EM-EVAL: AN EVALUATION PROCEDURE FOR SMARTPHONE-BASED TRAVEL DIARIES

K. Shankari
shankari@eecs.berkeley.edu

Hari Venugopalan
hvenugopalan@ucdavis.edu

David E. Culler
culler@berkeley.edu

Randy H. Katz
randy@cs.berkeley.edu

Word Count: 6476 words + 1 table(s) $\times$ 250 = 6726 words

Submission Date: August 2, 2019

Abstract

As smartphone-based Human Mobility Systems (HMSes) shift from informing individuals to influencing public asset allocation, it is critical to assess how well they perform. Evaluation techniques of HMSes have typically been ad-hoc, limited to single measurements and ignored power-accuracy tradeoffs. These tradeoffs are important due to built-in and context-sensitive sensing in the phones - e.g. “Find my iPhone" or “Doze mode".

Treating HMSes as physical instruments with inherent physical noise, we propose a novel evaluation procedure that uses artificial trips and multiple parallel phones to provide controlled, repeated inputs to the HMSes under test. Artificial trips mitigate privacy concerns and allow repeatability while efficiently exploring a wide variety of trip contexts. Parallel sensing with control phones mitigates the effects of context sensitive power consumption and inherent sensing error.

We use this procedure to create three artificial timelines with 15 different modes, including ebike and escooter. We travel these three timelines three times each to collect and publish a dataset with over 500 hours of ground truthed data. We use this data to explore the tradeoffs for power versus spatial and temporal accuracy.

Our results show the benefits of thinking statistically about HMSes. They control for outliers, revealing meaningful signals about the behavior of smartphone virtual sensors that are relevant to instrumenting human travel behavior. If adopted widely by the community, the resulting ground truthed, tradeoff-aware, public datasets can form the basis for additional HMS optimizations.

Contents

INTRODUCTION

Inspired by the popularity of smartphone-based personal fitness tracking, the transportation community aims to build Human Mobility Systems (HMSes) that can automatically track and classify multi-modal travel patterns. Such systems can replace expensive and infrequent travel surveys with long-term, largely passive data collection augmented with intermittent surveys focused on perceptual data. Such data collection can capture changes in travel behavior as they occur and open new avenues for responsive urban planning.

While there has been much work on building HMSes, both in academia and in industry, the procedure to *evaluate* them has largely been an afterthought. Careful evaluations are critical as we move from the personal to the societal domain. Users who make decisions based on self-tracking have an intuition of its accuracy based on their experienced ground truth. The decisions are low-stakes lifestyle changes, which may be personally meaningful, but are not societally contentious. However, a Metropolitan Transportation Agency picking projects and allocating millions of dollars in funding needs to know the accuracy of the data before making its decisions (*18*).

The typical HMS evaluation procedure (e.g. Quantified Traveler (*7*), prior versions of our work (*8, 17*)) is ad-hoc and also functions as a pilot - a small ($\approx$ 3-12) set of the author's friends and family are recruited to install the app component of the HMS on their phones, and go about their daily life for a few days or weeks while annotating the trips with "ground truth". The ground truth annotation can either directly happen on the app, or through a recap at the end of the day. Conscientious researchers may ensure that the set of evaluators are demographically diverse, in an attempt to evaluate against a richer set of travel patterns.

While this procedure imposes little additional researcher burden, it conflates the *experimental* procedure (understanding human travel behavior in the wild) with the *evaluation* procedure (evaluating the instrument that will measure the human travel behavior). The first is trying to understand *behavior*, so it needs human diversity. The second is trying to understand *sensing* parameters, so it needs diversity of *trip types*. The human functions as a phone transportation mechanism during evaluation and could be profitably replaced with a self-navigating robot if one was available.

An analogy with classic physical measurements may be useful. Consider the situation in which a researcher wants to collect data on the weight distribution of the population in a particular region. Since there are currently no certifying bodies for travel diaries, let us pretend that she cannot purchase a pre-certified scale. How would she evaluate the available scales before starting her experiment?

The analog to ad-hoc evaluation procedure would involve recruiting several of her friends and family to weigh themselves on the scales and compare the reported weight with their true weight. This analogy clearly reveals some of the limitations of the ad-hoc procedure: (i) How does she trust that the self-reported weights are "true"? (ii) If all her friends are adults weighing 55kg - 75kg, how does she know how the scales perform outside that range? She can overcome the range limitations by recruiting an broader set of testers, e.g. through an intercept survey. However, that modification makes the ground truth limitation worse, since it is less likely that strangers will reveal their true weight. A further modification might pay contributors to improve the self-reported accuracy, but at this point, she is essentially running the experiment.

A more robust evaluation procedure would involve choosing known weights across a broad range (e.g. 0kg to 300kg in 10kg increments) and comparing them to the reported weights. Since no instrument is perfect, there is likely to be some variation in the values reported. She would

likely repeat the experiments multiple times in order to establish error bounds.

HMS evaluation procedures need to be more sophisticated than simple physical measurements since: (i) their operation is based on prior behavior (e.g. HMS duty cycling, android doze mode) and the potential for feedback loops makes it important to control the *sequence* of evaluation operations, (ii) unlike a physical scale, which has a fixed one-time cost, they have an ongoing, variable cost in terms of battery drain, so the evaluation must assess the power/accuracy tradeoff, and (iii) unlike scalar weight data, HMSes generate strongly correlated timeseries data, which is extremely hard to deanonymize.

The main contributions in this paper are:

1. We propose an **evaluation procedure** for HMSes based on pre-defined, ground truthed, artificial trips and outline how it addresses the above challenges
2. We describe the design of a cross-platform **evaluation system** that can be used to perform such evaluations reproducibly and publish the results.
3. We use this system to evaluate the **power/accuracy tradeoffs** of the android and iOS location, motion accuracy and visit detection virtual sensors in the San Francisco Bay Area.

