Collecting and using always-on location data in surveys

Stephanie Eckman, RTI International, USA
Rob Chew, RTI International, USA
Herschel Sanders, RTI International, USA
Robert Furberg, RTI International, USA

How to cite this article : Eckmann, S., Chew R., Sanders H. & Furberg R. (2020). Collecting and using always-on location data in surveys in Survey Methods: Insights from the Field, Special issue: 'Advancements in Online and Mobile Survey Methods'. Retrieved from https://surveyinsights.org/?p=13330

DOI : 10.13094/SMIF-2020-00022

Abstract : As survey costs increase and response rates decrease, researchers are looking for alternative methods to collect data from study subjects. Passively collected data may offer a way to reduce the burden on research subjects while also collecting high-quality data needed for social science research. Examples of passive data collection tools are applications installed on mobile devices and sensors in subjects' homes or worn on the body. In this study, we focus on always-on location data collected from subjects' iPhones. To explore the promise of passively collected data to augment and improve survey data, we conducted a 2-week pilot study with 24 subjects. We discuss the utility of always-on location data and the challenges researchers may encounter when they incorporate location data in their analyses.

Introduction

The traditional model of survey research—a lengthy survey instrument collecting all measures of interest, a high response rate, a random sample from all population members—is in crisis (National Academies of Sciences, 2017, 2018). Research subjects are increasingly intolerant of long questionnaires (Mavletova & Couper, 2015; Tourangeau, Kreuter, & Eckman, 2015), response rates are falling (Brick & Williams, 2013; de Heer & de Leeuw, 2002; The Economist Group, 2018), and incentives are not always effective at offsetting these trends (Mercer, Caporaso, Cantor, & Townsend, 2015). In this environment, how can researchers collect the data necessary to understand society?

Passive data collection may offer one way forward. By passively collected data, we mean data gathered without the direct involvement of research subjects. For example, rather than being asked numerous survey questions about their exercise and sleep, subjects could wear sport watches to track steps, heart rate, sleep, etc. Such data would reduce the recall and time burden placed on subjects and might also provide more accurate data. Surveys could then focus on asking about attitudes, characteristics, and behaviours not available via passive data collection.

This paper explores one specific source of passively collected data: Global Positioning System (GPS) data collected from mobile devices. Data on where a device has been may contain useful information about

subjects. From these data, we might be able to infer, with varying degrees of accuracy, characteristics such as:

- what Census block the subject lives in, which is highly correlated with race and income in the United States;
- whether she regularly visits day-care or school (evidence that she has children);
- whether she attends religious services;
- where she works and what hours; and
- how frequently, how long, and in what ways she exercises (at the gym/running/ biking).

These variables are not present in the location data themselves: GPS sensors record latitude and longitude with a corresponding date-time stamp. However, traces of these characteristics and behaviours are present. If we can impute these characteristics with reasonable accuracy, we could remove them from the survey instrument, which could increase response rates and decrease data collection costs. In this paper, we describe our experiences in collecting and analysing always-on location data alongside survey data in a pilot study. We address the following research questions:

1. Can we determine where a subject was (grocery store, dentist) from the GPS coordinates?
2. Do passively collected measures of where subjects were agree with survey responses?
3. Can we identify subjects' workplace through mobile device GPS data?
4. What are subjects' attitudes toward passive data collection?

Small studies such as this pilot are a necessary first step in understanding how we might transition from survey to passive data collection. Although this case study does not offer definitive answers, we hope it will help inform future surveys interested in using passively collected location data.

Data collection

We recruited subjects via an e-mail to colleagues in two departments at RTI International. All participants in the study were RTI employees who owned iPhones. Subjects were in the study for 2 weeks between January 28 and February 24, 2017. Previous research has shown that 2 weeks of location data is enough to understand subjects' activity spaces (Stanley, Yoo, Paul, & Bell, 2018). The study protocol was approved by RTI's Office of Research Protection and the legal and human resources departments.

