COMPARATIVE ANALYSIS OF BIG AND SMALL (SURVEY) DATA FOR DERIVING HUMAN MOBILITY PATTERNS

Conference Paper · August 2018

CITATIONS 0

READS 51

4 authors, including:

Feilong Wang
University of Washington Seattle
11 PUBLICATIONS 29 CITATIONS

Xuegang (Jeff) Ban
University of Washington Seattle
136 PUBLICATIONS 2,429 CITATIONS

Some of the authors of this publication are also working on these related projects:

- Project
  Understanding transportation big data View project

- Project
  National Natural Science Foundation of China View project

All content following this page was uploaded by Feilong Wang on 17 January 2019.

The user has requested enhancement of the downloaded file.
COMPARATIVE ANALYSIS OF BIG AND SMALL (SURVEY) DATA FOR DERIVING HUMAN MOBILITY PATTERNS

Jingxing Wang
Department of Civil and Environmental Engineering
University of Washington
Box 352700
Seattle, WA 98195
Email: wangjx@uw.edu

Feilong Wang
Department of Civil and Environmental Engineering
University of Washington
Box 352700
Seattle, WA 98195
Email: flwang@uw.edu

Xuegang J. Ban
Department of Civil and Environmental Engineering
University of Washington
Box 352700
Seattle, WA 98195
Email: banx@uw.edu

Cynthia Chen
Corresponding Author
Department of Civil and Environmental Engineering
University of Washington
Box 352700
Seattle, WA 98195
Email: qzchen@uw.edu

Word count: 5571 words text + 2 tables = 6071 words

Prepared for Presentation at
98th Annual Meeting of Transportation Research Board
Submission Date: 8/1/2018
ABSTRACT
The next-generation household travel surveys, the core data generation mechanism for supporting both short- and long-term transportation planning applications, are poised to be transformed. It is now increasingly recognized that passively-solicited big data, or large amount of data generated through various types of subscription services, will play an important role in this transformation. Passively-solicited data in its various forms (e.g., mobile sightings, app-based data) not only differ substantially from the household travel survey data, but also among themselves. We argue that the very first step for the passively-solicited data to be integrated into the next-generation household travel surveys is to understand their differences. This paper proposes a three-order analysis framework to analyze these differences. Two case studies each involving a big dataset and a corresponding survey dataset are analyzed to demonstrate their respective properties. The analysis results confirm many distinct properties of such big data as compared between themselves and to the survey data.

Keywords: human mobility pattern, app-based data, mobile phone data, household travel survey

1. INTRODUCTION
Understanding and modeling human mobility patterns, i.e., how individuals and people move in time and space, plays a critical role in many applications including traffic monitoring and forecasting [17], Origin-Destination (OD) flows estimation [43][26][2], event management [3], and even the spread of biological viruses [9]. The traditional human mobility research relies on household travel surveys that log trips made during a 24-hour period for every sampled respondent. It is well-known that traditional household travel surveys have a number of limitations [7] including declining sample sizes [29]; increasing non-response rates [36]; non-representative samples [20][29]; missing activities and trips [22][37]; and imprecise travel time [30].

In the last ten years, the rise of Information and Communication Technology (ICT) has resulted massive amount of temporally- and spatially-resolved data (i.e. cellular data, GPS data) that can be used to derive human mobility patterns and such data is termed as passively-solicited. Passively-solicited data (big data) is generated for purposes that are not intended but can be potentially used for research [8]. It has a number of advantages over traditional household travel surveys including vastly reduced cost, its bigness (massive sample size), and the capability to capture both short trips and long trips [8]. The passively-solicited data has been used diversely in the past decade including understanding mobility patterns [10][11][27][6][18][31][23][44][45], OD estimation [2][16], travel mode detection [24][35], and traffic state monitoring and estimation [38][1][12][13][14][32].

Unlike the household travel survey data that is generated through a controlled, probabilistic sampling process, the passively-solicited data results from a process that is non-probabilistic and its associated properties highly depend on the underlying data generating mechanism that remains largely unknown to researchers [4]. It is non-probabilistic because users in the dataset are not probabilistically selected but self-select to subscribe to a certain service (e.g., phone service or an app on the phone). Consequently, it is unlikely that the dataset is representative of the underlying population, even though it is big. More importantly, its properties such as spatial accuracy and its ability to capture one’s mobility patterns depend on a number of factors such as usage patterns,
signaling channels (e.g., cellular triangulation vs. GPS), and operational characteristics that vary with providers.