The rest of this paper is structured as follows. In Section 2 we outline an experiment procedure that can control for data collection inconsistencies, and discuss some alternative approaches in Section 3. We place the procedure in the context of prior work in Section 4, and describe the reference implementation in Section 5. In Section 6 we outline a specific experiment involving 15 different modes, and use it to evaluate the power/accuracy tradeoff results for various virtual sensors, concluding with Section 7.

CONTROLLED EVALUATION OF CONTEXT-SENSITIVE BEHAVIOR

As discussed in Section 1, instruments are typically evaluated by repeatedly exposing them to controlled inputs to determine their error characteristics. In the case of complex systems such as HMSes, the evaluation needs additional controls for feedback loops, cost/accuracy tradeoffs and privacy considerations. In this section, we outline EM-EVAL a procedure for HMS evaluation that addresses these concerns with two techniques that are novel in this domain:

(i) pre-defined, **artificial trips** that support spatial ground truth, preserve privacy, increase the breadth of *trip types* and support repetitions for establishing error bounds, and

(ii) power and accuracy **control phones** carried at the same time as the experimental phones, that can cancel out context-sensitive variations in power and accuracy.

Artificial timeline

The core of the experimental procedure is the pre-defined specification of a sequence of artificial *trips*, potentially with multiple *legs* or *sections* per trip. The trajectory and mode of travel is also pre-defined. The *data collector* completes the timeline trips by strictly following the specified trajectory and mode while carrying multiple phones that collect data simultaneously using different configurations.

The specified pre-defined trajectories provide spatial ground truth. We do not pre-define temporal ground truth since it is extremely hard to control for differences in walking speeds, delays due to traffic conditions, etc. We use manual input from the data collector to collect *coarse* temporal “ground truth" of the transitions along the timeline. We do not use manual input for *fine-*

grained temporal ground truth along the trajectory because: (i) human response times are too slow for fine-grained temporal ground truth during motorized transportation, and (ii) distracting the data collector during active transportation can be risky.

Using an artificial timeline addresses several of the unique challenges associated with HMS evaluation.

Privacy Since the trips are artificial, they preserve the data collector's privacy. Even if his adversaries would download the trips, they would not be able to learn anything about his normal travel patterns.

Spatial ground truth Since even high accuracy (GPS-based) data collection has errors, pre-defining spatial ground truth allows us to resolve discrepancies (Section 3) and compute the true accuracy.

Breadth and variety of trips Artifical trips allow efficient exploration of the breadth of the trip space. For example, the trips could include novel modes such as e-scooters and e-bikes, or specify different contexts for conventional modes, such as express bus versus city bus.

Repetitions Since the trips are pre-defined, they can be repeated exactly. This allows us to use standard variance and outlier detection to estimate error bounds on the measured values.

Control phones

The artificial trips give us spatial ground truth, but they do not give us cost (power consumption) or temporal ground truth.

We control for the cost through the use of the use of multiple phones, carried at the same time by the data collector. The phones carried by the data collector are divided into control phones and experiment phones. The control phones represent the baseline along each of the axes in our tradeoff and the experiment phones implement a custom sensing regime that is at some intermediate point. The evaluation procedure allows us to determine those points.

Power The power control phone captures the baseline power consumption of a phone that is not being used for tracking by a HMS. This does not mean that the phone is idle - phone OSes (e.g. iOS or android) are complex, context-sensitive systems that perform their own location tracking (e.g. "Find my iPhone") and their own duty cycling (Figure 1). Using a power control allows us to identify the **additional power** consumed by the HMS, even if it is context sensitive.

Accuracy The accuracy control captures the upper bound on the accuracy of a particular class of smartphones given sensor and OS limitations. While we would like to compare the experimental accuracy to ground truth, (i) all sensors have errors, so ground truth is not achievable in practice, (ii) artificial trips give us spatial but not temporal ground truth, and (iii) GIS-based trajectory specifications do not have an associated power tradeoff. Using an accuracy control allows us to compare the experimental data collection against the best achievable data collection, in addition to the ground truth.

DISCUSSION OF ALTERNATIVE PROCEDURES

While EM-EVAL (Section 2) addresses the complexities of HMS evaluation, it also imposes a much higher researcher burden than the ad-hoc method. This raises the question of whether all these controls are necessary or merely sufficient. In this section, we discuss some alternative approaches and highlight the unexpected behavior that they would miss. This list is not comprehensive but provides a flavor of the arguments without tedious repetition.

