Available online at www.sciencedirect.com

SciVerse ScienceDirect

Transportation Research Procedia 00 (2015) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

21st International Symposium on Transportation and Traffic Theory, ISTTT21 2015, 5-7 August 2015, Kobe, Japan

Cellpath: fusion of cellular and traffic sensor data for route flow estimation via convex optimization

Cathy Wu [a,*], Jérôme Thai [a], Steve Yadlowsky [a], Alexei Pozdnoukhov [b], Alexandre Bayen [a,b,c]

[a] *Electrical Engineering and Computer Sciences, UC Berkeley, 652 Sutardja Dai Hall, Berkeley, CA 94720, USA*
[b] *Civil and Environmental Engineering, UC Berkeley, 109 McLaughlin Hall, Berkeley, CA 94720, USA*
[c] *Institute for Transportation Studies (ITS), UC Berkeley, 109 McLaughlin Hall, Berkeley, CA 94720, USA*

Abstract

A new convex optimization framework is developed for the route flow estimation problem from the fusion of vehicle count and cellular network data. The issue of highly underdetermined link flow based methods in transportation networks is investigated, then solved using the proposed concept of *cellpaths* for cellular network data. With this data-driven approach, our proposed approach is versatile: it is compatible with other data sources, and it is model agnostic and thus compatible with user equilibrium, system-optimum, Stackelberg concepts, and other models. Using a dimensionality reduction scheme, we design a projected gradient algorithm suitable for the proposed route flow estimation problem. The algorithm solves a block isotonic regression problem in the projection step in linear time. The accuracy, computational efficiency, and versatility of the proposed approach are validated on the I-210 corridor near Los Angeles, where we achieve 90% route flow accuracy with 1033 traffic sensors and 1000 cellular towers covering a large network of highways and arterials with more than 20,000 links. In contrast to long-term land use planning applications, we demonstrate the first system to our knowledge that can produce route-level flow estimates suitable for short time horizon prediction and control applications in traffic management. Our system is open source and available for validation and extension.

Peer-review under responsibility of the Scientific Committee of ISTTT21.

Keywords: Route flow estimation; cellular network data; convex optimization; traffic assignment; simplex constraints; isotonic regression

1. Introduction

While there is a wealth of literature in transportation science that is aimed at modeling, computing, and estimating the movement of people in terms of *link* flows and *origin-destination (OD)* flows, there is relatively little work focused on *route* flow estimation. The route flow estimation problem is particularly important because route flow estimates can capture phenomena in traffic behavior that link flows and OD flows (also called OD demands) cannot. For instance, route flows would enable analysis and re-routing of commuters who would be most affected by a link closure. Additionally, route flows provides a rich state estimate of the network which may be used to compute link

* Corresponding author. Tel.: +1-510-642-3585; fax: +1-510-643-3955.
E-mail address: cathywu@eecs.berkeley.edu

2214-241X
Peer-review under responsibility of the Scientific Committee of ISTTT21.

flows, OD flows, turning ratios, etc., thereby providing backwards compatibility with past and ongoing work that builds upon those estimates.

Simultaneously accurate and efficient methods for estimating route flows are crucial for large scale urban network analysis and planning demands. However, the first step for many approaches to estimating route flow requires enumerating all feasible routes, which is an unreasonable task for many urban road networks because it may require exponential time to compute (Ford & Fulkerson, 1962, §1.2). Classically, the set of potential routes may be extracted from the induced equilibrium in network flow models. At the cost of restrictive assumptions, *deterministic user equilibrium* (UE) (Wardrop & Whitehead, 1952) permits the modeling of unique link flows and feasible route (or path) flows without requiring full route enumeration (Sheffi, 1985, §3.3), (Bell & Iida, 1997, §5.2)). The *stochastic user equilibrium* (SUE) (probit-based (Daganzo & Sheffi, 1977; Maher & Hughes, 1997) and logit-based (Fisk, 1980; Bell & Iida, 1997)) addresses some of the shortcomings of the UE by modeling imperfect knowledge of the network and variation in drivers' preferences, which makes the estimation of route flows possible (Bell *et al.* , 1997). However, frequent perturbations in traffic networks indicate that real-world transportation networks may not be in equilibrium (or only approximately so) (Hato *et al.* , 1999), so we develop a data-driven framework that focuses on effectively utilizing the large amount of data available for estimation in traffic networks. Indeed, in recent years, the growing number of mobile sensors in urban areas enables the use of probe vehicles for route inference from GPS traces (Hunter *et al.* , 2009; Rahmani & Koutsopoulos, 2013).

