Transportation Research Part C xxx (2015) xxx–xxx

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 xxxx

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 crowdsourced 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

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

and triangulated geographical coordinate pairs where each call has a unique pair of coordinates accurate to within a few hundred meters. Market shares associated with the carriers that provide the data also vary. Personal information is anonymized through the use of hashed identification strings. For reference, 6 weeks of CDR data from the Boston area containing roughly 1 billion calls made by 1.6 million unique users consumes roughly 70 gigabytes of disk space in its raw format. In cities with longer observation periods, data size quickly becomes a performance issue.

2. *Census data:* At the census tract (or equivalent) scale, we obtain the population and vehicle usage rate of residents in that area. For US cities, the American Community Survey provides this data on the level of census tracts (each containing roughly 5000 people). Census data is obtained for Brazil through IBGE (Instituto Brasileiro de Geografia e Estatística) and for Portugal through the Instituto de Nacional de Estatística. All cities analyzed in this work have varying spatial resolutions of the census information.
3. *Road networks:* For many cities in the US, detailed road networks are made available by local or state transportation authorities. These GIS shapefiles generally contain road characteristics such as speed limits, road capacities, number of lanes, and classifications. Often, however, these properties are incomplete or missing entirely. Moreover, as such road inventories are expensive to compile and maintain, they simply do not exist for many cities in the world. In this case, we turn to OpenStreetMaps (OSM), an open source community dedicated to mapping the world through community contributions. For cities where a detailed road network cannot be obtained, we parse OSM files and infer required road characteristics to build realistic and routable networks. At this time, the entirety of the OSM database contains roughly 4 terabytes of geographic features related to roads, buildings, points of interest, and more.
4. *Survey and model comparisons:* Wherever possible, we obtain the most recent travel demand model or survey from a particular city and compare the results to those output by our methods. In Boston, we use the 2011 Massachusetts Household Travel Survey (MHTS) and upscale trips according to standard procedures, in San Francisco, the 2000 Bay Area Transportation Survey (BATS), in Rio de Janeiro, a recent transportation model output provided by the local government, and in Lisbon, the most recent estimates from the MIT-Portugal UrbanSim LUT model that uses the 1994 Lisbon transportation survey as input (Ferreira et al., 2010). We found no recent travel survey or model for Porto.

Table 1 compiles descriptive statistics for these data sources for each city we explore in the latter sections of this paper.

2. System architecture and implementation

2.1. Architecture

The system architecture to integrate the data sources above must be flexible enough to handle different regions of the globe which may have different data availability and quality and efficient enough to analyze massive amounts of data in a reasonable amount of time. The proposed system must also be modular, so that components can be updated easily as new technologies and algorithms become available. To meet these requirements, we choose an object-oriented approach with loose schema requirements. A final object is to make results accessible to a range of end users via online, interactive visualization. To satisfy these constraints, we propose the system architecture depicted in Fig. 1.

2.2. Parsing, standardizing, and filtering user data

One of the biggest challenges in parsing and analyzing travel survey data is the incredible variety in data schema, collection, and reporting practices. Each planning organization typically constructs its own set of data codes and definitions and provides data in unique formats. This makes it very difficult to compare surveys done in different cities. Call detail records, on the other hand, are typically available for many cities from the same provider and in the same format, and in most cases, translating between the formats of different carriers is simply a matter of shuffling columns. The first component of our system is a simple architecture to convert all CDR data to a standard format that can be expected by the rest of the components.

Given the size of these data sets and the rapidly evolving schema requirements of new models, choosing the proper data structure is critical. Google's open source Protocol Buffer library[2] is an ideal choice as they provide fast serialization for speed and space efficient file storage as well as flexible schemas that can be changed without compromising backwards compatibility. These structures were designed to serve some of the largest databases in the world and are more than enough for our task.

We take a user centric approach to CDR data. We define a *user_data* protocol buffer message that will form the core data structure for our custom User class in an object-oriented programing model. Each User object can be assigned a number of attributes such as the number of calls they make, their home and work locations, and mobility characteristics such as the average time between calls or the average distance traveled on each trip. More sophisticated methods can compute the number and distribution of their trips and even expand them based on census information. We define similar structures and classes for OD matrices, trips, and census data. The serialization routines built into the protocol buffer library ensures that storage of raw data is efficient. To analyze a new city, the user only needs to write two simple routines, one to parse a single

[2] Google Protocol Buffers https://developers.google.com/protocol-buffers/.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

Table 1
A comparison of the extent of the data involved in the analysis of the subject cities.

	City				
	Boston	SF Bay	Rio	Lisbon	Porto
Population (mil.)	4.5	7.15	12.6	2.8	1.7
Area (1000 km^2)	4.6	18.1	4.5	2.9	2.0
# of Users (mil.)	1.65	0.43	2.19	0.56	0.47
# of Calls (mil.)	905	429	1045	50	33
# of cell towers	N/A	892	1421	743	335
# of Edges (ths.)	21.8	24.3	22.7	28.1	15.1
# of Nodes (ths.)	9.6	11.3	22.1	16.1	8.6
# of Tracts	732	1139	729	295	272

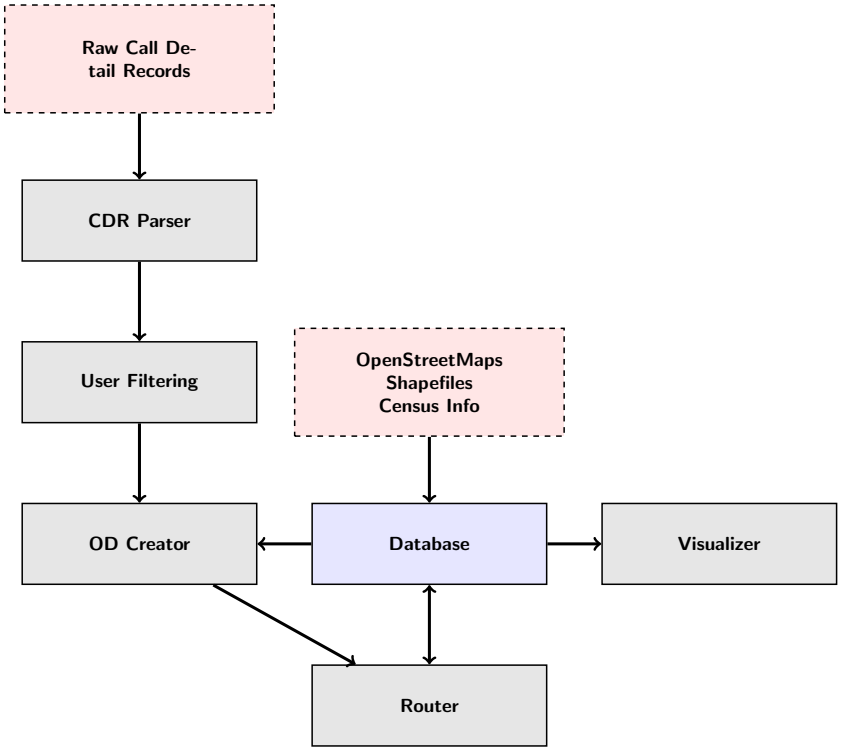

Fig. 1. A flowchart of the system architecture.

line of the CDR file and populate relevant user attributes and one to populate census data objects. Standardizing the CDR data format in this way makes it very easy to compare the output of our estimation models across different cities.

