Transportation Research Part C 58 (2015) 162–177

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

The path most traveled: Travel demand estimation using big data resources

Jameson L. Toole [a,1], Serdar Colak [b,*,1], Bradley Sturt [a], Lauren P. Alexander [b], Alexandre Evsukoff [c], Marta C. González [a,b]

[a] Engineering Systems Division, MIT, Cambridge, MA 02139, United States
[b] Department of Civil and Environmental Engineering, MIT, Cambridge, MA 02139, United States
[c] COPPE/Federal University of Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 1 June 2014
Received in revised form 19 April 2015
Accepted 21 April 2015
Available online 7 May 2015

Keywords:
Mobility
Location based services
Congestion
Road networks
Mobile phone data

ABSTRACT

Rapid urbanization is placing increasing stress on already burdened transportation infrastructure. Ubiquitous mobile computing and the massive data it generates presents new opportunities to measure the demand for this infrastructure, diagnose problems, and plan for the future. However, before these benefits can be realized, methods and models must be updated to integrate these new data sources into existing urban and transportation planning frameworks for estimating travel demand and infrastructure usage. While recent work has made great progress extracting valid and useful measurements from new data resources, few present end-to-end solutions that transform and integrate raw, massive data into estimates of travel demand and infrastructure performance. Here we present a flexible, modular, and computationally efficient software system to fill this gap. Our system estimates multiple aspects of travel demand using call detail records (CDRs) from mobile phones in conjunction with open- and crowdsourced geospatial data, census records, and surveys. We bring together numerous existing and new algorithms to generate representative origin–destination matrices, route trips through road networks constructed using open and crowd-sourced data repositories, and perform analytics on the system's output. We also present an online, interactive visualization platform to communicate these results to researchers, policy makers, and the public. We demonstrate the flexibility of this system by performing analyses on multiple cities around the globe. We hope this work will serve as unified and comprehensive guide to integrating new big data resources into customary transportation demand modeling.

1. Introduction

The accelerating growth of cities has made the estimation of travel demand and the performance of transportation infrastructure a critical task for transportation and urban planners. To meet these challenges in the past, methods such as the widely used four-step model and more recent activity based models were developed to make use of available data computational resources. These models combine meticulous methods of statistical sampling in local (Daganzo, 1980; Smith, 1979) and national household travel surveys (Stopher and Greaves, 2007; Richardson et al., 1995) to process and infer trip

* Corresponding author.
E-mail address: serdarc@mit.edu (S. Colak).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.trc.2015.04.022
0968-090X/

information between areas of a city. The estimates they produce are critically important for understanding the use of transportation infrastructure and planning for its future (Van Zuylen and Willumsen, 1980; Spiess, 1987; Maher, 1983; Lo et al., 1996; Hazelton, 2003, 2001, 2000; Lu et al., 2013; Cascetta, 1984; Bell, 1991).

While the surveys that provide the empirical foundation for these models offer a combination of highly detailed travel logs for carefully selected representative population samples, they are expensive to administer and participate in. As a result, the time between surveys range from 5 to 10 years in even the most developed cities. The rise of ubiquitous mobile computing has lead to a dramatic increase in new, *big data* resources that capture the movement of vehicles and people in near real time and promise solutions to some of these deficiencies. With these new opportunities, however, come new challenges of estimation, integration, and validation with existing models. While these data are available nearly instantaneously and provide large, long running, samples at low cost, they often lack important contextual demographic information due to privacy reasons, lack resolution to infer choices of mode, and have their own noise and biases that must be accounted for. Despite these issues, their use for urban and transportation planning has the potential to radically decrease the time in-between updated surveys, increase survey coverage, and reduce data acquisition costs. In order to realize these benefits, a number of challenges must be overcome to integrate new data sources into traditional modeling and estimation tools.

Analyzed on its own, data generated by the pervasive use of cellular phones has offered insights into abstract characteristics of human mobility patterns. Recent work has found that individuals are predictable, unique, and slow to explore new places (González et al., 2008; Brockmann et al., 2006; de Montjoye et al., 2013; Song et al., 2010a,b; Candia et al., 2008; Calabrese et al., 2013). The availability of similar data nearly anywhere in the world has facilitated comparative studies that show many of these properties hold across the globe despite differences in culture, socioeconomic variables, and geography. The benefits of this data have been realized in various contexts such as daily mobility motifs (Schneider et al., 2013; Sevtsuk and Ratti, 2010), disease spreading (Belik et al., 2011; Wesolowski et al., 2012) and population movement (Lu et al., 2012). While these works have laid an important foundation, there still is a need to integrate these data into transportation planning frameworks.

To make these new data useful for urban planning, we must clarify their biases and build on the progress made by transportation demand modeling even in the face of limited data resources. We must combine this domain knowledge with new algorithms and metrics to better understand travel behaviors and the performance of city infrastructure and we must update technologies to accommodate the computational requirements of processing massive geospatial data sets. Individual survey tracking and stay extraction (Asakura and Hato, 2004), OD-estimation and validation (Caceres et al., 2007; Nie et al., 2005; Wang et al., 2012; Iqbal et al., 2014), traffic speed estimation (Bar-Gera, 2007; Zhan et al., 2013), and activity modeling (Phithakkitnukoon et al., 2010; Reades et al., 2009) have all been explored using new massive, passively collected data. However, these studies generally present alternatives for only a few steps in traditional four-step or activity based models for estimating travel demand or fail to compare outputs to travel demand estimates from other sources. Moreover, many methods offered to date lack portability from one city to many with minimal additional data collection or calibration required.

Here we fill this gap with a modular, efficient computational system that performs many aspects of travel demand estimation billions of geo-tagged data points as an input. We review and integrate new and existing algorithms to produce validated origin–destination matrices and road usage patterns. We begin by outlining the system architecture in Section 2.1. In Section 2.3 we explain our methods of extracting, cleaning, and storing road network information from a variety of sources. We discuss recent advances in OD creation from mobile phone data in Section 3.1 and implement a simple, parallel incremental traffic assignment algorithm for these trips in Section 3.2. We present comparisons of these results to estimates from traditional survey methods in Section 4.1. Finally, in Sections 4.2, 4.3, 4.4 we present a variety of measurements that can be made with the proposed system as well as an online, interactive visualization for conveying these results to researchers, policy makers, and the public. To demonstrate the flexibility of the system, we perform these analyses for five metro regions spanning countries and cultures: Boston and San Francisco, USA, Lisbon and Porto, Portugal, and Rio de Janeiro, Brazil.

1.1. Description of data

Travel surveys are typically administered by state or regional planning organizations and are integrated with public data such as census tracts and the demographic characteristics of their residents, made available by city, state, and federal agencies. New data sources, however, come from new providers. Large telecommunications companies, private applications, and network providers collect and store enormous quantities of data on users of their products and services, presenting computational challenges for storing and analyzing them. Billions of phone calls must be processed, data from open- and crowd-sourced repositories must be parsed, and results must be made more accessible to individuals that generated them. At the same time, it is critical that measurements from these new sources are statistically representative and corrected for biases inherent in new data. This process requires integration of new pervasive data with reliable (though less extensive) traditional data sources such as the census or travel surveys. We combine the following data sets to illustrate the capabilities of the system architecture here proposed:

1. *Call Detail Records (CDRs):* At least three weeks of call detail records from mobile phone use across each subject city. The data includes the timestamp and the location for every phone call (and in some cases SMS) made by all users of a particular carrier. The spatial granularity of the data varies between cell tower level where calls are mapped to towers