1.1. Traffic data sources

Traditional traffic sensing systems such as loop detectors embedded in the pavement and cameras provide accurate volume and speed estimates, but their placements are typically sparse and their information content is too coarse. Most importantly, they measure total counts of vehicles passing through a road segment without distinguishing between vehicles following different routes. In order to partially address the shortage of information on the routes followed by vehicles, other types of static sensors have been deployed on the road network: cameras that measure split ratios at different intersections (Veeraraghavan *et al.* , 2003) and plate scanning systems (Castillo *et al.* , 2008, 2010). However these systems require costly infrastructure and only provide highly localized traffic information. Meanwhile, given the large penetration of mobile phones among the driving population and the ubiquitous coverage of service providers in urban areas, mobile phones have become an increasingly popular source of location data for the transportation community. For example, dynamic probing by means of in-car GPS traces (Work *et al.* , 2008; Herrera *et al.* , 2009; Hunter *et al.* , 2009) is a promising technology for trajectory recovery and travel time estimation. However, due to the read-only nature of GPS signals, the low penetration rate of GPS-enabled devices running a dedicated sensing application currently limits the ability to accurately estimate traffic volumes, and it is also unlikely that such data would become available to public agencies (Patire *et al.* , 2013).

Cellular network data, in contrast to GPS traces, benefit from dedicated communication between mobile phones and cellular network network base stations, and the (coarse) location data are available directly from cellular communication network operators. Cellular network infrastructures record a variety of phone to cell communication events, such as *handovers* (HO), *location updates* (LU) and *call detail records* (CDR) (Volinsky *et al.* , 2011a,b), and this data has already been shown to be effective in studying urban environments (Candia *et al.* , 2008a; Jiang *et al.* , 2013; Toole *et al.* , 2012). Since typical cellular networks in urban areas include thousands of cells, HO/LU/CDR events are dense enough to be used effectively to estimate the route choice of agents without requiring any additional infrastructure. When an agent is moving, HOs transfer ongoing calls or data sessions from one cell to another without disconnecting the session, and LUs allow a mobile device to inform the cellular network when the device move from one location (or cell) to the next. CDRs (mainly used by service providers for billing purposes) contain timestamped summaries of the cell through which each data transmission came, and therefore contain abundant mobility traces for a majority of the population. Due to the granularity of sensing, these records alone are not sufficient for recovering agent routes precisely. The spatial resolution of CDR, HO, and LU data varies with the density of antennas and is roughly proportional to the daytime population density. In the present work, we use a standard localization approach when dealing with cellular data based on Voronoi tessellation, a simple model solely based on the locations of the cell towers (Baert & Seme, 2004; Candia *et al.* , 2008b).

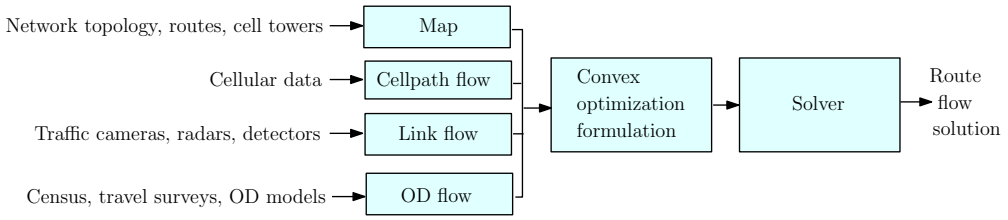

Fig. 1. Proposed route flow estimation pipeline, from raw data to route flows, including: 1) a scheme for route selection and resolution cellular and road networks into a unified map; 2) a trip analysis step to filter driver cellular traces from other traces and infer cellpath flows; 3) an aggregation of the link flow obtained from static sensors over a sizable duration (*e.g.* 1 hour) suitable to address the static estimation problem; 4) a state-of-the-art OD matrix estimation method; 5) a problem formulation that handles data fusion from disparate sources; and 6) the route inference method.