2.3. Creating and storing geographic data

A relational database is used to store road network and census information for every city in a standard format. Given the current cost of computing resources, these systems provide adequate performance for storing static GIS and census data and have convenient, mature interfaces for easy access. We also use this database to store aggregated results from our estimates so that they can be made available to interactive web APIs and visualization platforms. We use a Postgres and the open source spatial extension PostGIS to store and manipulate census and road network data.

While census tract or TAZ (Traffic Analysis Zone) polygons and demographic information are stored in this database, it is computationally inefficient to perform point-in-polygon calculations for each user or call record in our CDR dataset. To dramatically speed these computations, we rasterize polygons into a small pixel grid, where pixel values is a unique identifier for the census tract covering that pixel. This raster is then used as a look-up table to convert the latitude and longitude of calls into census tract IDs. The rasterization introduces some error along the boarders of tracts, but these errors are minimized by making pixel sizes much smaller than the size of the raster and resolution of the location estimates of calls (between 10 m and 100 m).

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

While the platform supports road networks supplied by local municipalities in the form of shapefiles, we have implemented a parser to construct routable road networks from OpenStreetMap (OSM) data due to its global availability. Transportation networks in OSM are defined by *node* and *way* elements. Nodes represent points in space that can refer to anything from a shop to a road intersection, while ways contain a list of references to nodes that are chained together to form a line. In our context, relevant ways are those used by cars and relevant nodes are intersections within the road network. Ways and nodes may also contain a number of tags to denote attributes such as "number of lanes" or "speed limit". Many roads, however, do not include the whole set of attributes necessary for accurate routing. For example, city roads often lack speed limit information required to estimate the time cost, which in turn is used to find shortest paths based on total travel time. To infer this missing data, our system supports the creation of user-defined mappings between highway types and road properties. For example, ways tagged as "motorways" are generally major highways and have a speed-limit of 55 mph in the Boston area. They tend to have 3 lanes in each direction. "Residential" roads, on the other hand, have a speed-limit of 25 mph and 1 lane in each direction. Each road segment is also given a capacity based on formulas suggested by the US Federal Highway Administration. Using these mappings, we parse the OSM xml data to create a routable, directed road graph with all properties required to estimate realistic costs driving down any given road.

We implement two additional cleaning steps to improve efficiency. The first filters out irrelevant residential roads. These small local roads are filtered from our network, as they are not central to the congestion problem, yet tend to increase computation time significantly. Finally, in OSM data, a node object can refer to many things, for example an actual intersection or simply a vertex on a curve used to draw a turn. The latter case results in a network node with only one incoming and one outgoing edge (assuming U-turns are not allowed). These nodes are superficial and increase network size and routing algorithm run times needlessly. We simplify networks by removing these nodes from the network and only connecting true intersections, keeping the geographic coordinates of the nodes so that link costs still reflect actual geographic length of roads rather than straight line distances between start and end points. The parsed and cleaned edges are then loaded into the Postgres database, preserving attributes and geometry. Pseudo-code of the algorithm to parse and simplify OSM networks can be found in Algorithm 1 in the supplementary materials.

3. Estimating origin–destination matrices

The following sections review algorithms for transforming billions of geo-tagged data points into validated origin destination matrices and assigning these flows to transportation infrastructure. Some of these algorithms are important for their deviation from traditional approaches and some are important for their computational efficiency, a requirement when faced with such massive data sets.

3.1. Measuring flow

Current methods to estimate the flow of people or vehicles from place to place in a city generally fall into two categories: four-step or activity based approaches. The former class of models breaks the process into a sequence of four steps from which it earns its name. The first three steps in a four-step model – trip generation, distribution, and mode choice – are designed to estimate origin–destination matrices containing the number of trips from place to place within a city. Traditional modeling approaches use data from travel surveys possibly combined with land use and point of interest information to generate estimates of trip production and attraction for locations. These trips are then distributed from their point of origin to destinations across the city using gravity or radiation models. Modes of transit are assigned using models estimated from survey data and information on the transit infrastructure. More recent activity based models approach travel demand from an individual level. Assuming that travel demand is created by the need to fulfill activities, these models use similar survey data to estimate utility curves for travels and predict behaviors using probit or logit models based on these preferences.

While new data sources such as CDRs do not provide the same detailed demographic and contextual information about individuals or trips, they do provide an opportunity to measure travel more directly. With billions of data points, high spatio, temporal resolution, and long observation periods, passive data collected by mobile devices provide unparalleled scale of observation. New methods to estimate travel demand must balance trade offs between small, but complete data for a short period of time and large, but incomplete data over a longer period of time. In both cases noise and biases must be carefully dealt with to produce valid measurements. In this section we adapt and integrate previous works that have tackled parts of this problem into a full implementation of travel demand estimation for cities.

Mobile phones offer good, but imperfect measurements of geographic position. The coordinates of a mobile phone event are either recorded as the location of a nearby tower through which the event was routed or as a triangulation based signal strength from multiple towers. This creates uncertainties of a few hundred meters in estimates of a user's location. Moreover, observations are only recorded when an individual uses his or her device, resulting in heterogeneous sampling frequencies between users and at different times for a given user. While sampling rates and data density are increasing rapidly with rising penetration rates and usage, these issues present statistical challenges.

Initial methods by Wang et al. construct *transient origin–destination matrices* by simply counting a trip for pair of consecutive calls made within the same hour from two different towers. However, this method lead to an abundance of short trips

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

and provided a very biased view of movement. Instead, mobile phone trajectories must be de-noised to remove spurious points or calls made in the middle of routes rather than origins or destinations. To extract meaningful locations, termed as *stays*, algorithms have been developed to smooth out this noise and control for these biases. Jiang et al. provide a thorough review of these techniques in Jiang et al. (2013) and we adapt the stay point algorithm originally described by Zheng and Xie (2011).

Given a user's trajectory of spatiotemporal points $P = \{p_1(x_1, y_y, t_1), \ldots, p_n(x_n, y_n, t_n)\}$, the goal is to discover meaningful locations at which a user repeatedly stays for a significant amount of time. The algorithm begins by considering each call in a time ordered sequence. Two consecutive (p_i, p_{i+1}) points are considered to form the start of a *candidate set* of points at the same semantic location if the distance between them is less than a threshold $\Delta r_{i,i+1} < \delta$. Subsequent points are added to this candidate set if they also meet this criteria, e.g. p_{i+2} is added if $\Delta r_{i+1,i+2} < \delta$. The result is a candidate set $S = \{p_s(x_s, y_s, t_s), \ldots, p_t(x_t, y_t, t_t)\}$ containing a number of consecutive calls. A candidate set is considered to represent a single *candidate stay* if time between the first and the last observation in the subsequence S are separated by a time greater than a threshold $\Delta t_{m,n} > \tau$. The geographic location of a candidate stay is set to be at the centroid of points in S. Due to noise in locations and daily call frequencies, multiple candidate stays that are actually the same place may be estimated at a slightly different geographic coordinate on different observation days. To account for this, a final agglomerative clustering algorithm is used to consolidate candidate stays to a single semantic location regardless of the temporal sequence of individual calls. Though many agglomerative clustering algorithms exist, we implement a simple, efficient grid based approach by assigning each filtered location to a grid cell and then defining a final *stay point* as the centroid of all filtered locations in each cell. A final pass through the original calls assigns any call within a distance δ from a stay point to that stay point regardless of whether or a not a consecutive call was recorded from that location. This algorithm removes noisy or spurious outliers from the data set while preserving as much information on visits as possible. It may also be run on both triangulated and tower-based CDR data, in the latter case it removes noise associated with calls from the same location being routed through different nearby towers due to environmental factors. Pseudo-code can be found in Algorithms 2–5 in the SI.

