Title: Correcting biases in using emerging big data for mobility research: a likelihood-based approach

Authors

Xiangyang Guan[1], Cynthia Chen[1] and Shuai Huang[2]

[1]Civil and Environmental Engineering, University of Washington, Seattle, WA

[2]Industrial and Systems Engineering, University of Washington, Seattle, WA

Introduction

As emerging data sources such as GPS, mobile phone and social media are increasingly leveraged to probe human mobility patterns, concerns rise regarding the representativeness of these data sources and the human mobility patterns inferred from them, and consequently the generalizability of the discoveries and policies made based on the inferred mobility patterns. In this study, bias is defined as the condition where the mobility patterns inferred from a data source are not representative of the population's mobility patterns. At least two levels of biases could occur in a data generation process. First, users self-select into using certain services, which means whether a user is included in a data source depends on a number of individual factors such as gender, age and income. This suggests the users in an emerging data source are likely biased toward certain demographic groups. Second, for a given user, his/her likelihood of using a certain service could be uneven for different activity types, which leads to biased representation of different trips in the data source. Biases underlying a data source violate the key random sampling requirement for a discovery or policy conclusion to be generalizable beyond the data source.

This study seeks to mitigate the biases in the mobility patterns inferred from single data sources. It is assumed that the two levels of biases – in service subscription and usage patterns, respectively – are correlated to three categories of factors: individual demographics, trip/activity characteristics and service attributes. As each data source covers a subset of these three factors, the basic idea in correcting the biases in inferred mobility patterns is to integrate multiple data sources that likely have different coverages. This study thus answers the following two questions.

1. How to integrate multiple different data sources to infer mobility patterns? and
2. How effective is integrating multiple data sources on mitigating biases in inferred mobility patterns associated with each single data source?

This study contributes to the literature by proposing a novel methodology for integrating data and designing conceptually-sound scenarios to test its performance on mitigating bias. The methodology in particular addresses the dependency between data sources (i.e. whether a mobility pattern is captured by two data sources may affect each other), which is mostly overlooked in existing research. More theoretically, we construct a conceptual framework for describing trip-making behaviors and data generation processes, and contrast the framework with what happens in the real world.

Methodology

In developing the methodology, we assume to integrate two data sources. The developed methodology can be easily extended to more than two data sources. And without loss of generality, the mobility

pattern of interest is set to be the number of trips N_a in a zone *a*. Supposing the probabilities of a trip *i* being captured by data source 1 (e.g. mobile phone) and data source 2 (e.g. social media) are $p_{i,1,a}$ and $p_{i,2,a}$, respectively, Table 1 lists all possibilities regarding trip *i*'s presence in the two data sources.

Table 1 Probabilities of a trip being captured or not by each data source

	Not captured by source 2	Captured by source 2
Not captured by source 1	$(1-p_{i,a,1})(1-p_{i,a,2})$	$(1-p_{i,a,1})p_{i,a,2}$
Captured by source 1	$p_{i,a,1}(1-p_{i,a,2})$	$p_{i,a,1}p_{i,a,2}$

Among these four possibilities, the trips in three of them can be observed: those captured by data source 1 but not data source 2, $n_{a,1-2}$, those captured by data source 2 but not by data source 1, $n_{a,2-1}$, and those captured by both data source 1 and data source 2, $n_{a,1+2}$. The number of trips not captured by either data source is not observable, but is related to the observed numbers and N_a, i.e., $N_{a,-1-2} = N_a - N_{a,1-2} - N_{a,2-1} - N_{a,1+2}$ where N_* is the number of trips corresponding to n_*. Consequently, the likelihood of observed trips can be written as in equation (1).

$$L\left(n_{a,*}\middle|p_{*,a,*},N_a\right) = \prod_{i\in n_{a,1-2}} p_{i,a,1}(1-p_{i,a,2}) \prod_{i\in n_{a,2-1}} (1-p_{i,a,1})p_{i,a,2} \prod_{i\in n_{a,1+2}} p_{i,a,1}p_{i,a,2} \prod_{i\in n_{a,-1-2}} (1-p_{i,a,1})(1-p_{i,a,2}), \quad (1)$$

