.12.019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chris Harding is a Planner in the Systems Analysis and Forecasting Office at Ontario's Ministry of Transportation. Both as a research intern in the Chaire Mobilité at École Polytechnique and during his PhD at University of Toronto, he worked on travel survey methods. Other areas of research are GHG modelling, public participation and its role in the planning process, and transportation and land use linkages (with a focus on active transportation).

Ahmadreza Faghih Imani is a research associate at Urban Systems Laboratory and Centre for Transport Studies at Imperial College London. Before joining Imperial College in October 2019, he was an NSERC Post-Doctoral Fellow at the University of Toronto Transportation Research Institute. He obtained his Ph.D. in Civil Engineering (Transportation) from McGill University in 2017. His dissertation on examining bicycle-sharing systems was recognized with the 2015 Benjamin H. Stevens Graduate Fellowship in Regional Science from North American Regional Science Association (NARSC). He is a member of the Transportation Research Board's (TRB) Committee on Bicycle Transportation (ANF20).

Siva Srikukenthiran is the Director of Data Science at Ratio.City. He previously was a Research Associate at the University of Toronto Transportation Research Institute, where he managed research programmes in transit simulation, schedule optimization and travel survey data collection, and developed software systems for survey data collection and transit network simulation. He has a PhD in Civil Engineering from the University of Toronto, and is a member of the Transportation Research Board's Committee on Passenger Intermodal Facilities.

Eric J. Miller (BASc, MASc University of Toronto; PhD Massachusetts Institute of Technology) has been a faculty member in the Department of Civil & Mineral Engineering, University of Toronto since 1983, where he is currently Director of the University of Toronto Transportation Research Institute. Research areas include activity-based travel modelling, integrated transport – land use modelling and agent-based microsimulation. He is the recipient of the 2018 International Association for Travel Behaviour Research Lifetime Achievement Award.

Khandker Nurul Habib is the Percy Edward Hart Professor in Civil & Mineral Engineering at the University of Toronto. He is a member of TRB committees AEP30, AEP50. He reworks on developing appropriate planning methodology and policy analysis tools (mathematical models) that can give realistic predictions of travel demands for short-, medium- and longterm transportation planning exercises.

Affiliations

Chris Harding[1] · Ahmadreza Faghih Imani[2] · Siva Srikukenthiran[3] · Eric J. Miller[4] · Khandker Nurul Habib[4]

Chris Harding
Chris.harding@ontario.ca

Siva Srikukenthiran
siva@ratio.city

Eric J. Miller
miller@ecf.utoronto.ca

[1] Systems Analysis and Forecasting Office, Ontario Ministry of Transportation, Toronto, Canada

[2] Civil and Environmental Engineering, Imperial College London, London, UK

[3] RATIO.CITY, Toronto, Canada

[4] Civil and Mineral Engineering, University of Toronto, Toronto, Canada