With de-noised trajectories of stay points, the next step is to infer contextual information about each location. Alexander et al. (2015) and Çolak et al. (2015) improve on methods by Wang et al. (2012) and Iqbal et al. (2014) by using visit frequencies and temporal data to infer contextual information such as a location's function or trip purpose. A user's *home* location is defined as the stay point they are observed at most frequently between the hours of 8 pm and 7 am on weeknights. Their *work* location is defined as the stay point other than home that a users visits the most between the hours of 7 am and 8 pm on weekdays. Because many individuals do not work, we leave the work location blank if the candidate location is not visited more than once per week or if the location is less than 500 m from their home location. All remaining non- home or work stay points are designated as *other*.

Daily trips are estimated from filtered users by analyzing consecutive observations at different stay points during a given time window. They begin by defining an *effective day* as a period between 3 am one morning and 3 am on the next consecutive morning. This definition is used to minimize the number of trips that are prematurely ended due to the assumption that users start and end each day at home. A home-based work (HBW) trip is counted if a user is observed to travel between home and work, a non-home based (NHB) trip is counted if a user moves between two non-home stay points, and a home-based other (HBO) trip is counted if a user is observed moving between their home location and a location labeled as other.

Though a user must have traveled between two different observed stay points at some in time, we do not know the precise departure time. We assign a random departure time based on the conditional probability that user departed during an hour between the time they were last observed at the origin and the time they were first observed at the destination. This conditional probability function for departure time can be derived from surveys such as the National Household Travel Survey or estimated empirically using observed call frequencies of all users over the course of the day. Alexander et al. show that this method produces CDR trip departure time distributions in line with multiple surveys for the Boston region. Having assigned departure times and purposes to each trip, we can construct trips made by a given user. Generally, we are interested in trips between geographic areas such as towns or census tracts so here we convert origin and destination points to IDs of the tract of zone they are in. The result is a vector of trips between locations in the city for each user in our data set.

While a trip represents an observation of movement of at least one person between two locations, we expand these trip counts to represent all individuals in a city. Expansion is a critical step in models relying on survey data where the sample sizes are typically less than 1% of the population. Here we generally have hundreds of thousands of users in our sample, but must still be careful to control for differences in market share and usage rates across a city. We first scale trips based on how often an individual uses their phone. For each user, we calculate the average number of trips made during a given time window by dividing the number of trips counted by the number of days that user was observed making a call. This step effectively measures the average number of trips a user makes between two locations on a day given that they are observed in our data set.

Due to differences in daily usage of mobile phones among the population, not every user makes enough calls on a typical day to infer their movement patterns. For this reason, we must filter out users that do make enough calls. This step requires trade-offs between sample size and amount of data we have on each selected user. Because we will eventually be routing these trips through the transportation network, it is important to correctly estimate the total number of trips taken as well as the distribution of trips across the city. In practice, we find that filtering out users who we measure to make fewer than 2.5

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

trips per day leaves a large sample size of active users and results in valid estimates of trip tables and OD matrices as shown in subsequent sections. Those implementing these methods may find that different filtering criteria produce samples suited for different tasks.

We then expand the average trip counts of filtered users to account for market penetration rates. As with survey participants, the ratio of cell phone users to the population is not uniform within the region. Each user is assigned a home census tract and expansion factors are computed for each tract by measuring the ratio of the number of users assigned there and the reported population. In cities such as Boston, these expansion factors tend to be less than 10, but can be higher in places with lower market share. They are generally much lower than surveys which may only choose two or three individuals to represent hundreds or thousands in an area. Each user's typical daily trip volumes are then multiplied by the expansion factor corresponding to their home tract and the now represent the movements of some fraction of the tracts population.

Finally, we may wish to consider only trips via a certain mode, e.g. vehicle trips. Though CDR data does not provide resolution required to measure mode choice, vehicle trips can be approximated by weighting person trips by vehicle usage rates in the home census tract of users. In this way, full OD matrices for vehicle or person trips are computed by summing the expanded trip volume computed for all users between all pairs of census tracts. We also construct partial OD matrices containing only trips of a certain purpose during a certain time window. Due to the relative consistency of CDR data around the world, we can adopt this same OD creation procedure in all cities. Pseudo-code to generate OD matrices has been adapted from Alexander et al. (2015) and Çolak et al. (2015) and can be found in Algorithms 6 and 7 in the SI. The results from this method are compared to the output of traditional models where applicable. Trip tables and correlations plots can be found below in Section 4.1.

3.2. Trip assignment

Having estimated OD flows, our next task is to efficiently assign these trips to transportation infrastructure, in this case a road network (Bast et al., 2007). The first step takes tract to tract OD matrices and distributes trips among nodes, or intersections. A trip originating in a census tract is assigned uniformly at random to an intersection in that tract and to an intersection within its destination tract. This distributes flows such as not to create artificial congestion points and reflects general uncertainty in the exact origin of trips. Other approaches, however, may consist of using abstract centroid nodes unique to each tract and connect to a number of other intersections within that tract using what's referred to as centroid connectors. With intersection to intersection flows, the next task is to assign traffic to routes.

Traffic assignment is another mature domain that has been studied extensively by urban and transportation planners. Static non-equilibrium models approaches consist of treating all users as homogenous agents who make route choices prior to departure based on some heuristic related to current traffic conditions (e.g. the path that minimizes travel time). Incremental Traffic Assignment (ITA) is a variant of these static non-equilibrium assignment models that assigns batches of trips serially and updates costs between increments, as an improvement over the simplest all-or-nothing assignment methods. However, it is known that dynamic equilibrium models are more realistic in assigning trips as outcomes are closer to the Wardrop principles (Wardrop, 1952), or Nash Equilibria, where drivers seek paths that minimize their travel time and in the final traffic conditions, no driver has an incentive to change their route. To take a step further from static models, Dynamic Traffic Assignment (DTA) (Merchant and Nemhauser, 1978) models take an iterative and temporally more coherent approach. The addition of these complexities help model traffic flow at finer granularity, enabling road segments to have different conditions within themselves and consequently the representation of phenomena like congestion spill-back, FIFO principle, and others (Colak et al., 2013).

Our system is modular so that it may implement any number of traffic assignment algorithms. Here, however, we take a simple ITA approach, as it is computationally efficient for many trip pairs in detailed road networks and allows us to keep track of each vehicle as it is routed through the network. We develop a set of tools to perform large scale routing and traffic assignment using parallelization for speedups. First, the parsed and optimized road network is loaded into a graph object. In our implementation, we use the Boost Graph Library for its flexibility and efficiency. We can then compute shortest paths based on a user defined cost (in this case travel time on road segments). We choose the A* algorithm among the wide range of shortest path algorithms, as it's widely used in routing on geographic networks for its flexibility and efficiency. The A* algorithm implements a *best-first-search* using a specified heuristic function to explore more promising paths first. The euclidian distance between nodes provides an intuitive heuristic that ensures optimal solutions are found. While this algorithm provides the same results as Dijkstra's algorithm, we find that it becomes more efficient to compute paths one by one for sparse OD matrices.