1.2. Related work

Several problems within traffic estimation have already benefited from incorporating data from cellular networks: OD matrix computation using cell phone location data (Caceres *et al.* , 2007; Calabrese *et al.* , 2011) such as CDRs (White & Wells, 2002), link flow estimation (Yadlowsky *et al.* , 2014), and travel time and type of road congestion (Janecek *et al.* , 2012). These studies vary in scale and assumptions, but they indicate the promise of non-pervasive sensing to provide a richer understanding of mobility. In particular, cellular network data has been used to improve the accuracy of OD matrix estimation (Caceres *et al.* , 2007; Calabrese *et al.* , 2011). There are many surveys on the subject in the past decades (Bell & Iida, 1997; Abrahamsson, 1998; Ortuzar & Willumsen, 2001), and the accuracy of OD estimates will continue to improve. Additionally, convex optimization techniques have been used quite frequently by the transportation community for diverse purposes, including several of these problems. For example, the classical Wardrop equilibrium approach to the traffic assignment problem can be formulated as a convex optimization program given some typical assumptions on the link performance (or delay) functions (Sheffi, 1985). Recent works often combine convex optimization with machine learning techniques (Blandin *et al.* , 2009; Shen & Wynter, 2012; Mardani & Giannakis, 2013).

An early study on the use of cellular network data for traffic assignment (Tettamanti *et al.* , 2012) estimates the route choice for each user in the cellular network using a distance measure to determine the best matching route. Their small experiment (2-4 routes) performed via a macro-simulator indicates the potential of cellular network data for solving this problem. However, a recent survey on the use of wireless signals for road traffic detection (Mathew & Xavier, n.d.) concludes that there is thus far no existing system that can estimate traffic densities in a practical sense, that is, in terms of scalability, coverage, cost, and reliability, thus motivating our work on estimating route flows.

1.3. Contributions of this article

One of the key innovations of the present work is generalizing the common notion of an OD matrix to a general form of coarse (route) flow measurements (here collected from cellular network data). As mentioned above, the problem of traffic assignment is historically highly underdetermined because the OD matrix and link flows (even when all the links are observed) contains relatively little information as compared to the number of available routes. We introduce the concept of *cellpaths*, which generalizes 2-point network flow, which we call *OD flow*, to n-point network flow, which we call *cellpath flow*. Where OD flow is characterized by two centroids (illustrated in Figure 3), cellpath flow can be characterized by n region centroids through with vehicles pass on a single trip. In this article, the centroids for cellpath flow correspond to cellular base stations, and the centroids for OD flows correspond to centroids of Traffic Analysis Zones (TAZ). Since our approach includes a "strict" generalization of ODs to cellpaths, the methodology presented in this article can be applied to a variety of traffic modeling and estimation problems.

Now, we define our problem as follows: given a large-scale road network in the quasi-static regime, a set of OD demands, a set of admissible routes between each OD pairs, cellpath flow measurements along the network, and link flow measurements on a subset of links in the network, our goal is to develop a method to estimate the distribution of flow over the set of routes. We pose the route flow estimation problem as a mathematical program optimizing the fit to link sensor data over feasible route flow distributions, constrained to those which are consistent with measured cellpath flows in the network.