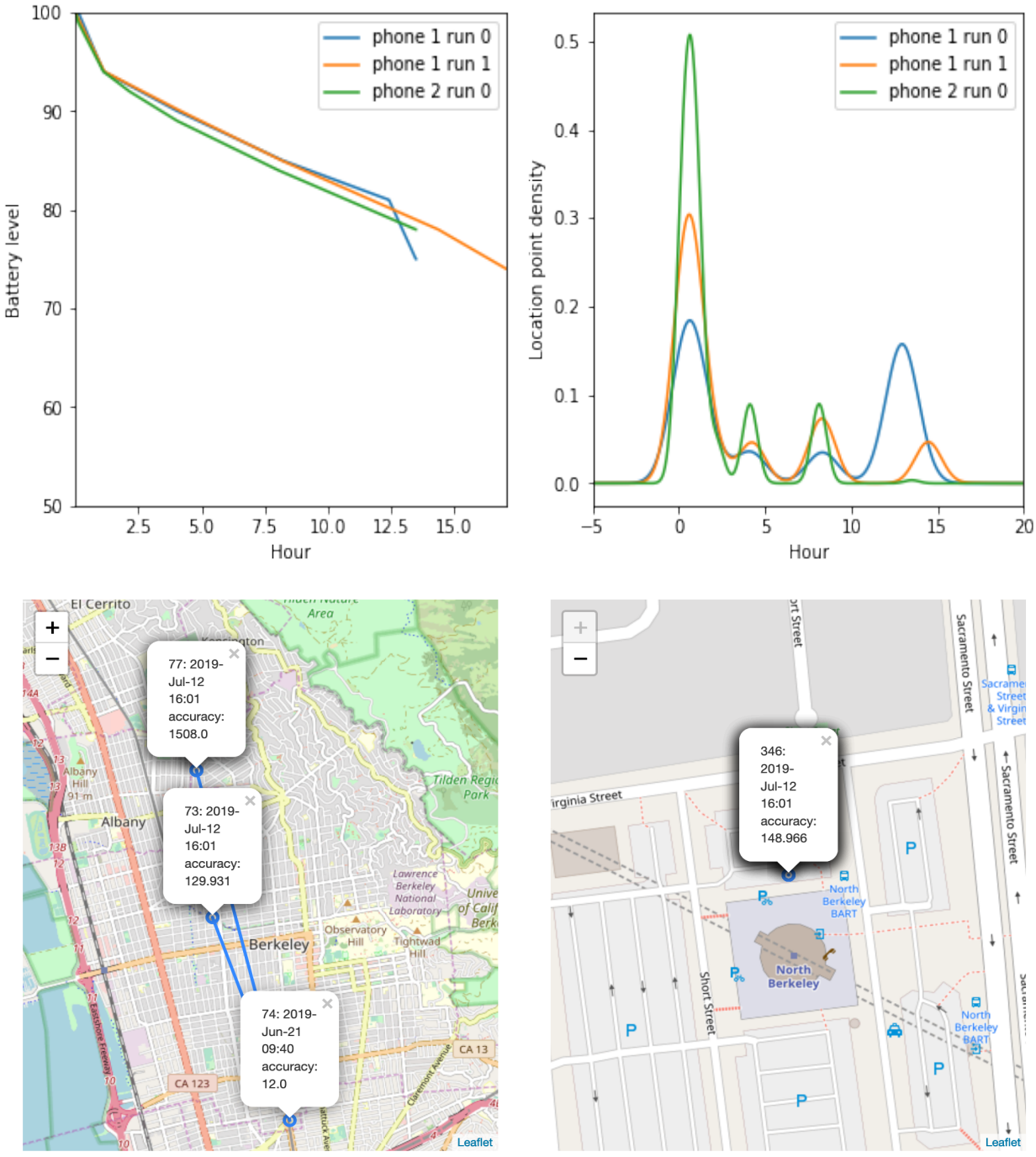

FIGURE 1: **Top:** Power variation illustrated by duty cycling on android. All the phones were configured identically, and placed in the same environment. The built-in duty cycling on android switches all phones to low power mode at around 1 hr. However, phone 1, on run 1 alone, switches back to high power mode at around 12.5 hours. Repeating experiments allows us to distinguish the first consistent duty cycle and the second outlier. **Bottom:** Accuracy variation illusrated by mismatched timestamps during trajectory data collection. Both trajectories are collected from identical phones during a subway trip. Point 74 has an accuracy radius of only 12, but its timestamp is in June instead of July! Spatial ground truth allows us to sort out the varying accuracies here.

No artificial trips
Creating pre-determined trips requires an upfront investment in effort, and requires the data collector to take trips just for data collection. An alternative would use multiple phones, but allow data collectors to go about their regular routines and tag the modes only. We could use the accuracy control phones to determine the ground truth trajectory.

No privacy Capturing the data collector's regular routines compromises their privacy. Even if the data does not include their name or phone number, a list of their commonly visited places and trips can form a unique fingerprint that can uniquely identify them (*2, 19*). This sensitivity precludes evaluation data from being published and used for reproducible research.

No repetition The behavior of the same phone with the same configuration can vary over time, both for power and for accuracy (Figure 1). Repeating the same trip multiple times allows us to detect and remove outliers. With ad-hoc trips, it is unclear whether any difference in behavior is real or caused by context-sensitive variation. And without pre-determined trips, it is challenging to repeat the same trips and trajectories over time.

No spatial ground truth No sensor is perfect and even the accuracy control phones can have sensing errors. If we see a divergence between an experiment phone and the accuracy control, it is unclear which one has the error (Figure 1).

No control
Using control phones requires the researcher to purchase multiple phones of the same make, model and approximate age. While used smartphones are relatively cheap (USD 50 - USD 100), 4 android phone and 4 iPhones combined will still cost USD 400 - USD 800. An alternative would be to use one phone each for each OS, perform the timeline trips, and look at the app-based power consumption reported by the phone OS?

Sensor access attribution and the meaning of the % Sensor access in modern phone OSes (android and iOS) is also context-sensitive, making it unclear how it is counted for per-app consumption. For example, if multiple apps request a sensor reading, the OS delays returning a result until it can batch related requests and serve all of them with a single sensor access. This is why the OSes treat the sensing frequency as a hint instead of a guarantee. Second, if sensor access is mediated by a service (e.g. fused location in Google Play Services), it is unclear whether the sensor access is counted for the service or the app (Figure 2). And finally, although android reports per app consumption as a % of the battery *capacity*, iOS does so as a % of the battery *consumption*. This indicates that on dedicated phones, the HMS under test will always show close to 100%, whether it is the power control or the accuracy control (Figure 2). Using a control phone for the power will cancel out these context sensitive effects and estimate the difference in power drain with and without the HMS app component installed.

Custom duty cycling increases power drain Sensing is not the only source of power consumption - CPU usage can also have a significant impact on power usage. HMSes can use smart local processing to reduce local sensing, but the increased power consumption from the CPU can cancel out the savings from the sensing. Including an accuracy control showed that the basic duty cycling algorithm in our experiment paid for itself in low frequency sensing but actually increased power usage for high frequency sensing (Figure 2).