On most city roads, free-flow speeds are rarely achieved due to congestion. As a result, traffic patterns may significantly change the time costs associated with using a particular route. To address this, we implement an Incremental Traffic Assignment (ITA) algorithm (Ortúzar and Willumsen, 1994). A simplified schematic explaining the procedure can be seen in Fig. 2. This algorithm assigns trips in a series of increments and updates the costs of edges in the network based on the number of vehicles that were previously assigned to that road between increments. For example, the first increment assigns 40% of trips for each pair assuming each driver experiences free-flow speeds. The travel time cost associated with every road segment is then adjusted based on how many drivers were assigned to that road and the total number of cars a road can accommodate in unit time. The next 30% of drivers are then routed in the updated conditions. This process is repeated until all users have been assigned a route. The shortcoming of this method is that once a driver has been assigned

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

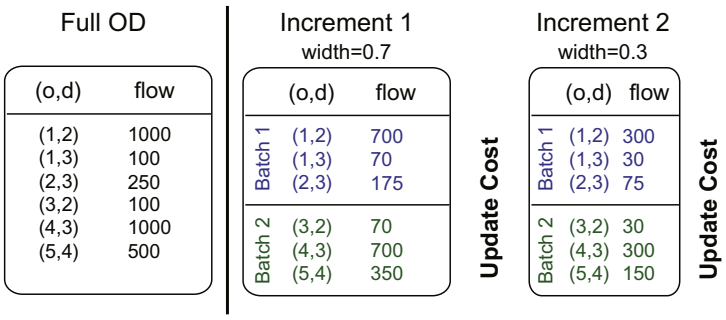

Fig. 2. Our efficient implementation of the incremental traffic assignment (ITA) model. A sample OD matrix is divided into two increments and then split into two independent batches each.

a route it does not change, and consequently the approach does not converge to Wardrop's equilibrium even for very small increment sizes. Yet we use it here due ease of implementation and the fact that it is still insightful for the purposes of demonstrating the implementation of a modular data-driven travel demand model. Future work will explore the use of newer methods.

Relating travel performance to traffic conditions has been a long standing problem in transportation. Many different characterizations exist, ranging from conical volume-delay functions to more complex approaches (Branston, 1976; Spiess, 1990; Akcelik, 1991). One of the most simplistic and common metrics used in determining the travel time associated with a specific flow level is the ratio between the number of cars actually using a road (volume) and it's maximum flow capacity (volume-over-capacity or V/C). At low V/C, drivers enjoy large spaces between cars and can safely travel at free-flow speeds. As roads become congested and V/C increases, drivers are forced to slow down to insure they have adequate time to react. Based on the volume-over-capacity (V/C) for each road, costs are updated according to Eq. (1), where $\alpha = 0.15,\ \beta = 4$ are used per guidelines set by the Bureau of Public Roads.[3]

$$t_{current} = t_{freeflow} \cdot (1 + \alpha(V/C)^{\beta}) \tag{1}$$

Though increments must be routed in serial, all routes discovered within an increment are independent. To speed up the routing process, we divide all trips in an increment into batches and send these batches to different threads for parallel computation. Because the road network remains fixed in each increment, we only need to store a single graph object shared by all threads. When a shortest path is found, we walk that path and increment counts of the number of vehicles that were assigned to each road and sum the counts from all batches after the increment has finished. We also keep track of the origin and destination census tracts of the assigned vehicles in a bipartite graph for later analysis. After all trips have been routed, we compute final V/C ratios and other metrics of each segment and update these values in the database so they can be used for other applications or visualization. Pseudo code for this ITA procedure can be found in Algorithm 8 in the SI.

4. Results

In the following sections we demonstrate the range of outputs provided by our system. We first report trip tables and compare origin–destination matrices produced by our system to available estimates made using travel surveys. We then report road network performance as well as characteristics of road usage patterns enabled by the construction of a bipartite road usage network.

4.1. Trip tables and survey comparison

In order to understand when and where these new data will be effective and how the results differ from traditional approaches, we compare the output of our system to previous travel surveys wherever possible. In four of the cities studied, we find estimates of travel demand from surveys: the 2011 Massachusetts Household Travel Survey (MHTS) in Boston, the 2000 Bay Area Travel Survey (BATS) in San Francisco, a 2013 transportation plan in Rio de Janeiro, and estimates from a 2012 LUT model in Lisbon (Ferreira et al., 2010). While these surveys do not always produce all estimates we are able to generate with our system, we make comparisons wherever possible.

Trip tables report the total number of trips of a given purpose or during a given time of day for a city and represent the total load placed on transportation infrastructure. In Table 2, we report trip tables for each city in this study. We find close agreement with trip tables estimated using CDR data and surveys in Boston and the San Francisco Bay Area and less agreement in Rio de Janeiro. We note, however, that the 3.74 million person trips estimated for Rio is far too low given the population of the region and highlights the difficulty in finding reliable planning resources in many areas. Finally, we note that in

[3] Travel Demand Modeling with TransCAD 5.0, User's Guide http://www.caliper.com/PDFs/TravelDemandModelingBrochure.pdf.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

Table 2
Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based word (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7–10 am (AM), 10 am–4 pm (MD), 4–7 pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. In general, trip magnitudes align closely, with the exception of Rio de Janeiro, where the survey results report far too few trips, illustrating the difficulty of obtaining sensible measurements via certain techniques. No comparisons could be found for Porto.

City	HBW	HBO	NHB	AM	MD	PM	RD	Total
Boston	5.76	8.99	6.72	3.71	7.68	5.75	4.33	21.47
MHTS	3.22	12.83	9.49	5.32	8.87	8.20	3.15	25.54
SF Bay	4.07	10.05	7.04	4.47	7.81	5.35	3.53	21.16
BATS	4.60	11.54	4.66	4.18	6.90	4.22	3.00	20.80
Rio	9.92	17.17	11.46	7.71	14.09	10.47	6.29	38.55
Survey	2.06	–	–	1.31	1.19	1.24	–	3.74
Lisbon	1.08	2.01	1.21	0.79	1.67	1.26	0.58	4.30
Survey[a]	0.61	–	–	–	–	–	–	–
Porto	0.49	0.87	0.46	0.32	0.70	0.54	0.27	1.83
Survey	–	–	–	–	–	–	–	–

[a] Note that the Lisbon Survey only contains estimates of vehicle trips in millions.

Lisbon, the survey results represent vehicle trips only, while we report person trips. When adjusting for mode car ownership rates in Portugal, our numbers align more closely. We were unable to find a survey or model for comparison in Porto.

In addition to trip tables, it is also necessary to compare the distribution of trips from place to place around the city. In order to make this comparison, the area unit of analysis for the survey and our model must be aligned. Given the resolution of mobile phone data, our system is designed to create ODs at the census tract (or equivalent) level while many surveys aggregate to larger traffic analysis zones or super districts. For comparison, we aggregate the OD matrices from CDRs to the coarser grained resolution provided by the survey and compare results. Fig. 3 shows correlation histograms comparing OD matrices at the largest spatial aggregation available produced by our methods and those produced by traditional methods. In general we find very high correlations in Boston, San Francisco, and Rio, with lower correlations in Lisbon. Lisbon, however, has the smallest units of aggregation and these results demonstrate the limitations of these comparisons at very high spatial resolutions. We hope future work explores how these correlations relate to the modifiable area unit problem. Finally, there is significant uncertainty in all models and we hope future works will explore this uncertainty further.