To meet the aims of our study, we used a combination of survey and passive data collection. All subjects completed daily surveys: participants downloaded an application to their phones which asked survey questions each day. The survey was just two questions long. At the end of the 2 weeks of data collection, subjects completed an outtake survey. This web survey asked about experiences with the study and the subjects' familiarity with common digital and Internet topics and products. Subjects also installed the Moves application (Evenson & Furberg, 2017), which passively collected location data from the phone: time, date, and GPS coordinates. The Moves application is no longer available from the Apple App Store. Arc App (https://www.bigpaua.com/arcapp/) was developed specifically to replace the Moves application and to offer similar functionality. We have not replicated our data collection with Arc App, however. Both Moves and Arc App are only available for iOS devices. Location data were collected whenever the phone was on and moving.

Forty-six subjects expressed initial interest in the study. After reading the informed consent document for this study, four who had expressed interest chose not to participate. Others dropped out without explicitly

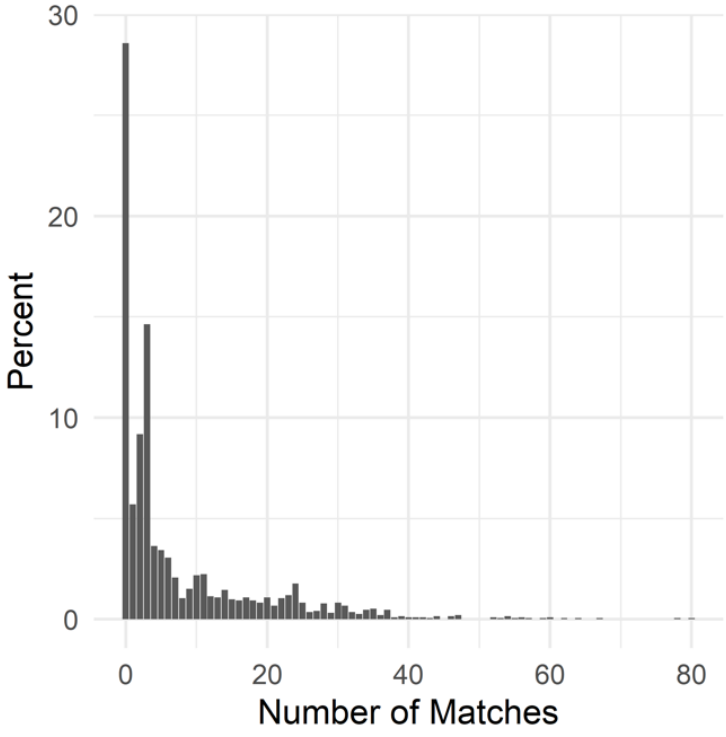
giving a reason. We have location data from 24 subjects, and 21 completed all phases of data collection.

The output of the Moves app is not raw sensor data, but rather processed travel or location data. The Moves app generates two datasets: (1) a places file—a list of coordinates where the subjects stopped and spent some time, and (2) a traces file—a database of subjects' travel paths. The algorithm used to determine what constitutes a place is not published by the application developers. (Arc App, however, seems to publish its algorithms on GitHub, though we have not rigorously reviewed them.) We only use the places file because of our focus on locations subjects visited. Across all subjects and days, we collected coordinates of 1,928 places. The number collected per day and subject varied from 1 to 11, median 5, mean 5.6.

RQ1: Determining where a subject was

The places dataset contains the date, start time, end time, and latitude and longitude of each location a subject visited. To replace survey responses from these data, however, we must first figure out the real-life sites (office, store, park) each subject visited. We queried three popular online databases of business and other points of interest (PoIs): Google Places (Google Maps Platform, 2018), Yelp (Yelp, 2018), and Foursquare (Foursquare, 2017). We used each site's Application Programming Interface to generate candidate PoIs within 100 meters of each place coordinate.

Figure 1: Number of Matches per Place



Matching a given place to a single PoI was challenging. Figure 1 shows the distribution of the number of matches to each of the 1,928 places. The modal number of matches was zero. Nevertheless, 1,377 coordinates had at least one matched PoI. The databases returned 14,643 PoIs within 100 meters of these places. For the purposes of this article, we retained only the closest PoI from each of the three databases,

leaving 2,536 candidate PoIs for 1,377 places. There is more than one PoI for each place, on average, because the databases often did not agree on which PoI best matched a place. For example, one coordinate matched to a grocery store in the Google database, a liquor store in the Foursquare database, and a dentist's office in the Yelp database.