And yet, it is crucial to understand the properties of the passively-solicited data that are generated through different underlying mechanisms for at least two important reasons. First, it is conceivable that the next generation household travel survey will no longer be business as usual. It is now almost certain to envision that it will be a hybrid between surveys and passively-solicited data to leverage advantages of both. The integration of surveys and passively-solicited data requires a thorough understanding of the latter, particularly that its underlying data generation process is unknown. Second, it is also safe to say that there will be multiple data providers each with its pros and cons when their data is concerned. In fact, this is already the case for today with data providers such as Inrix, Streetlights, AirSage and others. Thus, comparative understanding of these different datasets is useful as each may be more or less useful for different application needs. Furthermore, a future fusion effort that aims to leverage their pros together will also require a full understanding of their respective properties.

The paper is organized as follows. Section 2 introduces the two case studies and the associated four datasets that were used in this study. Section 3 provides our three-order analysis framework, followed by our comparative results in Section 4. We conclude the paper and discuss the relevant implications in Section 5.

2. EMPIRICAL DATA
In this study, we analyze two case studies each involving a big dataset and a survey dataset in a particular geography. The two big datasets are of different kinds: app-based data and mobile phone sightings data.

The app-based data result from the usage of various apps on smartphones. In general, the spatial and temporal properties of the app-based data depend on five components, including the types of mobile devices, communication networks, positioning (e.g., GPS vs cellular triangulation), service and application providers, and data & content providers [28]. Since those critical components vary greatly, the resulting app-based datasets generally have mixed temporal- and spatial-features. The app-based dataset used in this study was collected from April to June 2017 in the Puget Sound Region (Washington State, US). Spanning 63 days, it comprises 462,401 unique device IDs and contains 66,793,915 observations. Each observation (data record) contains six attributes, including ID, Unix time, Latitude, Longitude, Accuracy, and Time zone.

The other big dataset is the mobile phone data covering 933,508 users in the Buffalo metropolitan area during the month of April 2014. The observations are resulted when users make phone calls or send text messages. It is unclear how frequent the observations are logged, as it depends on the phone operator as well as the data provider that processes the data from the phone operator. Each observation in this dataset contains information including mobile ID, Unix time, and a triangulated location estimation.

To correspond with the two big data sets, we also collected two household survey datasets for the Seattle and Buffalo regions. They are the 2017 Puget Sound Region Household Travel Survey and
Wang, Wang, Ban and Chen

Table 1. Summary of datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Study period</th>
<th>Study area</th>
<th>Number of samples</th>
<th>Sampling rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>App-based data</td>
<td>April 4th-June 5th 2017</td>
<td>Puget Sound Region (Washington State, US)</td>
<td>462,401 unique IDs</td>
<td>11.9% (# of people sampled divided by the population, 2015 PSRC)</td>
</tr>
<tr>
<td>Mobile phone data</td>
<td>April 2014</td>
<td>Buffalo-Niagara Region (New York State, US)</td>
<td>933,508 users</td>
<td>82.2% (# of people sampled divided by the population, 2010 Census)</td>
</tr>
<tr>
<td>2017 PSRC household travel survey data</td>
<td>April-June 2017</td>
<td>Puget Sound Region (Washington State, US)</td>
<td>3,277 households; 6,235 persons</td>
<td>0.16% (# of people sampled divided by the population, 2017 PSRC)</td>
</tr>
<tr>
<td>2002 Buffalo-Niagara regional transportation survey data</td>
<td>2002</td>
<td>Buffalo-Niagara Region (New York State, US)</td>
<td>2,779 households; 6,636 persons</td>
<td>0.59% (# of people sampled divided by the population, 2000 Census)</td>
</tr>
</tbody>
</table>

3. ANALYSIS FRAMEWORK

We propose an analysis framework that consists of three categories of properties: zeroth-order, first-order, and second-order properties [4]. The zeroth-order properties describe the overall properties of the data, such as the temporal and spatial distribution of observations (data points). They provide general data characteristics on how observations are distributed spatially and temporally. The first-order properties refer to data characteristics and patterns related to single activity locations, or trip ends, such as the stop time at a trip end (or termed as activity location). The locations are defined based on traffic analysis zones (TAZs) or census tracts. The second-order properties illustrate properties associated with two locations - usually one is trip origin and the other is trip destination. Properties associated with trips or more generally, OD patterns and associated travel times will be investigated. Since trip ends (first-order) and trips (second-order) are derived from the observations in the dataset (zeroth-order), it is clear that these properties are inherently related—the spatial and temporal characteristics of the data (zeroth-order properties) directly affect the quality of first- and second-order attributes.