4.2. Road network analysis

The first output of this procedure is volume, congestion (volume-over-capacity), and travel times for all road segments. Using the outcomes of our analyses, we calculated the distributions of volumes on roads, along with V/Cs in Fig. 4. Interestingly, the results suggest qualitatively similarly distributed volumes and V/Cs for our five subject cities. Moreover, our findings are consistent with general congestion studies that identify Rio de Janeiro as one of the most congested cities in the world and the San Francisco Bay Area not far behind. Smaller cities such as Boston and Porto have fewer problems with congestion.

4.3. Bipartite road usage graph

In addition to measuring physical network properties of roads, the system architecture enables detailed analysis of individual road segments and neighborhoods within a city. Though the transient OD matrices constructed by Wang et al. (2012) correlate poorly with OD matrices developed by the methods above and traditional surveys, their work highlights new metrics of road usage patterns that can be measured via these new data sources. To this end we create a bi-partite usage graph. Every time a route between two location is assigned, we traverse the path and keep a record of how many trips from each driver source (census tract) used each road. This record is then used to construct a bipartite graph containing two types of nodes: road segments and driver sources, as shown in Fig. 5. Roads are connected to driver sources that contribute traffic to that segment and census tracts are connected to roads that are used by people who live here.

$$k_s^{road} = \sum_o A_{o\to s}, \quad k_o^{source} = \sum_s A_{o\to s} \tag{2}$$

$$A_{o\to s} = \begin{cases} 1, & \text{if vehicles from tract } o \text{ use road } s \\ 0, & \text{otherwise.} \end{cases}$$

We then examine the degree distributions of roads and census tracts using Eq. (2) in this bipartite graph to reveal patterns of road usage in Fig. 6. The number of roads used by residents of a given location is much more consistent between different cities and appears less affected by the size of the road network. On the other hand, the number of driver sources contributing traffic to a given road segment is broadly distributed, suggesting that most roads are *local* in that they serve only a few

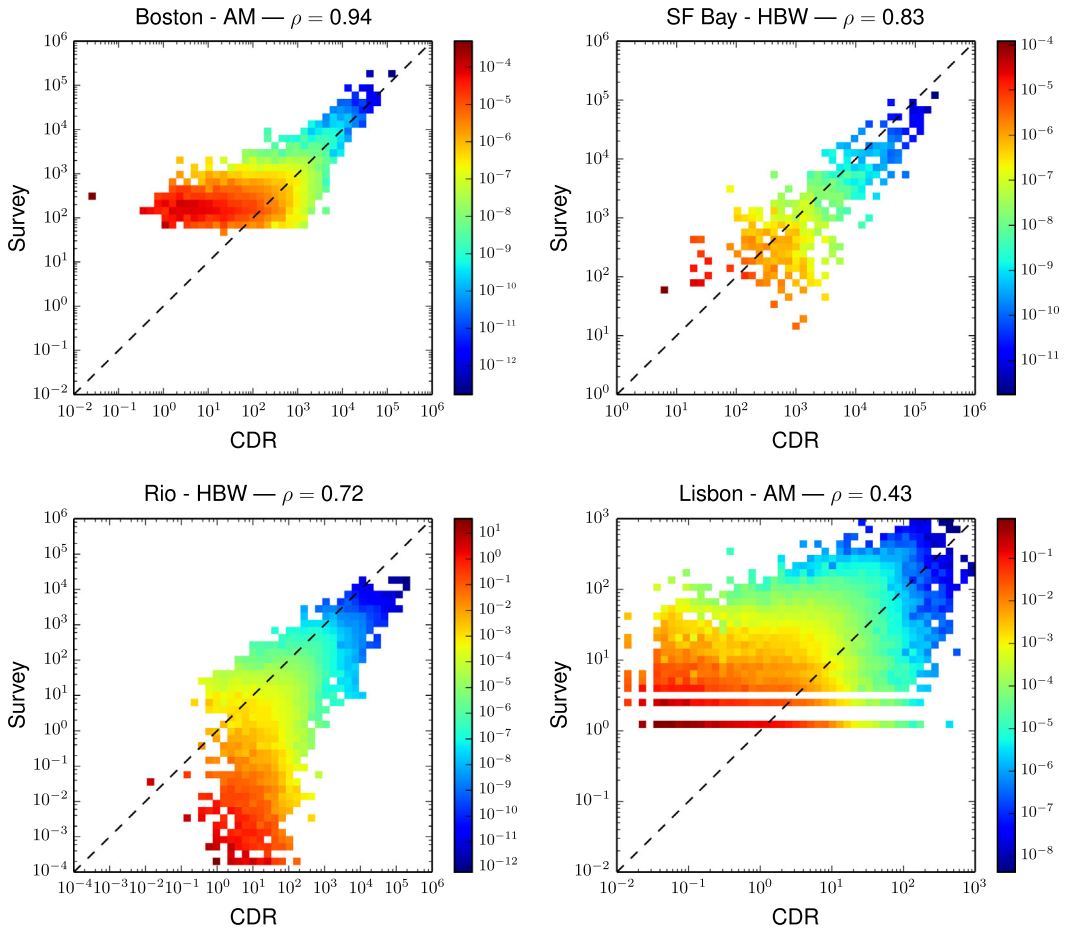

Fig. 3. Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, freguesia-to-freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates (e.g. freguesias in Portugal), correlations are lower. However, more work must be done to understand uncertainties in estimates provided by both models.

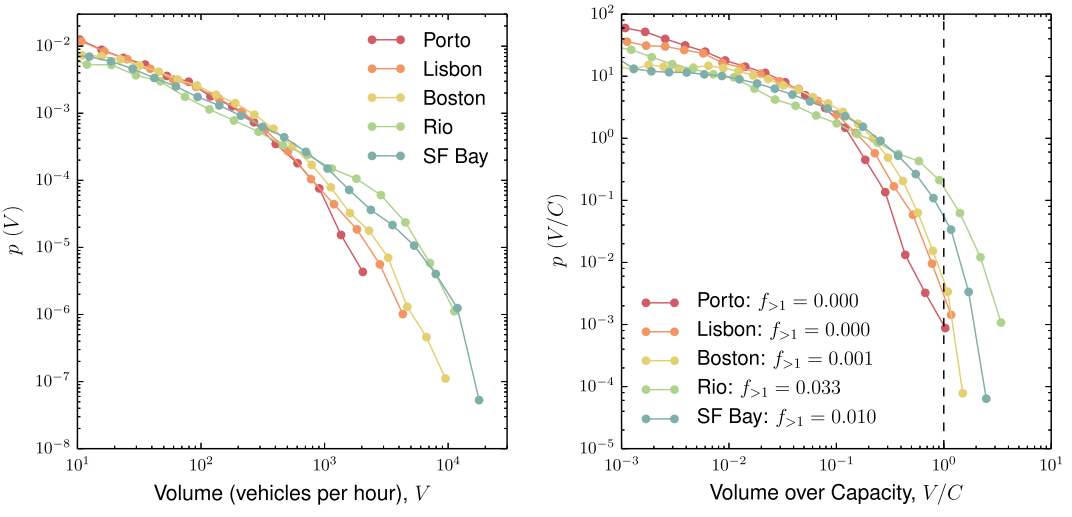

Fig. 4. Distributions of travel volume assigned to a road and the volume-over-capacity (V/C) ratio for the five cities. The values presented in the legend refers to the fraction of road segments with $V/C > 1$.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

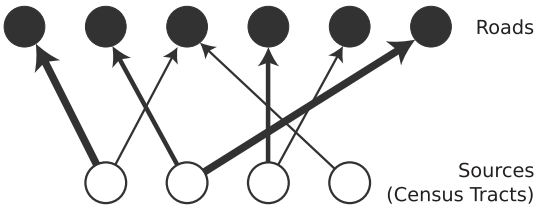

Fig. 5. A graphical representation of the bipartite network of roads and sources (census tracts), with edge sizes mapping the number of users using the connected road in their individual routes.

locations, while a few roads in the tail of the distribution are used for large fractions of the population. While this result is intuitive given that highways are designed for just this purpose, we hope future work explores the relationship between this bipartite usage graph and road network topology further.