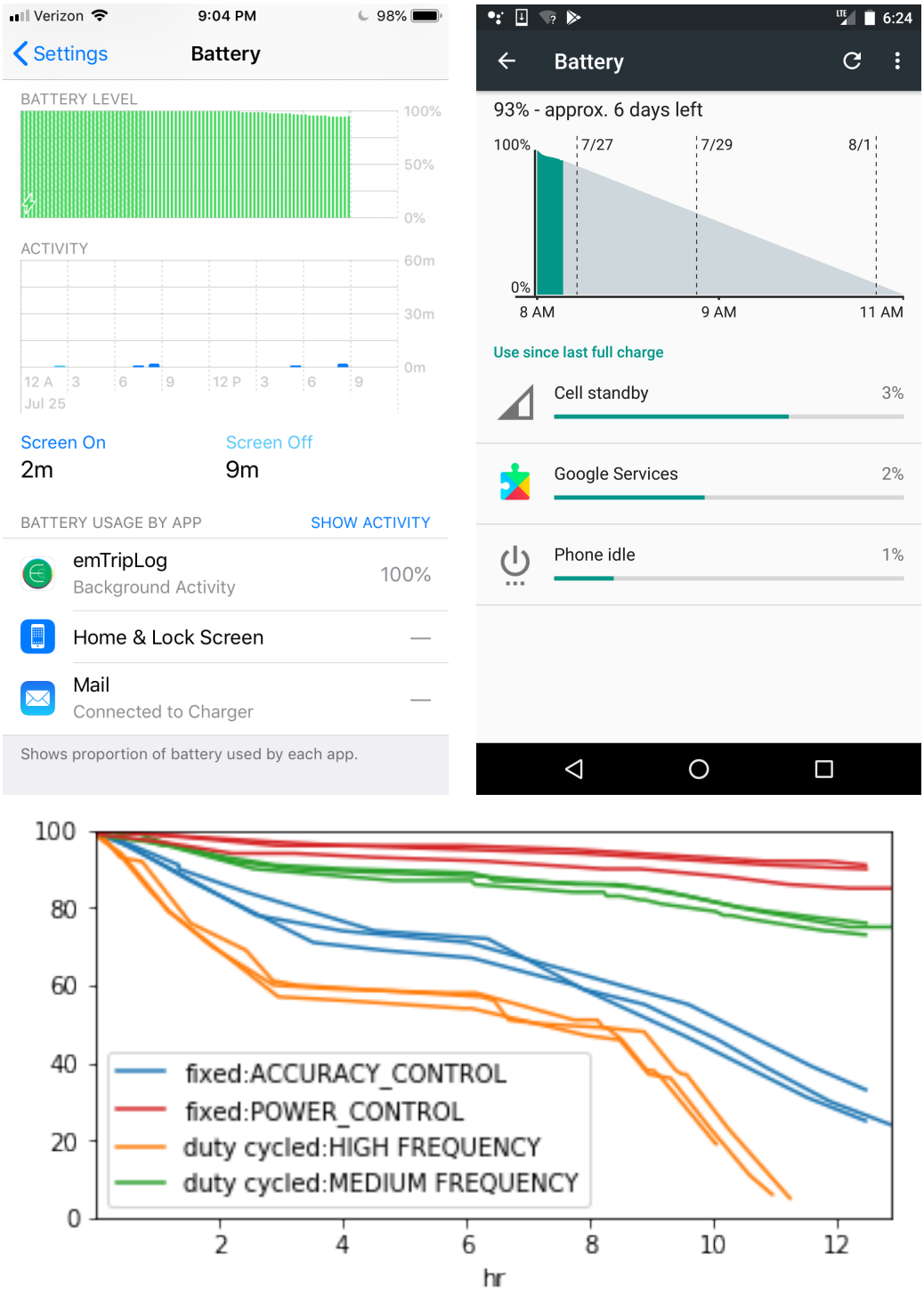

FIGURE 2: **Top left**: iOS power control phone with the sensing app consuming 100% but of a power drain of only 2% over the entire day. **Top right**: android phone showing Google Play services as a separate power consumer.**Bottom**: Explicit duty cycling causes increased power drain at high frequencies, possibly due to greater CPU power consumption. Note that the battery drain flattens out on all curves during the middle, stationary part. The main difference is in the rate of power drain while moving: low frequency sensing consumes the least power, and the checks for android's built-in duty cycling appear to be optimized to be more efficient than our simple implementation.

RELATED WORK

Human Mobility Systems (HMSes) are complex to evaluate (Section 3). Other researchers have identified similar challenges as part of survey papers (e.g. phone context, privacy, learning, scaling (*9*), varying metrics and time scales across research areas (*13*)). However, to the best of our knowledge, there is no proposed solution that addresses all of them.

Papers related to instrumenting travel behavior fall into three main categories; we list some work from each as an example. A comprehensive classification of papers into categories is beyond the scope of this paper.

Context sensitive sensing algorithms: power without accuracy

This research area focuses on context sensitive, adaptive power management of sensors. Papers such as ACE (*11*) and Jigsaw (*10*) compare their power requirements to naive sensing techniques. However, their accuracy evaluations focus on the localization error (Jigsaw), or comparison to naive inferred results[1] (ACE).

Travel diary systems: compare to manual surveys

There is a vast variety of one-off travel diary systems that combine smartphone based sensing with cloud-based processing to generate travel diaries. Systems such as such as Data Mobile (*12*), Future Mobility Study (FMS) (*1, 6*) and rMove (*5*) aim to replace the paper and telephone based Household Travel Surveys with smartphone and cloud based systems. So they evaluate the accuracy of their systems against the traditional methods, not against ground truth. This can show that smartphone based methods are significantly better than traditional methods, but not provide a quantitative estimate of the accuracy of their system. Similarly, they do not include quantitative power evaluations - preferring statements like "Among the three types of discrepancies, the second type, data gap due to battery drainage, was most frequently observed." (*6*) or "The battery consumption test was simply whether, under regular usage, the phone could make it through the day without having to be charged." (*12*). So they do not rigorously evaluate either the power or the accuracy side of the tradeoff.

Mode inference: accuracy without power, non-uniform data

Mode inference of travel mode based on sensor data is an extremely popular subject in the literature[2]. Researchers have used decision trees (*14, 20*), Hidden Markov Models (*15, 21*), and neural networks (*3, 4*) to distinguish between various subsets of travel modes. However, although the inference algorithms are different, most such papers use similar methods for evaluating their accuracy. They typically recruit a small sample of their friends (e.g. 16 users over one day (*14*), 4 users over two weeks (*21*)) to collect naturalistic data along with annotations of the ground truth. The data collection focuses on the sensors used for analysis and omits the battery. This kind of evaluation does not meet the any of the requirements outlined above, except privacy, which is addressed by not publishing the dataset.