Manual matching of the PoIs across the three datasets by name revealed only 53 three-way matches. The three sources were more likely to agree when the PoI was large and isolated, such as a large retail store, a university, or a large church. Sources were less likely to agree about smaller locations such as restaurants, professional offices, and coffee shops. Foursquare was the clear outlier, having more and different types of matches, such as "Work Break Room" and "Paul's Apartment." It contained separate PoIs for many buildings on the RTI campus and the softball field. Foursquare is a user-driven community more so than Google and Yelp, which may account for these differences.

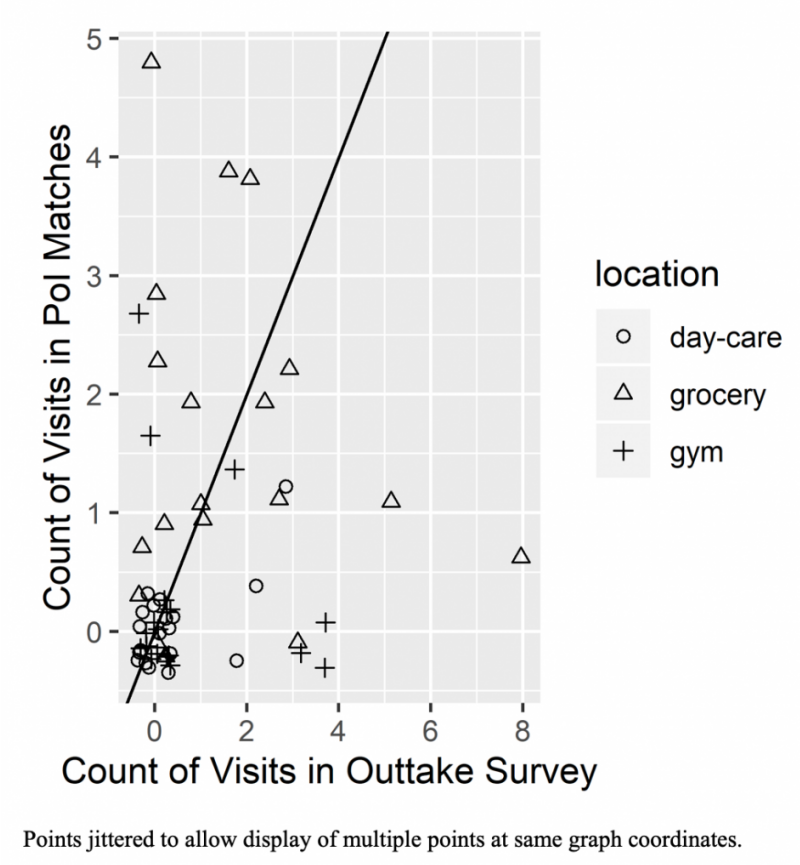
Several sources of error could lead a place to be matched to the wrong PoI. The collected GPS coordinate of the place recorded in the Moves data could be off by several meters: smartphone GPS is accurate to approximately 5 meters under ideal conditions (Van Diggelen & Enge, 2015). The PoI databases could also be wrong about the name or location of a PoI—perhaps the ice cream store recently went out of business and the subject in fact visited a chiropractor's office. Lastly, accurate matching becomes more difficult in dense commercial areas where there are several PoIs near a recorded coordinate. This issue is particularly problematic in mixed-use developments: visiting friends or family living in an apartment above a row of stores could trigger a false positive detection of a ground-floor PoI.

RQ2: Agreement between survey responses and passively collected data

Despite the disagreement between PoI databases, we are still interested in how closely the inferred PoI visits from the location data match reported survey responses. For the survey, we used items from the outtake questionnaire asking subjects to report the number of times they had been to day-care centres, grocery stores, and gyms during the study period. For the passively collected location data, we manually coded the closest PoIs for each coordinate to flag the ones falling into these three categories. When the PoI sources disagreed, we considered a subject to have visited a PoI whenever any of the sources indicated that she had. We summed within subjects to get the number of visits to day-care centres, grocery stores, and gyms in the PoI data.