An example of such an application was proposed by Wang et al. to classify road segments based on the relationship between topological and demand based metrics. Comparing the topological properties of roads in the physical network to the bipartite usage graph provides insights into their role in the transportation system. Edge betweenness centrality (Newman, 2005) captures the importance of a road by counting how many shortest paths between any two locations σ_{OD} must pass through that edge $\sigma_{OD}(e)$ (Eq. (3)). While this measure captures some aspects of importance, it treats all potential paths as equally likely and tends to be biased towards geographically central links. The degree of a road in the bipartite usage graph reflects the number of locations in the city that actually rely on that road because trips were assigned there from

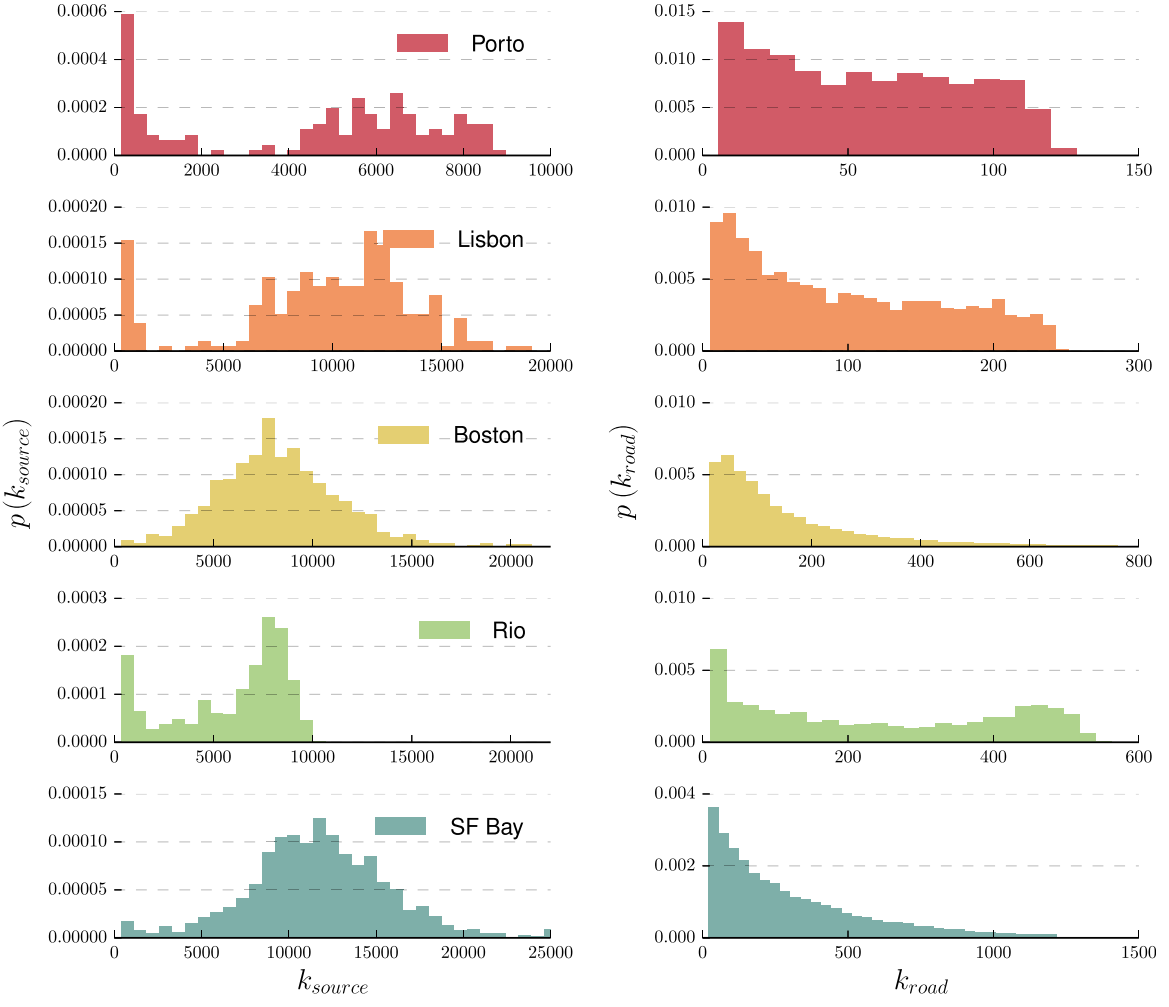

Fig. 6. Distributions of k_{road} and k_{source} for the five cities.

actual travel demand. With these two metrics, betweenness centrality and a roads degree in the usage network, we can classify the role of a road in the cities transportation network.

$$bc_s = \sum_{o,d} \frac{\sigma_{OD}(s)}{\sigma_{OD}} \tag{3}$$

A simple classifier divides the betweenness usage degree space into four quadrants surrounding the point representing the 75th percentile for betweenness centrality and usage degree. Roads with betweenness and usage degree above the 75th percentile are both physical connectors and are used by large portions of the region. These roads tend to be bridges or urban rings. Roads with low betweenness, but high usage degree are attractors, receiving a higher proportion of trips than would be expected assuming uniform demand. Roads with high betweenness and low usage are physical connectors and serve an important purpose geographically, but may not be utilized by actual demand. Other roads, with low betweenness and low usage are local roads and primarily serve populations living and working nearby. Fig. 7 shows each road according to this classification using data from the ODs calculated via mobile phones.