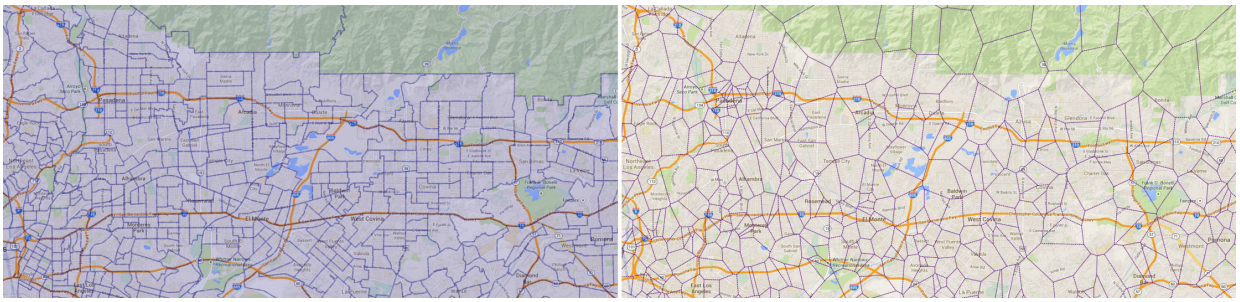

Fig. 2. I-210 corridor in Los Angeles county used for the numerical work presented in §5. Left subfigure: The 700 regions are origin/destinations areas called Traffic Analysis Zones (TAZ) used for the numerical experiments. Right subfigure: Corresponding Voronoi partition of the cellular network based on 1000 cell towers. Best viewed in color.

Our analysis of the structure of the constraints in the optimization program allows us to present a more efficient solution method that scales to full-sized networks. By recognizing the constraints as block-simplex constraints, we apply a standard equality constraint elimination technique (Boyd & Vandenberghe, 2004, §4.2.4) with a particular change of variables to convert the non-negativity constraints on the variables into ordering constraints. In the new space induced by the change of variables, we show that the projection on the feasible set (characterized by the ordering constraints) can be performed in linear time via bounded isotonic regression (see (Tibshirani *et al.* , 2011) for a short survey on isotonic regression), where n is the number of routes per OD pair. This is an improvement over the $O(n \log n)$ time required by the projection onto the simplex (J. Duchi, 2008; Wang & Carreira-Perpin, 2013). Then we solve our convex optimization program with a first-order projected descent algorithm. The change of variables presents two main advantages: our projection requires $O(n)$ time and the dimensionality is reduced (sometimes by a factor of $1/3$), which is critical for large-scale problems. In addition, it is worth noting that a wide variety of problems can benefit from this methodology. First, the use of algorithms that feature a projection step, e.g. projected descent methods and alternating direction methods, is very popular since they often provide a simple and efficient way to solve constrained convex optimization problem as opposed to more specialized active set methods. There is also a great deal of applications that feature simplex constraints, such as the aforementioned traffic assignment problem, many game theory settings for the computation of strategy distributions, and ℓ_1-based approach in machine learning (J. Duchi, 2008).

Practical considerations that traffic flow in urban areas may not be in equilibrium motivate our emphasis on a data-driven approach that benefits from the sheer amount of cellular network data without relying on equilibrium-based models or other route choice models. Aiming at a real-world production system pipeline summarized in Figure 1, we demonstrate the versatility and data-driven nature of the proposed approach via validation on three datasets produced by two simulators of route assignment, where the positions of the cell towers are sampled randomly on the urban networks. To assess our approach on a variety of possible route choice models that may be realized in a real-world setting, we develop a small equilibrium-based model that generates user equilibrium (UE) and system-optimal (SO) flows on the I-210 corridor near Los Angeles, CA. We also use MATSim agent-based transport simulator[1] on a large-scale urban road network near Los Angeles, CA (with more than 20K links and 290K routes) to showcase the performance of our methodology on large datasets. We demonstrate that our full pipeline, from the simulators to the procedures to estimate static route flows on small and large-scale urban networks, can be extended easily to incorporate other types of data such as link capacities, split ratios etc. Hence we hope that our framework will be a benchmark for many future studies of estimation problems in transportation science.[2]

We summarize the contributions of the presented work:

- We propose a convex optimization formulation for the route flow estimation problem which uses a new data fusion approach for loop detectors counts and cellular signal traces (ubiquitous among the driving population).

[1] MATSim is an open source project: `http://www.matsim.org/publications`.
[2] Our full system is open source and available at `https://megacell.github.io/` for validation and extension.

Table 1. Notation for route estimation problem. We have m observed links, q cellpaths, n routes.

