Using conferences to create open, ground truthed easily accessible travel datasets: a simple proposal

K. Shankari Paul Waddell David Culler Randy Katz *

November 17, 2017

Abstract

There is a severe lack of open, ground truthed travel datasets that can be used to develop travel behavior analysis software tools. Most existing open datasets do not include ground truth for travel behavior. Projects that relate more directly to travel behavior do not make their datasets open because of privacy considerations.

Travel-related conferences can be used to collect ongoing, open, privacy preserving, ground truthed travel data. Since the travel patterns of attendees are conference-specific, publishing them will not leak information about their regular habits. Involving transportation researchers in the data collection will also ensure that the ground truth evolves to be relevant to new travel options and upcoming research areas.

Creating a long-running, constantly updated travel dataset can lower barriers related to reproducibility and reuse in the travel space. We propose a pilot of this approach at the Transportation Research Board (TRB) Annual Meeting in January 2018.

Although researchers in both the transportation and computing communities have proposed novel techniques for extracting meaning from location sensed data, there is a lack of open datasets that can be used to build on these techniques.

The primary challenges for generating such datasets include privacy and relevance of ground truth. Existing open datasets that include location information such as the Nokia Data Challenge [LGPA+12] and the Reality Mining [EP05] datasets do not include ground truth for travel behavior. Projects that relate more directly to travel behavior such as mode inference [ZWHI15, ZA15, FZP15] typically use small friends and family datasets (n=4,35,20) and do not publish them due to privacy concerns.

The primary privacy concern is re-identification through unique travel patterns, which can then be extrapolated and used for malicious or embarassing ends. For example, given precise, fine-grained travel history, it is possible to determine home and work locations. If there is only one possible resident with that home/work pair, a malicious observer can then determine that she stops by an ice cream shop every day at 5pm, and shame her about her eating habits. Even if trip start and end locations are fuzzed to increase privacy, 95% of the population can be uniquely identified by no more than 6 coarse locations [dMHVB13].

Collecting data at conferences removes this attack vector. Since conference attendees stay in hostels and work at the conference, their real home and work locations cannot be determined from conference travel history. Any travel patterns that are observed at the conference will not generalize to the attendees' regular patterns at home. At the same time, as attendees, specially at large conferences, travel to the conference venue, socialize after hours and sightsee in the days before or after, they can generate rich and varied travel data with ground truth.

Ground truth collection in the context of travel behavior comes with its own challenges. As travel options expand and novel travel analyses are proposed, new forms of ground truth need to be collected. For example, the Geolife open dataset [ZXM10] contains ground truth for 4 travel modes (walk/driving/bus/bike) [ZCL+10],

*{shankari, waddell, culler, randykatz}@berkeley.edu

but it does not include trip purpose, or newer modes such as ride hailing services. The utility of the increased ground truth collection also needs to be balanced with the increased burden on travelers.

Using transportation researchers to provide ground truth can mitigate several of these challenges. During the data collection planning process, researchers can request the inclusion of new options or the collection of new types of data. This will ensure that the data collected is relevant to the current state of the art. The rate at which participants contribute ground truth is an upper bound on the rate in the wild - if the ground truth burden is too large for transportation researchers, it is certainly too large for the general public.

We propose the use of the e-mission platform[1] for collecting and publishing this data. The platform supports a configurable user interface, end of trip surveys and targeted surveys based on travel history. Both surveys can either be custom, or linked to third party survey tools such as qualtrics. The raw data is stored as an immutable time series that analysis tools can ingest to generate output results. This aids reproducibility and comparison of various implementation of the same analysis.

We envision that for every conference that agrees to participate, a group of transportation researchers would determine the data collection parameters for that conference. They would then apply for IRB approval, customize the data collection app, and publish a custom version of the app for the conference. As part of the conference logistical information, attendees would be provided a link to the app, and encouraged to install it upon arrival at the city where the conference is held. They would be instructed to uninstall it when leaving the conference. The conference-specific version of the app would connect to a special server that would make all stored data available publicly. Participating researchers could analyse the data in real-time and use targeted surveys to request additional information from attendees in a timely fashion. An initial exploration of this data collection method can be seen at https://github.com/interscity/open-data-challenge-2017

The resulting datasets can be accumulated on the public data server, enabling consistent, longitudinal analysis of travel data. A public ipython notebook server, provisioned with analysis and data access libraries, and a read-only connection to the database, will enable easy analysis without significant setup. The collection of notebooks can also act as an analysis repository and aid reproducibility. The server can be initially provisioned using credits allocated by cloud infrastructure providers for the publication of open data. As researchers use the datasets and we can empirically derive data usage patterns, we can decide whether to switch to a temporary server for the duration of the conference, and long-term storage using a github repository.

We propose a pilot of this approach at the Transportation Research Board (TRB) Annual Meeting in January 2018. As a proof of concept, once the data collection design is complete, the e-mission team will file the IRB protocol and perform any customization required. The Travel Survey Methods committee can determine the ground truth to collect and coordinate with the organizers on publicity and recruitment. The 2018 TRB meeting includes 5000 presentations, so the potential pool is at least 5000 people. If even 10% of them are interested in travel surveys, or are willing to help out the travel survey team, we have the potential to get a dataset from 500 people over 7 days (3500 person days).

This pilot can establish processes for this data collection method for use in future conferences. If successful, the availability of a long-running, constantly updated travel dataset can lower barriers related to reproducibility and reuse, and hopefully, accelerate the development of the field.

[1] https://e-mission.eecs.berkeley.edu

References

[dMHVB13] Yves-Alexandre de Montjoye, Csar A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013.

[EP05] Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, November 2005.

[FZP15] 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, January 2015.

[LGPA+12] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, and others. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, 2012.

[ZA15] Zahra Ansari Lari and Amir Golroo. Automated Transportation Mode Detection Using Smart Phone Applications via Machine Learning: Case Study Mega City of Tehran. 2015.

[ZCL+10] Yu Zheng, Yukun Chen, Quannan Li, Xing Xie, and Wei-Ying Ma. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web*, 4(1):1–36, January 2010.

[ZWHI15] Mingyang Zhong, Jiahui Wen, Peizhao Hu, and Jadwiga 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.

[ZXM10] Yu Zheng, Xing Xie, and Wei-Ying Ma. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.