EVALUATION SYSTEM AND EXPERIMENT DESIGN

The EM-EVAL procedure (Section 2) allows us to estimate the power/accuracy tradeoff of various sensing configurations used in Human Mobility Systems (HMSes). One of the novel components

[1] e.g. based on speed

[2] Probably because it is hard, and nobody has really solved it well yet for modes other than walking

of the procedure involves the specification of pre-defined, artificial trips with ground truthed trajectories and modes.

This section explores the nuances of implementing such a procedure. We first describe a publicly available reference implementation of a system - EM-EVAL-ZEPHYR - that can be used perform this procedure. We then discuss challenges encountered while using the system to perform an experiment in the San Francisco Bay Area (Section 6). Some of these challenges were addressed by system improvements, while others are documented as best practices for future data collectors.

System overview

EM-EVAL is a generic *procedure* for HMS evaluation - it does not actually collect any data. To use it, we need a concrete system that configures data collection based on the spec configurations, collects coarse temporal ground truth, periodically reads battery levels and stores data for future analysis.

As part of our evaluation (Section 6), we built a system EM-EVAL-ZEPHYR that combines our prior work on power evaluations (*16*) with our existing HMS platform (*8*) and supports performing the EM-EVAL procedure. The system consists of three main parts:

Evaluation Specification The *spec* describes an evaluation that has been performed or will be performed in the future. In addition to mode and trajectory ground truth, it includes the app configurations to be compared and the mapping from phones to evaluation *roles*. The spec automatically configures both the data collection app and the standard analysis modules.

To reduce evaluator burden, we provide pre-processing functions to fill in trajectory information based on route waypoints for road trips and OSM relations for public transit. We also provide sample notebooks to verify timelines and their components before finalizing and uploading the evaluation spec.

Auto-configured Smartphone App We have generated a custom UI skin for our E-MISSION platform (*8*) that is focused on evaluation. It allows evaluators to select the current spec from the public datastore, and automatically downloads the potential comparisons to be evaluated, the role mappings and the timeline.

Since the e-mission platform data collection settings are configurable through the UI, the sensing configurations defined in the spec are automatically applied based on the phone role when the data collector starts an experiment. For example, when starting an experiment to compare high accuracy (HAHFDC) versus medium accuracy (MAHFDC) data collection, the second experiment phone will automatically be set to MAHFDC settings. Finally, when the data collector performs the trips, he marks the transition ground truth in the UI, and the app automatically displays the next step in the timeline (Figure 3).

Public Data + Sample Access Modules Since there are no privacy constraints, EM-EVAL-ZEPHYR uploads all collected data to a public instance of the E-MISSION server. The associated repository contains sample notebooks that can download, visualize and evaluate the data associated with a particular spec. All the data used in this paper is publicly available, and the notebooks can be manipulated interactively using binder.

Note that although the EM-EVAL **procedure** is general, the current implementation of the EM-EVAL-ZEPHYR **system** is integrated only with the E-MISSION platform. Using the procedure with other HMSes will require re-implementing the EM-EVAL procedure with the other HMS, or using a combination of systems for the evaluation. For example, EM-EVAL-ZEPHYR can still read the battery level periodically, display the trip sequence to the data collector, and be used to mark

FIGURE 3: **Top**: Spec components in EM-EVAL-ZEPHYR include configuration, timeline and trip details. **Bottom**: Sample spec for a multi-modal trip, including transfers and waits for public transit.

the transition ground truths. However, the evaluator needs to configure the settings for the app being tested manually, and to download, clean and analyse the resulting data.

System iterations and lessons learned

As we started collecting data, we had to resolve some ambiguities around exactly when the transition ground truth should be collected. We also discovered best practices that increased the likelihood of successful data collection. This section outlines these lessons learned.

System change: capture transition complexity One of the big promises of using HMSes for instrumenting human travel is that we don't have to focus only on the primary mode. Instead, with fine-grained data collection, we can understand the full complexity of end to end travel.

In fact, the only true unimodal trips are walking trips. Everything else is multi-modal. Thus, a significant change to the system was to restore the hidden complexity that is elided from user descriptions of travel diaries. For example, consider the trip description "Drive from Mountain View Library to Los Altos Library". Although that appears to be a unimodal trip, it is actually a multi-modal trip which involves implicit walk access sections to and from the car at the source and destination respectively.

It is not possible to pre-determine the ground truth for these walk access sections since we cannot control which parking spaces are available when we perform the trip. We address such issues by adding *shim sections*, and expanding the start and end from points to 100m polygons. We can then relax the constraints around ground truth within the polygon by only using the reference dataset, but still check the accuracy of the mode inference (Figure 3).

Best practice: Pilots are critical In spite of reviewing the pre-determined trajectories ahead of time as part of the validation process, and also having them displayed on the EM-EVAL-ZEPHYR UI, we found that we frequently made small mistakes, during the first round of data collection for a new timeline. Sometimes, we found that the route suggested by OSRM felt unsafe to bicycle on and we had to accordingly change our specification. The second repetition generally resolved these issues. In order to avoid a stressful data collection experience, we suggest running through a new timeline with a trial run before starting full-featured data collection.

Best practice: Mindfulness Remembering to mark the transition ground truths was one of the hardest parts of the ongoing data collection and really highlights the challenges of ground truth collection. In spite of the fact that she was performing artifical trips to collect data for her own project, one of the authors forgot to mark wait -> move transitions during the pilot for the long multi-modal timeline because she had started checking her email while waiting. It is important to be present in the moment and pay attention to the context while collecting data.

EVALUATION

Experiment design

EM-EVAL is a generic evaluation procedure and can be used with any kind of HMS. EM-EVAL-ZEPHYR is a reference implementation of EM-EVAL that used to evaluate many experimental settings relevant to HMSes over any set of trajectories. The evaluator can pick her settings based on her research goals. In this section, we outline our goals for this evaluation, and use them to define three timelines that cover 15 separate modes, including recently popular modes such as escooter and ebike.