Notation	Description	Notation	Description
$\mathcal{O}$, $\mathcal{D}$	Set of origins/destinations $\mathcal{D} = \mathcal{O}$	$d \in \mathbb{R}_+^{\lvert\mathcal{O}\rvert^2}$	Vector of OD flows, $d = (d_k)_{k\in\mathcal{O}^2}$
$\mathcal{L}$	$\lvert\mathcal{L}\rvert = m$, links with observed flow	$b \in \mathbb{R}_+^{\lvert\mathcal{L}\rvert}$	Observed link flow vector, $b = (b_l)_{l\in\mathcal{L}}$
$\mathcal{P}$	$\lvert\mathcal{P}\rvert = q$, observed cellpaths	$f \in \mathbb{R}_+^{\lvert\mathcal{P}\rvert}$	Cellpath flows vector, $f = (f_p)_{p\in\mathcal{P}}$
$\mathcal{R}$	$\lvert\mathcal{R}\rvert = n$, set of routes	$x \in \mathbb{R}_+^{\lvert\mathcal{R}\rvert}$	Vector of route flows $x = (x_r)_{r\in\mathcal{R}}$
$\mathcal{A}$	Set of all links in the network	$v \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Full link flow vector, $v = (v_a)_{a\in\mathcal{A}}$
$A \in \{0,1\}^{\lvert\mathcal{L}\rvert\times\lvert\mathcal{R}\rvert}$	Link-route incidence matrix	Subset $\mathcal{R}^p$	Subset of $n_p := \lvert\mathcal{R}^p\rvert$ routes with cellpath p
$A^{\text{full}} \in \{0,1\}^{\lvert\mathcal{A}\rvert\times\lvert\mathcal{R}\rvert}$	Full link-route incidence matrix	$\tilde{x}^p \in [0,1]^{\lvert\mathcal{R}^p\rvert}$	Ratios of flows across routes $r \in \mathcal{R}^p$
$U \in \{0,1\}^{\lvert\mathcal{P}\rvert\times\lvert\mathcal{R}\rvert}$	cellpath-route incidence matrix	$x^p \in \mathbb{R}_+^{n_p}$	x_r^p is the flow of route $r \in \mathcal{R}^p$
$T \in \{0,1\}^{\lvert\mathcal{O}\rvert^2\times\lvert\mathcal{R}\rvert}$	OD-route incidence matrix	$\mathcal{R}^k \subset \mathcal{R}$	Subset of n_k routes between OD pair k

- We demonstrate that our formulation is also compatible with several other approaches to this problem, including equilibrium concepts, which may be used in conjunction for improved estimation.
- We introduce the concept of *cellpaths* and demonstrate its application to traffic estimation problems. We address the issues with highly underdetermined link flow based methods (which was already raised in the traffic assignment literature) by formalizing cellular data as cellpaths and incorporating them as constraints. Though we focus on the route flow estimation problem, many traffic problems may benefit from such an approach.
- Using a reduction scheme, we design an algorithm to solve the route flow estimation problem and large-scale traffic assignment problems in general. In the resulting formulation, the projection step can be performed in $O(n)$ via isotonic regression, an improvement over $O(n \log n)$, where n is the number of routes per OD pair.
- We present a full system pipeline from cellular network and link flow data to estimate the static route flow (and as a by-product, link flow) on a large-scale urban network. We demonstrate the first system to our knowledge that can produce route-level flow estimates suitable for short time horizon prediction and control applications in traffic management from the fusion of cellular network data and data from static sensors along roads.
- We present numerical results from different sets of small and large-scale datasets for the Greater Los Angeles Area (see Figure 2). In particular, the emphasis is placed on a data-driven approach: it is versatile to different types of underlying agent behavior models.

The remainder of the article is organized as follows: In Section 2, we present the setup and assumptions of our work, then formulate our route estimation problem in the framework of convex optimization. We also provide a re-formulation necessary for the algorithmic approach described in Section 3. Further in Section 3, we develop a specialized projected gradient method to solve convex optimization programs with simplex constraints. Section 4 is dedicated to the setting of our experiments. Section 5 presents our numerical results. Section 6 concludes the paper by placing the presented method within a general data-driven traffic estimation framework and identifying directions for future work.