Figure 2 shows scatterplots of the number of visits reported in the outtake survey (on the horizontal axis) and the number found in the PoI data (on the vertical axis). The diagonal black line shows the points of agreement between the two sources. Away from the origin, we see very little agreement between the two sources. The greatest deviation in agreement is for grocery visits. There are several points both below the black line (respondents reported more visits than we see in the data) and above the black line. The smallest deviation occurs for day-care visits, although most of the agreement in this category is from points at the origin, where neither source indicated a visit to a day-care facility. Interestingly, although we expected overidentification of PoIs in the passively collected location data, based on our generous assignment across databases, the survey counts are often greater.

Figure 2. Comparison of Visits Counts in Survey Responses and Location Data



Points jittered to allow display of multiple points at same graph coordinates.

RQ3: Identifying work location

Although successfully matching subjects' passively collected coordinates to PoIs is difficult without context, we can begin drawing more compelling insights if we include additional information on subjects or if we include behavioural assumptions as part of matching algorithms. To demonstrate this approach, we used a combination of unsupervised learning methods and common-sense decision rules to identify where subjects work from the places dataset. In our analysis, we included only subjects who work at RTI headquarters (n=22). However, we used this information only to validate the model, not to build it. Thus, our results are applicable to other surveys where the work location is unknown and only GPS coordinates are available. We used the Python programming language for wrangling and analysing the data and the scikit-learn library (Pedregosa et al., 2011) for clustering.

First, we filtered each subject's places coordinates to only those occurring between the hours of 8:00 AM and 6:00 PM, Monday through Friday. Records were included for further analysis if (1) both the start and end times lie entirely within the interval (e.g., 8:30 AM–11:00 AM); (2) the start time begins before and end time finishes after the interval (e.g., 7:00 AM–6:45 PM); or (3) either the start or end times fall within the interval and more time was spent inside the interval than outside (e.g., 7:00 AM–11:30 AM, corresponding to 1 hour outside and 3.5 hours inside). These hours were chosen to reflect the most common business hours for industries using a 40-hour workweek in the United States. Researchers working with different populations (e.g., students, workers in the hospitality industry) should modify their query to better reflect the expected work patterns for their sample.

Next, we truncated the latitude and longitude coordinates to the thousandths place. Truncating the digits helps smooth and reduce noise in the clusters; we want to identify a workplace instead of detecting different spots in the parking lot. We used the DBSCAN algorithm (Davis, 2014; Ester et al., 1996) to

develop clusters of coordinates. DBSCAN is a density-based clustering algorithm that groups dense neighbouring points. The algorithm has several nice properties that are useful for this type of task: (1) DBSCAN does not require specifying the number of clusters up front, as opposed to other popular clustering methods like K-means (Hartigan & Wong, 1979); (2) DBSCAN has a natural notion of outliers and assigns all points lying in low-density areas to a catch-all "outlier" cluster; and (3) the algorithm's tuning parameters have a useful interpretation for GPS coordinates. The two parameters are (1) the radius surrounding each point that should be considered when determining neighbouring points for cluster assignment, and (2) the minimum number of points that must be densely connected to be considered a cluster. Generally, the larger the radius, the larger the cluster membership will be. For our model, the radius parameter was set to 0.2 km and the minimum points parameter was set to 5.

All clustering algorithms require a distance metric to determine similarity between points. Because our points have a geographic interpretation, we used the haversine formula (Bullock, 2007) to calculate pairwise "great-circle" distance in kilometres between each coordinate for each subject. The "great-circle," or orthodromic, distance is the shortest distance between two points on the surface of the Earth. Although using a Euclidean distance approximation is likely fine for determining candidate work locations within a short commute, we opted for the orthodromic distance to help with edge cases where distances travelled are longer, especially in areas farther from the equator where the distortion between distance calculations is more pronounced.

We ran DBSCAN independently on each subject to create clusters. We then coded the cluster where the subject spent the most time 8:00 AM–6:00 PM, Monday through Friday, as the workplace location. If the subject spent the most time in the outlier category, then the cluster with the second longest duration was assigned as the workplace.