To derive first- and second-order attributes such as trip ends and trip rates, algorithms must be applied. Since this paper is an overview paper on comparative characteristics of different datasets, we will not dwell on the algorithms used to derive such properties in detail. A number of methods have been applied to derive trip ends including threshold-based trace-segmentation methods [15][41], machine learning-based methods [40][46], and revised incremental clustering method [33]. In this study, we used a Divide, Conquer and Integrate’ (DCI) framework to process the app-
based data [34], and the revised incremental clustering method [33] to process mobile phone data. The DCI framework was designed to handle data sources that present high variances in their spatio-temporal properties such as the app-based data [34]; and the incremental clustering method was designed to handle the locational uncertainty in the mobile phone data [33]. Both methods are able to identify common activity locations that are frequently visited at different times (e.g., home and workplaces), which enable us to analyze individual mobility patterns such as regular returns [11].

Table 2 lists the three order properties investigated for this study. Not all properties apply to all four datasets, i.e., the zeroth-order properties mostly apply to big datasets.

Table 2. Analysis framework: zeroth-, first, and second-order properties

<table>
<thead>
<tr>
<th>Order</th>
<th>Metrics</th>
<th>Definition/content</th>
<th>App-based data</th>
<th>Mobile phone data</th>
<th>PSRC survey data</th>
<th>Buffalo survey data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0th</td>
<td>Intra-day temporal sparsity</td>
<td>Fraction of IDs with their locations revealed at time of a day</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0th</td>
<td>Inter-day temporal sparsity</td>
<td>Number of days observed</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0th</td>
<td>Weekly pattern of number of observations</td>
<td>Daily distribution of number of observations/sightings</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>Activity duration</td>
<td>Activity duration distribution (in hour)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>1st</td>
<td>Home location correlation</td>
<td>Correlation between inferred home locations and census data</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>Trip rate</td>
<td>Distribution of trip rates</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2nd</td>
<td>Departure time</td>
<td>Distribution of departure times</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2nd</td>
<td>Travel time</td>
<td>Distribution of travel times</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2nd</td>
<td>OD correlation</td>
<td>Correlation between MPO OD and estimated OD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

4. RESULTS

4.1 Zeroth-order properties

*Intra-day temporal sparsity*

The intra-day temporal sparsity is characterized by the metric: *fraction of IDs being sighted during an hourly interval*. Figure 1 compares this metric between the app-based data and the mobile phone data on weekdays and weekends. It can be seen that on weekdays, the app-based data has three small peaks within a day, between 7:00 and 9:00 am, 12:00 and 1:00 pm, and 4:00-6:00 pm. On the other hand, the weekend days follow a unimodal distribution with only one peak around 12:00 and 1 pm. For mobile phone data, the weekday data shows two peaks around 12 pm and 4 pm.
while the weekend days show only one peak around 1 pm. Interestingly, the two curves for the mobile phone data show a sustained plateau in the afternoon between 12 pm and 4 pm, as compared to the app data that experienced sharper increase or decline from one hour to another. Additionally, from mid-night to early morning, the fraction of IDs of app-based data are much larger than those of mobile phone data. The differences between these two datasets are inherently related to the underlying data generation process: app-based data are derived from apps usage while the mobile phone data are from phone calls and text messages. During night and early morning, the number of phone calls and text messages largely subsided while app use is still substantial (around 25%).

Figure 1. Fraction of IDs with their locations revealed at time of a day

*Inter-day temporal sparsity*

Figure 2 shows the distribution of *number of days observed* for all IDs during the entire study period as an indicator describing the intra-day data sparsity. For app-based data, nearly 45% of IDs appeared fewer than two days and nearly 65% of IDs appeared seven days or less. Almost 30% of users of mobile phone data were observed fewer than two days and half of them were captured seven days or less. In addition, about 12% of users in the app-based data and 36% in the mobile phone data can be observed between 7 days (one week) and 21 days (three weeks). Such data can be useful for capturing activities and trips (e.g., grocery trips) that are conducted infrequently, i.e., weekly or every two or more weeks. It has been observed that data of at least three weeks is needed to capture over 80% of the activities and trips individuals make routinely [5][21][25].

While the portion of users being observed for more than a few days declines rapidly for both datasets, the app-based data has a small peak toward the end of the two-month period and this does not apply to the mobile phone dataset. We have yet to understand the causes of this small peak for the app-based data and our speculation is that this small peak at the end does not reflect changes
in mobility patterns but due to a possible change of the technology used to capture data or the number of apps that contributed to the data generation process.