2. Problem formulation

2.1. Problem setup and assumptions

We define the terminology used in the article, the notation is presented in Table 1, and the setup is illustrated in Figure 3. It is important to distinguish between four types of flows: cellpath flow, link flow, route flow, and OD flow. Our setup consists of:

- **Origins**: traffic regions each with an associated centroid, defined by a partitioning of the *road network*. Each region is both an *origin* (its centroid is a source from which trips emanate) and a *destination* (its centroid is a sink at which trips terminate). To demonstrate a possible implementation, the numerical work in this article

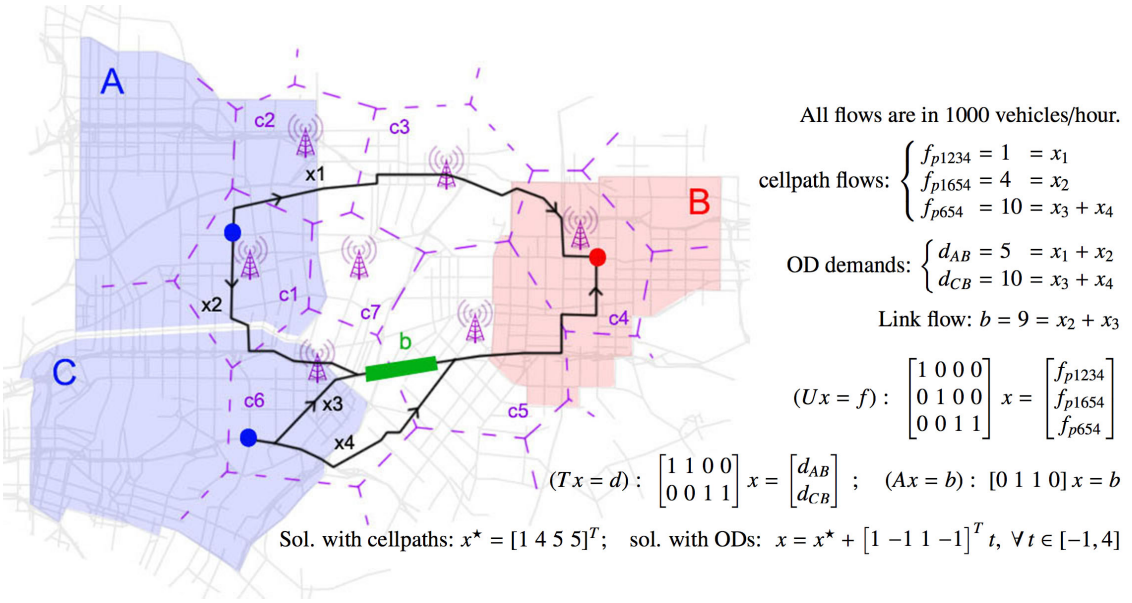

Fig. 3. In this illustration of the cellular and loop data fusion, we have two origins A and C (the blue traffic regions and their centroid as blue dots) and one destination B (the red traffic region). We have routes r_1, r_2, r_3, r_4 with flows $x = (x_1, x_2, x_3, x_4)$ such that r_1, r_2 go from A to B and r_3, r_4 go from C to B. Cells $c_1, \cdots, c_7$ are shown in purple dashed regions. Since route r_1 goes through cells c_1, c_2, c_3, c_4, its associated cellpath is p_{1234}. Similarly, routes r_2, r_3, r_4 have cellpaths $p_{1654}, p_{654}, p_{654}$ respectively. Let f_{p1234}, f_{p1654}, f_{p654} be the cellpath flows (obtained from cellular network data), *i.e.* there are f_{p1234}=1000 veh/h going through c_1, c_2, c_3, c_4. Let d_{AB} and d_{CB} be the OD demands. Cellpaths p_{1234} and p_{1654} disambiguate routes between AB: $f_{p1234} = x_1$, $f_{p1654} = x_2$, contrary to the ODs: $d_{AB} = x_1 + x_2$. However, cell towers are not dense along r_3, r_4, hence $d_{CB} = f_{p654} = x_3 + x_4$. The cellpath-route incidence matrix generalizes OD matrices since we consider the sequence of intermediate regions (cells here) that intersect with trips. We also have $x_2 + x_3 = b$, with b the flow on the green link (from loop detectors). There is a unique route flow inducing flows b, f_{p1234}, f_{p1654}, f_{p654} that is $x^\star = [1\ 4\ 5\ 5]$, while there are infinitely many flows inducing b, d_{AB}, d_{CB}: $x = x^\star + [1\ -1\ 1\ -1]^T t$, $\forall t \in [-1, 4]$, so the problem has *one degree of freedom* and is *underdetermined* with only the OD demands as data. Best viewed in color.