Dwell time Instead of focusing only on trips, we wanted to evaluate a timeline that included significant dwell time. We could see from our calibration runs that android appears to have

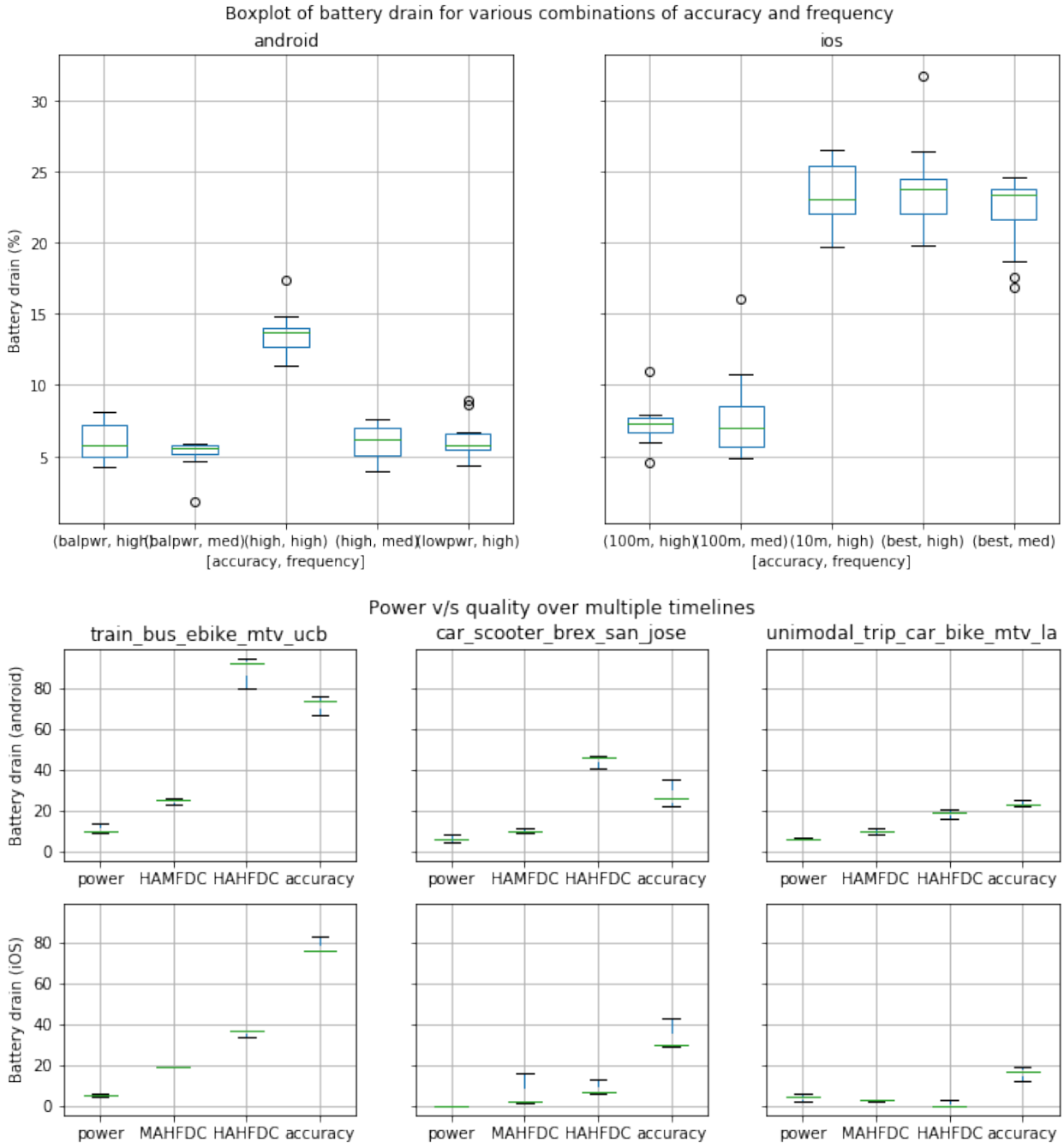

FIGURE 4: Power calibration and evaluation. **Calibration**: All phones were stationary, configured identically and had the app-based sensing turned always on. Although there are outliers, repeated experiments allow us to see clear behavior differences. High accuracy, high frequency (HAHF) is the only high power configuration on android, lowering either the accuracy or the frequency leads to essentially the same power drain. On iOS, the frequency does not matter at all, the only way to reduce the power drain is to lower the accuracy. **Evaluation**: Based on these calibration results, our experiment design compares high v/s med frequency on android and high v/s low accuracy on iOS. Given the calibration results, we would expect the drain to increase with quality, and we do see that behavior on the longest timeline for iOS and the shortest timeline for android. On shorter timelines for iOS, the differences are small and fall within our current error bounds. On longer timelines for android, the CPU drain of our basic duty cycling overwhelms any improvements from duty cycling, so the HAHFDC performs worse than the accuracy control.

id	Description	Outgoing trip modes	Incoming trip modes	Dwell time	Overall time
`unimodal trip car bike mtv la`	Suburban round trip	car (city streets)	bike	1 hrs	3 hrs
`car scooter brex san jose`	Downtown library	car (freeway)	escooter Bus Rapid Transit	3 hrs	5.5 hrs
`train bus ebike mtv ucb`	Multi-modal trip across the bay	commuter train subway city bus	ebike (shared) express bus downtown walk light rail commuter rail	6 hrs	12.5 hrs

TABLE 1: Brief description of timelines, covered modes, dwell times and overall times

built-in duty cycling and including significant dwell time would allow us to capture the impact of this context sensitive behavior. Therefore, we structured our timeline trips as round trips to libraries with an intermediate dwell time $\approx 3\times$ the mean travel time to the location.

Broad range of modes HMS evaluations should cover a broad spectrum of trip types, and since we are creating artificial trips, we can structure them to maximize mode variety. In order to efficiently cover this space, we tried to ensure that no mode was repeated. We only had to include commuter rail twice since there were few other transit options to reach the starting point chosen.