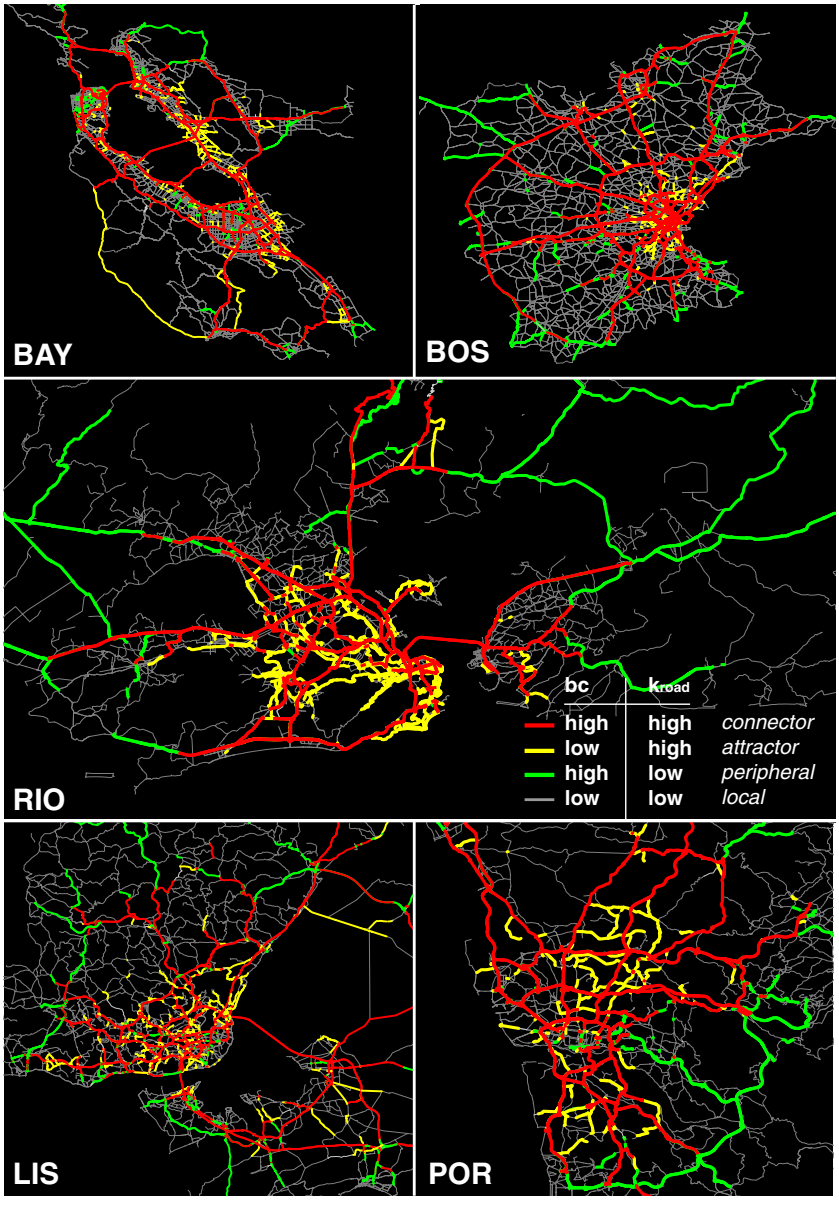

Fig. 7. Maps depicting the proposed road classification, summarized in the legend, for the five subject cities.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

Finally, this bipartite framework of analysis allows us to augment visualizations of congestion maps in two ways. The first focuses on a single road segment. For example, when we identify a segment of a highway that becomes highly congested with traffic jams each day, we can easily query the bipartite graph to obtain a list of census tracts where drivers sitting in that traffic jam are coming from and where they are going to. The census tract nodes can also be given attributes from containing any demographic data a user wishes. With this information, it is possible to identify leverage points where policy makers can offer alternatives to these individuals or even power applications such as car sharing, by notifying drivers that others sharing the same road may be going to and from the same places. Moreover, businesses considering products or services based on who may be driving by or near different locations may find value in these detailed breakdowns.

Rather than selecting a road segment node, we may also select a single census tract, and check its neighbors to construct a list of all roads used by individuals moving to or from that location. For example, for a given neighborhood in a city we can identify all major arteries that serve that local population. This information provides a detailed look at a central location based on how much road usage it induces. Moreover, geographic accessibility, critical to many socio-economic outcomes, can now be measured in locations that were previously understudied.

Fig. 8. Two screen images from the visualization platform. (a) The trip producing (red) and trip attracting (blue) census tracts using Cambridge St., crossing the Charles River in Boston. (b) Roads used by trips generated at the census tract including MIT. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

4.4. Visualization

To help make these results accessible to consumers and policymakers, we build an interactive web visualization to explore road usage patterns in each city. Most GIS platforms can connect directly PostGIS databases to visualize and analyze road networks with our estimated usage characteristics. While these platforms are preferred by advanced users familiar with GIS data, they are opaque to many consumers who may benefit from more detailed information on road usage. A simple API is implemented to query the database and generate standard GeoJSON objects containing geographic information on roads as well as computed metrics such as level of service. We also implement queries to answer questions such as "What are all the census tracts used by drivers on a particular road?" or "What are all roads used by a given location in the city?". These data are then parsed and displayed on interactive maps using any of the available online mapping APIs and D3js allowing users, with functionality that enables one to select individual roads and areas. Two screen images of this system is shown in Fig. 8.

5. Discussion and limitations

This paper has presented a full implantation of a travel demand model that uses new, big data resources as input. We have presented a system that combines and improved upon many disparate advanced in recent years to produce fast, accurate, and inexpensive travel demand estimates. We began by outlining methods to extract meaningful locations from noisy call detail records and estimate origin–destination matrices by counting trips between these places. Normalized and scaled trips counts are compared to estimates made using survey data in both trip tables and at the OD pair level. These flows are then assigned to road networks constructed from OpenStreetMap data using an incremental traffic assignment algorithm. As routes are assigned, a number of metrics on road usage are measured and stored.

While these results show great progress in making big data useful for transportation engineering, there are still limitations inherent in this data and our model. Specifically, we highlight three areas that are ripe for further study.

1. We have shown the level of aggregation applied to OD matrices can affect the correlation observed between model outputs. This is a standard manifestation of the modifiable area unit problem and a more detailed exploration may indicate which levels of analyses were better suited for different data sources. Moreover, a more detailed analysis of uncertainty in model estimates may make it easier to assess their correlation and validity.
2. Our traffic assignment algorithm is efficient, but simple. In the future, a stochastic dynamic user equilibrium assignment methods should be explored and compared. Moreover, route choice modeling may be significantly improved by the availability of high resolution GPS trajectories of drivers. We believe our system's modular design makes it easy to incorporate these new models.
3. Our mode choice model remains simple and will likely require more sophistication for modeling trips not taken in private vehicles. This, combined with improvements in route choice, may make it possible to estimate multi-modal trip demand, as public transportation, bike lanes, and even water transportation networks are included in OpenStreetMap data.

We hope future work will address these and improve further on the methods presented here.

6. Conclusions

Transportation engineers and urban planners have a rich history estimating flows of people within cities and mapping this flow onto transportation infrastructure. However, these efforts are often constrained by limited data resources. The rise of ubiquitous mobile sensors has generated a wealth of new data on human mobility, but new tools must be developed to integrate these data and insights into traditional transportation modeling approaches. To this end, we have demonstrated a full implementation of a travel demand model utilizing mobile phone data as an input. We presented algorithms to generate routable road networks from open source data repositories, generate validated OD matrices and trip tables from CDR data, and route these trips through road networks using a paralleled ITA algorithm. We have demonstrated a number of possible analyses that can be performed on the output of this system including network performance and classification measurements and an online, interactive visualization platform.

As more data becomes available in the form of calls, gps traces, or real time traffic monitoring systems, we are excited at the prospect of updating and improving these systems further.