where $n_{a,*} \equiv \{n_{a,1-2}, n_{a,2-1}, n_{a,1+2}\}$ and $p_{*,a,*} \equiv \{p_{i,a,1}, p_{i,a,2} \text{ for } i \in n_a\}$. One challenge in evaluating the likelihood in equation (1) is that the trip set $n_{a,-1-2}$ is unobserved, and thus it is unknown which trip *i* belongs to the set $n_{a,-1-2}$, making the fourth multiplication term in equation (1) difficult to compute. To overcome this challenge, we ignore the heterogeneity among trips in $n_{a,-1-2}$, and use the mean probabilities to replace $p_{i,a,1}$ and $p_{i,a,2}$, i.e., $p_{a,1} = \int_i p_{i,a,1}\Pr(i)\,\mathrm{d}i$ and $p_{a,2} = \int_i p_{i,a,2}\Pr(i)\,\mathrm{d}i$, where $\Pr(i)$ represent the probability distribution for trip/activity characteristics. $\Pr(i)$ can be obtained from observed trips. The fourth multiplication term can thus be written as

$$\prod_{i\in n_{a,-1-2}} (1-p_{i,a,1})(1-p_{i,a,2}) = \left[(1-p_{a,1})(1-p_{a,2})\right]^{N_{a,-1-2}},$$

which makes N_a the only unknown quantity in equation (1). N_a can be further integrated out using equation (2).

$$L\left(n_{a,*}\middle|p_{*,a,*}\right) = \int_{N_a} L\left(n_{a,*}\middle|p_{*,a,*},N_a\right)\Pr(N_a)\,\mathrm{d}N_a, \quad (2)$$

where $\Pr(N_a)$ represents the probability distribution of N_a. $\Pr(N_a)$ can take an assumed form (e.g. Poisson distribution), and the integral in equation (2) can be calculated using MCMC.

The key of estimating N_a is to estimate $p_{*,a,*}$, as the posterior of N_a can be obtained through $\Pr\left(N_a\middle|n_{a,*}\right) = \Pr\left(N_a\middle|n_{a,*},p_{*,a,*}\right) \propto L\left(n_{a,*}\middle|p_{*,a,*},N_a\right)\Pr(N_a)$. Based on the aforementioned assumption on the relationship between mobility bias and the three categories of factors, we parameterize the probability of a trip being capture as a function of these three categories of factors.

$$p_{i,a,s} = f(x_i, x_a, x_s, x_{s^-}; \boldsymbol{\beta}), \quad (3)$$

where x_i, x_a and x_s are the factors associated with trip *i*, zone *a* and service *s*, respectively, with coefficients $\boldsymbol{\beta}$. The factor x_{s^-} indicates whether trip *i* is captured by data sources other than *s*, and reflects the dependency among the data sources. Equation (2) can thus be written as $L(n_{a,*}|p_{*,a,*}) = L(n_{a,*}|x_i, x_a, x_s, x_{s^-}, \boldsymbol{\beta})$, and $\boldsymbol{\beta}$ can be estimated using Bayesian inference following $\Pr(\boldsymbol{\beta}|n_{a,*}, x_i, x_a, x_s, x_{s^-}) = L(n_{a,*}|p_{*,a,*})\Pr(\boldsymbol{\beta})$. With $\boldsymbol{\beta}$ estimated, $p_{*,a,*}$ and N_a can be estimated accordingly as aforementioned.

Expected results

We test the performance of the above methodology in a simulation setting. By presuming the values of $\boldsymbol{\beta}$, the factors *x*'s and N_a, we generate the simulation data, which amounts to the set of trips captured by each of the two data sources. When applying the methodology to the simulation data, we design the following scenarios to mimic realistic situations where certain factors are not available.

- Baseline scenario: all factors *x*'s are available;
- Scenarios 1-2: the trip-level factor x_i and the zone-level factor x_a are missing in equation (3), respectively;
- Scenario 3: both x_i and x_a are missing in equation (3).

It is expected that when the model formulation mirrors the data generation process (i.e. in the baseline scenario), integrating the two data sources using the above methodology will lead to unbiased estimation of the number of trips N_a in zone *a*. When certain factors are missing in the model formulation (i.e. in scenarios 1-3), the estimated N_a will be unbiased toward the remaining factors, but biased regarding the missing factors.

Conclusion

The proposed methodology for integrating multiple data sources to mitigate bias in inferred mobility patterns have three novelties. First, its formulation rests upon the data generation process and the fundamental mechanisms for bias to emerge; second, dependencies among data sources are accounted for; and third, by adopting a likelihood-based approach, uncertainties in the inferred mobility patterns can be quantified. This methodology, once tested and validated, could provide a promising venue for future applications of emerging big data to gain useful, unbiased mobility insights.