International Conference on Transport Survey Methods (ISCTSC 2017)

On the User Experience and Performance of Smartphone Apps as Personalized Travel Survey Instruments: Results from an Experiment in Toronto

Chris Harding[a], Siva Srikukenthiran[a], Khandker Nurul Habib[a] & Eric J. Miller[a]*

[a]*University of Toronto Transportation Research Institute, 35 St George street, Toronto, Ontario, M5S 1A4, Canada*

Abstract

In the context of a major shift away from traditional household survey data collection methods in the Toronto region, an experiment was designed to assess the state of the art in terms of smartphone app performance and design for location logging and trace processing. Instead of testing a single app, we contacted leaders in app design on a global scale and invited them to provide their apps and processing suites/expertise to make possible a comparison of performance and user experience. We reached out to academics, commercial app developers, and the open source community, both specifically with regard to travel surveys, as well as more broadly in the field of location logging apps, and were met with a high number of respondents willing to participate. From this initial list of potential participants, we chose a smaller group we thought represented different approaches, as well as apps and processing suites which could appropriately be deemed representative of the state of the art; this was based on project maturity and track record, current version user interface and quality of output. 8 Android and 9 iOS apps were assessed. Only one standalone trace processing suite was made available, a problem explored.

While much has been written about travel survey smartphone apps, the assessment of performance leaves much to be desired. With respect to the reporting on app performance, we have found (at least) five major problems. First, there is no universal set of assessment metrics; second, the results are only relevant to the travel patterns observed in a given area; third, assessment is usually carried out by the same team who designed a particular app; fourth, there is either no 'ground truth' data against which the apps are measured, or the ground truth data are so small in terms of sample size that it is hard to tell if the methods applied are robust; and fifth, battery drain is either loosely reported on as a problem to be overcome later, or reported as being high but offset by incentives, with some form of reporting on who complied with the expected protocol.

We set out in our assessment to minimize the influence of the problems laid out above. A data collection protocol was designed that allowed us to have a 'ground truth' set of data against which to compare all recorded and processed traces, while also taking frequent measurements of battery levels (1,063 in total) and assessing the usability of apps in a variety of dimensions. The latter, more design-focused evaluation, ranges from assessment of installation procedure and overall user interface, to task-specific learning curve, degree of user interaction required for proper functioning, leg and trip validation interface and associated burden, use of maps and load time/responsiveness. By outlining best practices in terms of key app design dimensions, we hope to provide guidance to maximize respondent engagement and data collection protocol compliance, while minimizing burden, bias, and errors.

* Chris Harding. Tel.: 1-647-963-6950.
E-mail address: chris.harding@mail.utoronto.ca .

To ensure performance scores for apps were robust for each handset type, operating system, mode of transportation and urban form, data were collected using 21 smartphones simultaneously. Each device had one app installed at any given time, while the devices were carried around together at all times, such that the same trips would be recorded. Apps were cycled through devices three times per week – installed on different phones-, while trips consisting of a variety of distances, transportation modes and destinations were made. This procedure was carried out for 314 trips over 34 days, accounting for 3,655 km, all the while carefully logging start and end times, mode change points, and modes of transportation used on each leg. Comparisons were then made between the output of each app (processed using app-specific and generic trace processing suites) and a record of the ground truth. Every trip was carefully recreated in ArcGIS, identifying every link along the route so that we could later verify whether the data output by different apps could be assigned to the correct path once processed. Inferred mode of transportation for all trips was also compared to the ground truth.

Using data on trips collected using 17 different apps on 21 devices at once, we quantify relative app and processing suite performance in terms of i) the percent correctly identified leg or trip ends, ii) percent of trips with mode accurately inferred; iii) percent route overlap; and iv) battery drain relative to travel time and accuracy. Taking our analysis a step further in order to better understand what these app and processing suite performances mean in context, we also transposed our findings to the Toronto region. We estimated what percent of travel generated by residents of different areas of the city would be accurately detected using these different apps, as well as what the associated battery drain given travel reported in previous surveys.

Results indicate that while trip end detection is accurate for most apps, mode inference is considerably less so, at least in denser, more congested environments where differences in speed between transit, car and bike trips is less distinctive. Key design decisions revolving around recording frequency, continuous logging and location accuracy are explored, as non-linear relationships are found between increased frequency, accuracy and ability to reproduce ground truth data. With respect to battery drain, our results indicate it is possible to design an app that provides both high quality data while not imposing an excessive burden on respondents with respect to battery life.

One of the most important conclusions is that a lack of standards for recording and processing data limits the potential for use of smartphone data. Differences in performance between apps, whether commercial or academic, is also problematic, as this introduces a large potential instrument bias when using current generation smartphone apps for data collection.

As one of the goals of this research was to provide guidance to agencies deciding how to proceed with integration of smartphone apps within their data collection ecosystem, a framework for app evaluation was generated, listing features to request, as well as best practices with regards to design. One feature which is of particular importance in light of disappointing or highly variable observed mode inference accuracy is validation. This can be carried out in real-time, triggered by prompts, or take the form of travel diaries that are reviewed periodically. Until an improvement in inferred travel episode attributes is observed, however, it is recommended to include validation of travel attributes as part of the survey instrument design, at least for a subset of collected travel data. Apps have great potential to contribute to regional travel data collection efforts, but diaries extracted from purely passive location data are not yet accurate enough to fully unburden respondents.

Results should be of value to any municipality or organization aiming to carry out a smartphone data collection exercise. They should also assist any organization wishing to better understand what portion of their resources they should allocate to smartphone data collection, and what type of app to employ so as best to maximize their return on investment.

Keywords: Data collection; smartphone travel surveys; location logging; trace processing; respondent burden; user experience