Acknowledgments

This work was partially funded by the BMW-MIT collaboration under the supervision of PI Mark Leach,[4] the World Bank-HuMNet collaboration agreement under the supervision of PI Shomik Mehndiratta[5] and the Center for Complex

[4] mark.leach@bmw.de.
[5] smehndiratta@worldbank.org.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022

Engineering Systems (CCES) at KACST under the co-direction of Anas Alfaris.[6] We thank Pu Wang for technical support, Shan Jiang for her help obtaining LUT model results for Lisbon, Nelson F.F. Ebecken for support with data, the Rio de Janeiro State Agency (FAPERJ) for the grant on this project, and the Rio City Hall for the support and the data they have provided. Our work was also supported, in part, by the UPS Center for Transportation and Logistics Graduate Research Fellowship awarded to Serdar Çolak and by the National Science Foundation Graduate Research Fellowship awarded to Jameson L. Toole. Lauren P. Alexander and Bradley Sturt are supported by the Austrian Institute of Technology and the MIT-Smart program, respectively.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.trc.2015.04.022.

References

Akcelik, R., 1991. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function. Aust. Road Res. 21 (3).
Alexander, L.P., Jiang, S., Murga, M., González, M.C., 2015. Validation of origin–destination trips by purpose and time of day inferred from mobile phone data. Transport. Res. Part C: Emer. Technol.
Asakura, Y., Hato, E., 2004. Tracking survey for individual travel behaviour using mobile communication instruments. Transport. Res. Part C: Emerg. Technol. 12 (3), 273–291.
Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from israel. Transport. Res. Part C: Emer. Technol. 15 (6), 380–391.
Bast, H., Funke, S., Sanders, P., Schultes, D., 2007. Fast routing in road networks with transit nodes. Science 316 (5824), pp. 566–566.
Belik, V., Geisel, T., Brockmann, D., 2011. Natural human mobility patterns and spatial spread of infectious diseases. Phys. Rev. X 1 (1), 011001.
Bell, M.G., 1991. The estimation of origin–destination matrices by constrained generalised least squares. Transport. Res. Part B: Methodol. 25 (1), 13–22.
Branston, D., 1976. Link capacity functions: a review. Transport. Res. 10 (4), 223–236.
Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439 (7075), 462–465.
Caceres, N., Wideberg, J., Benitez, F., 2007. Deriving origin destination data from a mobile phone network. Intell. Transp. Syst. IET 1 (1), 15–26.
Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr., J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transport. Res. Part C: Emer. Technol. 26, 301–313.
Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A: Math. Theor. 41 (22), 224015.
Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. Transport. Res. Part B: Methodol. 18 (4), 289–299.
Colak, S., Schneider, C.M., Wang, P., González, M.C., 2013. On the role of spatial dynamics and topology on network flows. New J. Phys. 15 (11), 113037.
Çolak, S., Alexander, L.P., Alvim, B.G., Mehndiretta, S.R., González, M.C., 2015. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. Transport. Res. Rec. J. Transport. Res. Board.
Daganzo, C.F., 1980. Optimal sampling strategies for statistical models with discrete dependent variables. Transport. Sci. 14 (4), 324–345.
de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. 3.
Ferreira, J., Diao, M., Zhu, Y., Li, W., Jiang, S., 2010. Information infrastructure for research collaboration in land use, transportation, and environmental planning. Transport. Res. Rec. J. Transport. Res. Board 2183 (1), 85–93.
González, M.C., Hidalgo, C.A., Barabasi, A.-L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.
Hazelton, M.L., 2000. Estimation of origin–destination matrices from link flows on uncongested networks. Transport. Res. Part B: Methodol. 34 (7), 549–566.
Hazelton, M.L., 2001. Inference for origin–destination matrices: estimation, prediction and reconstruction. Transport. Res. Part B: Methodol. 35 (7), 667–676.
Hazelton, M.L., 2003. Some comments on origin–destination matrix estimation. Transport. Res. Part A: Policy Pract. 37 (10), 811–822.
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transport. Res. Part C: Emer. Technol. 40, 63–74.
Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr., J., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM, p. 2.
Lo, H., Zhang, N., Lam, W.H., 1996. Estimation of an origin–destination matrix with random link choice proportions: a statistical approach. Transport. Res. Part B: Methodol. 30 (4), 309–324.
Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proc. Nat. Acad. Sci. 109 (29), 11576–11581.
Lu, C.-C., Zhou, X., Zhang, K., 2013. Dynamic origin–destination demand flow estimation under congested traffic conditions. Transport. Res. Part C: Emer. Technol. 34, 16–37.
Maher, M., 1983. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. Transport. Res. Part B: Methodol. 17 (6), 435–447.
Merchant, D.K., Nemhauser, G.L., 1978. A model and an algorithm for the dynamic traffic assignment problems. Transport. Sci. 12 (3), 183–199.
Newman, M.E.J., 2005. A measure of betweenness centrality based on random walks. Soc. Netw. 27 (1), 39–54.
Nie, Y., Zhang, H., Recker, W., 2005. Inferring origin–destination trip matrices with a decoupled GLS path flow estimator. Transport. Res. Part B: Methodol. 39 (6), 497–518.
Ortúzar, J.D., Willumsen, L.G., 1994. Modelling Transport. John Wiley & Sons, Chichester, England.
Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware map: identifying human daily activity pattern using mobile phone data. Hum. Behav. Underst., 14–25
Reades, J., Calabrese, F., Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environ. Plan. B: Plan. Des. 36 (5), 824–836.
Richardson, A.J., Ampt, E.S., Meyburg, A.H., 1995. Survey Methods for Transport Planning. Eucalyptus Press, Melbourne.
Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unravelling daily human mobility motifs. J. Roy. Soc. Interface 10 (84), 20130246.
Sevtsuk, A., Ratti, C., 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J. Urb. Technol. 17 (1), 41–60.
Smith, M.E., 1979. Design of small-sample home-interview travel surveys. Transport. Res. Rec. 701, 29–35.
Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010a. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.
Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010b. Modelling the scaling properties of human mobility. Nat. Phys. 6 (10), 818–823.

[6] anas@mit.edu.

Spiess, H., 1987. A maximum likelihood model for estimating origin–destination matrices. Transport. Res. Part B: Methodol. 21 (5), 395–412.
Spiess, H., 1990. Technical note – conical volume-delay functions. Transport. Sci. 24 (2), 153–158.
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: where are we going? Transport. Res. Part A: Policy Pract. 41 (5), 367–381.
Van Zuylen, H.J., Willumsen, L.G., 1980. The most likely trip matrix estimated from traffic counts. Transport. Res. Part B: Methodol. 14 (3), 281–293.
Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Sci. Rep. 2 (1001). http://dx.doi.org/10.1038/srep01001.
Wardrop, J.G., 1952. Road paper. some theoretical aspects of road traffic research. ICE Proceedings: Engineering Divisions, vol. 1. Thomas Telford, pp. 325–362.
Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O., 2012. Quantifying the impact of human mobility on malaria. Science 338 (6104), 267–270.
Zhan, X., Hasan, S., Ukkusuri, S.V., Kamga, C., 2013. Urban link travel time estimation using large-scale taxi data with partial information. Transport. Res. Part C: Emer. Technol. 33, 37–49.
Zheng, Y., Xie, X., 2011. Learning travel recommendations from user-generated GPS traces. ACM Trans. Intell. Syst. Technol. (TIST) 2 (1), 2.

Please cite this article in press as: Toole, J.L., et al. The path most traveled: Travel demand estimation using big data resources. Transport. Res. Part C (2015), http://dx.doi.org/10.1016/j.trc.2015.04.022