Multi-modal transfers Detecting multi-modal transfers in a HMS is tricky because there isn't a clear signal similar to a trip end. We ensure that there are many transition examples by emphasizing multi-modal transfers.

With those goals in mind, we decided on three artificial timelines of varying lengths that cover a total of 15 separate modes. We chose each timeline to be round trips to libraries so as to not include identifiable location data (e.g. home location) in our experiments. A description of each timeline with the associated modes and dwell times is given in Table 1.

Since this paper focuses on the evaluation procedure, to avoid bias, we do not use it to evaluate any particular HMS. Instead, we use it to evaluate the virtual sensors provided by smartphones themselves. These virtual sensors are exposed by smartphone operating systems (OSes) by combining underlying physical sensors using proprietary algorithms. iOS already restricts developer access to the GPS, instead supporting a virtual location sensor that chooses underlying sources based on developer-supplied accuracy constants. As phone OSes impose greater background restrictions on apps, we can only expect the use of such sensors to grow. In this section, we analyze spatial accuracy and motion activity accuracy from our collected data, and draw inferences from them to demonstrate the benefits that can be gained by using a procedure such as EM-EVAL.

Spatial accuracy

Box plots of the distribution of spatial error of the measured locations against the groundtruth in the different timelines are shown on the top row of Figure 5. It is interesting to note the skewed nature of the plots, and the presence of different outliers, whose values seem to vary across phones across timelines. Ad-hoc evaluation schemes would probably identify them by comparing error values against a threshold. Choosing this threshold without the knowledge of the distribution of error values would be difficult and introduce ambiguity to the evaluation. Our statistical and systematic approach eliminates these difficulties.

Motion activity accuracy

The distribution of temporal error between the ground truth section transitions and sensed activity transitions is shown in the bottom row of Figure 5. The plot shows that despite the presence of outliers, the quality of sensing does not seem to impact any difference in accuracy. Thus, a study focusing on places and modes but not trajectories could choose a lower quality sensing configuration and reduce power drain.

CONCLUSION AND CALL TO ACTION

Human Mobility Systems (HMSes) are complex software systems that run on equally complex smartphone operating systems (OSes). This complexity implies that there is rarely a simple linear relationship between their inputs and outputs, which complicates their evaluation.

We outline a procedure, based on repeated travel over pre-defined artificial *timelines* carrying experiment and control phones, to control this complexity. We show that it can control for outliers, and also reveal meaningful signals about the behavior of smartphone virtual sensors that are relevant to instrumenting human travel data.

The procedure is privacy-preserving, so it does not need human subjects approval. It focuses on trip diversity, not demographic diversity, so it can be undertaken by a small research group, or even a single researcher, as a pre-pilot before recruiting study participants. It uses predetermined trips and modes, so it can efficiently explore complex or newly emerging travel patterns and modes, such as escooters. The procedure, and the associated reference implementation can simplify the testing required before a study is launched.

The procedure is not a replacement for a pilot, since it does not capture user experience or other behavioral factors. However, it can shorten and simplify the pilot, since the pilot does not need evaluate quantitative “hard" factors. As a general principle, this controlled procedure evaluates computational features, such as power drain or accuracy, while the pilot assesses “soft" features such as usability and motivation.

The procedure primarily supports direct comparisons between configurations or systems for the tested contexts. Based on the procedure, we may be able to model individual features of the HMS or the underlying system (e.g. the rate of power drain while moving). However, since both hardware specifications and phone OS implementations can and do change[3], researchers will need to periodically re-tune their models, which will entail re-running the procedure.

Finally, we have used the procedure to evaluate metrics related to sensing accuracy under various configurations. Post-processing algorithms can potentially compensate for low sensing accuracy through integration with other sources, such as GIS systems. Large-scale datasets, such as