To assess this method, we compared the predicted workplace location to the location of RTI's Research Triangle Park, NC, campus. If the predicted workplace location fell within 0.5 km of the RTI campus centroid, we called the prediction a success. Twenty-two of our 24 subjects had location data in the Research Triangle Park area. For these 22, this heuristic correctly identified workplaces for all but 2 subjects (90.9%). Upon further inspection, the two misclassified clusters were apartments. Those subjects may have worked from home more often than on campus during the field period. For both misclassified subjects, the second most common cluster was the RTI campus. Thus, our approach was largely successful at identifying subjects' workplaces in this relatively homogenous population.

RQ4: Subjects' attitudes toward passive data collection

The outtake survey collected information about the subjects' experience with the passive data collection. Table 1 shows the negative effects of the passive data collection reported by the subjects. A majority indicated no negative effects on their smartphone performance.

Table 1. Frequency of Problems Encountered by Subjects with Passive Data Collection (n=22)

Problem[a]	# Indicating	%
Battery ran down faster	10	47.6
Device was slower	1	4.8
Device was less responsive	1	4.8
None of the above	11	52.4

[a] Subjects could check all problems they encountered.

One concern with passive data collection is that subjects might change their behaviour when the data are being collected. Similar effects occur with survey data collection (Bach & Eckman, 2018; Crossley et al., 2017; Dholakia, 2010; Traugott & Katosh, 1979). Estimating such survey conditioning effects is quite challenging (Bach, forthcoming), and our study was not designed to do so.

Table 2. Frequency of Thinking about Passively Collected Data (n=22)

How Often	# Indicating	%
Never	3	14.3
Every time I was asked to respond to a daily survey	9	42.9
A lot in the beginning, but less so over time	7	33.3
All the time	2	9.5

However, the outtake survey asked some questions to touch on this issue. One question asked how often the respondents thought about the passive data that the study was collecting via their mobile phones (Table 2). Two respondents reported that they thought about it all the time. Another question asked if respondents changed their behaviour at all in response to the passive data collection (Table 3). One subject indicated that he or she changed behaviour while the data were being collected, but most (81%) said they did not. In addition, 19 participants (90%) said they definitely or probably would participate in another survey that combined survey and passive data collection, including the two who reported that they thought about the collection all the time.

Table 3. Changed Behavior Because of Location Data Collection (n=22)

How Often	# Indicating	%
Yes, throughout the field period	1	4.84
Yes, but only at first	3	14.3
No	17	80.6

Of course, we do not have responses to these questions from those who chose not to take part in the survey. Thus, we do not know what aspects of the study caused them to opt out.

Discussion

Although our study was small and limited to our colleagues, it reveals important lessons for other researchers who are considering passive data collection. We suggest that researchers interested in incorporating passive data collection in their studies also start with small data collection studies to gain a hands-on understanding of the challenges involved. It is not cost-effective or ethical to collect data from subjects without a good plan for processing, storing, and analysing them. Researchers should also think carefully about how the data will be stored and transferred and who will have access at each stage.

An important finding from this study was that the location data we collected were challenging to work with. Our approaches to matching coordinates to PoIs was not always successful: we found both too few and too many matches. Even interpreting agreement between passively collected location data and survey responses is complex, because both may have errors. Much more research is necessary before researchers can use location data to impute subject characteristics. Future research should investigate the use of the trace data as well as the places data.

Always-on location data must be collected, stored, and used properly, with full knowledge and consent on the part of the study participants. Data such as the places and traces files collected in this study cannot help but reveal where subjects live and work, and where and when they travel around their neighbourhoods. The data should probably be considered personally identifiable information and should not be released to any researchers outside of the study team. The usual methods of anonymizing survey data such as review of outliers and separation of identifiers from survey data do not work with location data (Cassa, Wieland, & Mandl, 2008; Zang & Bolot, 2011). We anticipate a growing interest in passive data collection in the future and encourage researchers to develop standards and best practices for the collection, handling, storage, and release of such data.

Social science researchers are not the only ones working on understanding the places that people visit. Google, Yelp, Facebook, and other technology firms are far ahead in developing these capabilities, in large part because of their extensive data resources and business interest in selling targeted advertising. These firms are unlikely to share their proprietary algorithms with social science researchers, and we cannot share the confidential data we collect from our subjects with them. We hope to find a way for these two sets of researchers to combine efforts in the future. We are encouraged by the potential of passive location data and support the multidisciplinary research effort needed to make continued progress.