![Figure 2. Number of days observed](image)

Weekly pattern of number of observations

Figure 3 shows a consistent weekly pattern of daily number of observations for both big datasets, suggesting a daily variation for mobile app use or mobile phone use. The fraction of observations within the week is calculated as the ratio between the total number of observations on one day and number of observation within that week. Generally, weekdays have more sightings than weekends and Sundays have the least. The app-based data also shows an uptick on Friday consistently for three weeks, suggesting heavy usage of apps on Fridays. Phone call usage is also the highest on the last (third) Friday, as shown in Figure 3.

![Figure 3. Weekly patterns of number of observations](image)
4.2 First-order properties

Activity duration

Since household travel surveys are for residents in the area and the big data is generated for everyone using the network service, we processed the two big datasets to extract only those with a home location identified in the area. Thus, the two types of data (survey vs big data) are comparable since both are for residents in the area.

For both big datasets, activity durations are inferred and compared with those in the corresponding household travel surveys (the Puget Sound Regional Council (PSRC) household travel survey for app-based data, and the Buffalo household survey for the mobile phone data). Figure 4 shows the comparison. While the comparisons are largely similar in both cases, there are some notable differences. The app-based dataset yielded about 48% of the durations of less than one hour corresponding to about 40% from the PSRC travel survey (see Figure 4 (a)). The reverse seems to be true for the mobile phone data – about 32% activities identified from the mobile phone data lasting less than one hour corresponding to 40% from the Buffalo travel survey (see Figure 4 (b)). The app-based data are temporally less sparse (see Figures 1-2) than mobile phone sightings, so that more stays may be captured in the app-based data than the mobile phone data. And yet, this also points out a potential bias that has been revealed by a number of studies: trip rate tends to be positively correlated with number of sightings [4]. In both cases (Buffalo and Seattle), there seems to be an over-estimation of activities lasting between 2 and 6 hours and an underestimation of activities lasting between 7 and 14 hours for both big data sets when compared to their corresponding survey data.

(a) App-based data and PSRC survey
Figure 4. Distribution of activity duration

Home location correlation

In this study, the home location for a user is identified as one that is most visited during evening periods (22:00 to 6:00 the next day). Again, we refer the readers to other papers [2][33][34] for details on how the home location is derived. We then correlate the number of residents from our data at the census tract level to the population statistics at the same level and they are 0.91 for the app-based data and 0.433 for the mobile phone data. Two underlying differences between the app-based data and mobile phone data are likely responsible for this: 1) the mobile phone data generally have lower data accuracy compared with app-based data (cellular triangulation for mobile phone data vs. a mix of positioning techniques such as cellular triangulation, GPS, WLAN, Bluetooth for app-based data). Lower data accuracy makes it harder to correctly identify home locations at a small spatial scale such as census tracts; and 2), as noted earlier, app-based data is temporally less sparse than the mobile phone data (Figure 1-2).

4.3 Second-order properties

Trip rate

Figure 5 illustrates the distribution of trip rates across four datasets, (a) for app-based data and PSRC survey; and (b) for mobile phone data and Buffalo MPO survey. For both app-based data and mobile phone data, the estimated trip rates are significantly lower than those obtained from survey data. In the case of the app-based data, the estimated mean trip rates (per day) are 3.01 for weekdays, while the PSRC survey data show 3.45 for weekdays (see Figure 5 (a)). In case of mobile phone data, the estimated mean trip rates (per day) are about 1.78 for weekdays and 1.6 for weekends, while Buffalo travel survey data show 3.89 for weekdays (see Figure 5 (b)). Two observations can be made: first, both have an under-estimation bias; and second, app-based data appears to be able to capture more than the mobile phone data due to reasons discussed previously.
Figure 5. Distribution of trip rates

(a) App-based data and PSRC survey  
(b) Mobile phone data and Buffalo MPO survey

Departure time

Figure 6 shows the distribution of departure times derived from the two big datasets as compared to their household travel survey counterparts for weekdays. Both surveys show clearly three peaks in the morning around 8 am, at noon and in the afternoon between 3 and 5 pm, reflecting peaks in travel patterns (morning commute peaks, noon for lunch breaks and afternoon commute peaks). These are not readily reflected in the distribution for the two big datasets. The one consistent feature between the big data and the survey data is the afternoon peak, probably due to that in the afternoon between 4 and 6 pm is the peak time for afternoon commute, app use and phone calls.

Travel time

Figure 7 depicts the cumulative distribution of travel times obtained from the four datasets. The curves of two survey datasets share almost the same distribution. In the very early stage (0 – 100 minutes), the two survey curves grow much faster than the app-based data and the mobile phone
data, meaning that a much smaller percentage of users in the survey data have long travel time. Between the app-based data and the mobile phone data, 82% of the trips were completed within an hour for the former compared to 50% for the latter. This finding is consistent with the much lower trip rate calculated for the mobile phone data as compared to the app-based data (see Figure 5) as well as the higher levels of temporal sparsity observed for the mobile phone data (see Figures 1-2).