[3]battery optimization android neural network

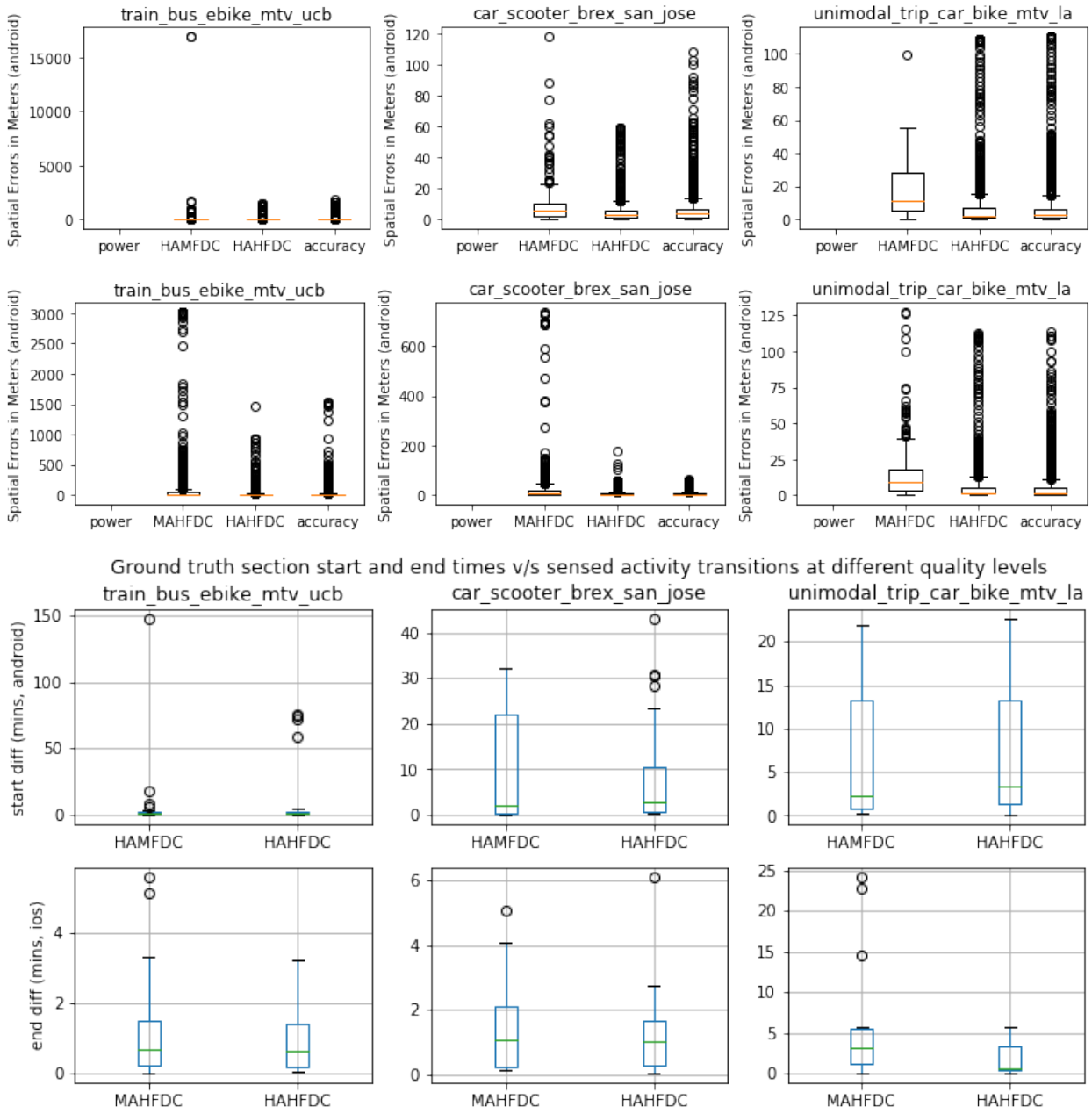

FIGURE 5: Spatio-temporal accuracy evaluations. Top: Distribution of the *spatial* error across the timelines. The medium sensing accuracy is suprisingly close to the high accuracy, although it has many more outliers, specially in the timeline with the underground segments. **Bottom**: Distribution of the *temporal* error between the ground truth section transitions and sensed activity transitions. There is essentially no difference between high and low quality sensing, even for the longest timelines. The vast majority of the transitions occur within 5 mins, although there are significant outliers.

ImageNet, and the resulting toolchains, have greatly accelerated the development of algorithms in other domains. If HMS reseachers adopt this procedure and publish resulting data, the aggregation over multiple evalutions can generate a large scale, ground truthed, public dataset. This dataset can accelerate the development of analysis algorithms, such as mode and purpose inference, for HMSes.

REFERENCES

1. C. D. Cottrill, F. C. Pereira, F. Zhao, I. F. Dias, H. B. Lim, M. Ben-Akiva, and P. C. Zegras. The Future Mobility Survey: Experiences in developing a smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, (2354):59–67, 2013.
2. Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.
3. Fei Yang, Zhenxing Yao, and Peter Jin. Multi-mode Trip Information Recognition Based on Wavelet Transform Modulus Algorithm by Using GPS and Acceleration Data. In *TRB 94th Annual Meeting Compendium of Papers*, Washington, DC, Jan. 2015.
4. P. Gonzalez, J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi, and R. Perez. Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. *IET Intelligent Transport Systems*, 4(1):37, 2010.
5. E. Greene, L. Flake, K. Hathaway, and M. Geilich. A seven-day smartphone-based gps household travel survey in indiana. Washington, D.C, Jan. 2016. Transportation Research Board.
6. Hongik University, J. S. Lee, P. C. Zegras, F. Zhao, D. Kim, and J. Kang. Testing the Reliability of a Smartphone-Based Travel Survey: An Experiment in Seoul. *The Journal of The Korea Institute of Intelligent Transport Systems*, 15(2):50–62, Apr. 2016.
7. J. Jariyasunant, M. Abou-Zeid, A. Carrel, V. Ekambaram, D. Gaker, R. Sengupta, and J. L. Walker. Quantified Traveler: Travel Feedback Meets the Cloud to Change Behavior. *Journal of Intelligent Transportation Systems*, 19(2):109–124, Apr. 2015.
8. K. Shankari, Mohamed Amine Bouzaghrane, Samuel M. Maurer, Paul Waddell, David E. Culler, and Randy H. Katz. E-mission: An open-source smartphone platform for collecting human travel data. *Transportation Research Record: Journal of the Transportation Research Board*, 2018.
9. N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.
10. H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell. The jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 71–84, 2010.
11. S. Nath. ACE: exploiting correlation for energy-efficient and continuous context sensing. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 29–42. ACM, 2012.
12. Z. Patterson and K. Fitzsimmons. Datamobile. *Transportation Research Record: Journal of the Transportation Research Board*, 2594:35–43, 2016.
13. A. C. Prelipcean, G. Gidófalvi, and Y. O. Susilo. Transportation mode detection – an in-depth review of applicability and reliability. *Transport Reviews*, 37(4):442–464, July 2017.
14. S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks*, 6(2):1–27, Feb. 2010.
15. R. C. Shah, C.-y. Wan, H. Lu, and L. Nachman. Classifying the mode of transportation on mobile phones using GIS information. pages 225–229. ACM Press, 2014.

16. K. Shankari, J. Fürst, Y. Wang, P. Bonnet, D. E. Culler, and R. H. Katz. Zephyr: Simple, ready-to-use software-based power evaluation for background sensing smartphone applications. Technical Report UCB/EECS-2018-168, EECS Department, University of California, Berkeley, Dec 2018.
17. K. Shankari, M. Yin, D. Culler, and R. H. Katz. E-Mission: Automated transportation emission calculation using smartphones. In *Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 268–271, Mar. 2015.
18. A. Vij and K. Shankari. When is big data big enough? Implications of using GPS-based surveys for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, 56:446–462, July 2015.
19. H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2011.
20. Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web*, 4(1):1–36, Jan. 2010.
21. M. Zhong, J. Wen, P. Hu, and J. Indulska. Advancing Android activity recognition service with Markov smoother. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on*, pages 38–43. IEEE, 2015.