![Figure 7. Distribution of travel times](image)

**Correlation between MPO OD and estimated OD**

We further compare the MPO OD demands and estimated OD demands at the Traffic Analysis Zone (TAZ) level. The MPO OD here refers to the OD matrices provided by PSRC and Buffalo MPO, and estimated OD are the inferred results from app-based dataset and mobile phone dataset. The results show correlations in the range of 0.6-0.7. This indicates the estimated OD demands do not sufficiently represent the MPO demand matrices. Notice here that MPO demand matrices are also the modeling results from survey and other types of transportation data, and are thus not necessarily the “ground truth”. We assume that they are more representative of the true OD demand matrices since the data collection and estimation processes are more controlled and carefully designed. However, given the mediocre correlations between big-data-estimated OD matrices and MPO demand matrices, more research and investigations are needed to further study the data properties and develop more sophisticated OD estimation methods to produce better representative OD matrices from the big data sources.

**5. CONCLUSIONS AND DISCUSSIONS**

Using two case studies each in a specific geography and each involving a big dataset and a traditional household travel survey dataset, we illustrate some similarities and many differences
between the two big datasets and with the survey data, using a three-order analysis framework. The framework we proposed describes the temporal and spatial properties of the big data themselves (zeroth-order), attributes associated with a single location such as home locations and activity durations (first-order), and those associated with two locations such as trip rates, travel times and departure times (second-order). As noted earlier, these properties are inherently related with each other, as demonstrated in the analysis results—for example, the lower level of temporal sparsity for the app-based data compared to mobile phone data leads to higher correlation with the population statistics at the census tract level, more trips captured, and a closer resemblance to the survey data on travel time and departure time distributions for the former. It is also clear that the differences between the big data and the survey data result from the fundamental differences between the two in the underlying data generation process—the big data results from phone calls and app usage, or passively solicited for potentially transportation planning purposes such as deriving human mobility patterns while the survey data is actively solicited to directly capture individual movements. Thus much caution is needed when using the passively-solicited big data for planning purposes whose focus is on trips.

It shall be also noted that data processing is much needed for the big datasets and such processing must be clearly described prior to describing the results. Though this paper is not on the techniques and algorithms used to process such data, our results, in particular the zeroth-order ones, shall illustrate the challenges (e.g., temporal sparsity and lower trip rates as compared to survey data) that exist for such datasets to capture a full spectrum of activity and travels.

As noted earlier, the next-generation household travel surveys are most likely a hybrid of both traditional household travel surveys where sampled respondents are asked to log their activities and trips with great details for a full day (so actively-solicited) and passively-solicited big data such as mobile phone or app-based data. These two types of datasets have their own pros and cons: small and expensive but very rich information for the actively-solicited small survey data and big and inexpensive but very limited information for the passively-solicited big data. The key question is how to integrate them such that the advantages of each dataset are leveraged. This means that the rich information contained in the survey data can be retained and potentially extrapolated for the big data to support travel demand forecasting purposes; the size of the big data can be retained to capture non-residents and residents who are not sufficiently sampled in the survey data; and the longitudinal nature of the big data is also an appealing nature so that planning agencies and DOTs can respond to policy questions dynamically, even in real time when needed (e.g., under disaster scenarios). Since the underlying data generation process for big data is unknown, the very first step toward this integration is to understand their various properties. This paper seeks to fulfill that goal.

ACKNOWLEDGEMENTS

This study acknowledges support from the Federal Highway Administration (FHWA). C. Chen and F. Wang thank National Institute of Health (NIH) (1R01GM108731-01A1) and the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116. J. Ban and J. Wang thank supports from the National Science Foundation (CMMI-1719551) The authors also
appreciate help from the discussions with Xi Yang Guan (Ph.D. candidate at University of Washington), Jinzhou Cao and Wenxiang Li (visiting students at University of Washington). Views in this paper do not represent those of sponsors and the authors are responsible for all errors that may exist.

AUTHOR CONTRIBUTIONS

C. Chen and J. Ban designed the research framework; J. Wang processed the app-based data and PSRC household travel survey data; F. Wang processed the app-based data, mobile phone data and Buffalo household travel survey data; J. Wang and C. Chen wrote the manuscript; All authors read, commented and approved the final version of the manuscript. The authors also thank Dr. Brian Lee from PSRC for helpful discussions about PSRC survey data.

REFERENCES

Wang, Wang, Ban and Chen


Wang, Wang, Ban and Chen