uses the Traffic Analysis Zones (TAZ) (see Figure 2) as origins/destinations. We define *OD flow* to be the flow (vehicle count per time) that originates and terminates with an OD pair.

- **Cells**: regions defined by the Voronoi partition of the locations of the *cellular network base stations*; they are generally a different set of regions from the origins.
- **Cellpath**: a sequence of cells, and we define *cellpath flow* to be the flow along a cellpath.
- **Link**: a segment of road in the network, and the *link flow* is the flow through a link.
- **Route**: a sequence of links from an origin to a destination. Each route also has a particular associated cellpath, as well as a particular associated OD; this insight is important for the structure of our convex optimization formulation. The *route flow* is the flow on the route.

The link-route incidence matrix A encodes the network topology (which routes $r \in \mathcal{R}$ contains which links $l \in \mathcal{L}$); the cellpath-route incidence matrix U encodes the collection of routes with the same cellpaths (which routes $r \in \mathcal{R}$ is associated to which cellpath $p \in \mathcal{P}$); and the OD-route incidence matrix T encodes which routes $r \in \mathcal{R}$ is between OD pairs $k \in \mathcal{O}$[2].[3]

$$\text{link-route: } A_{lr} = \begin{cases} 1 & \text{if } l \in r \\ 0 & \text{else} \end{cases}; \quad \text{cellpath-route: } U_{pr} = \begin{cases} 1 & \text{if } r \in \mathcal{R}^p \\ 0 & \text{else} \end{cases}; \quad \text{OD-route: } T_{kr} = \begin{cases} 1 & \text{if } r \in \mathcal{R}^k \\ 0 & \text{else} \end{cases} \tag{1}$$

The model assumptions are as follows:

- We consider a quasi-static setting, where traffic demands (flows) remain constant over time, and we focus on the noiseless case, with a short commentary on the noisy case in Section 5.

[3] The lowercase letters l, r, p, k written as subscripts refer to the indices associated to links, routes, cellpaths, and ODs respectively.

- Since enumerating all routes is not tractable, we consider the top routes between each OD pair following different criteria depending on the setting of the numerical experiment (see Section 4).
- We can reliably determine the cellpath flow f_p from the cellular traces along each cellpath p.
- All cellpaths $p \in \mathcal{P}$ are *contiguous*: each pair of consecutive cells in p shares a boundary.
- The set of cellpaths $\mathcal{P}$ is *well-posed*: each route $r \in \mathcal{R}$ corresponds to exactly 1 cellpath $p \in \mathcal{P}$, and we have a cellpath flow measurement f_p for each $p \in \mathcal{P}$.

2.2. Formulation and analysis of the model

The fusion of cellular and loop data for route flow estimation is one of the key contributions of this article. We pose the route flow estimation problem as a mathematical program optimizing the fit to link sensor data over feasible route flow distributions, constrained to those which are consistent with measured cellpath flows in the network. We formulate this in the framework of convex optimization as a minimization of a quadratic program:

$$\begin{array}{l} \min \ \frac{1}{2}\|Ax - b\|_2^2 \\ \text{s.t. } Ux = f, \ x \geq 0 \end{array} \tag{2}$$

The problem is a constrained linear inverse problem in which we want to estimate a signal of length n (the route flows) given that we have m measurements (the observable link flows). We additionally have q cellpath flow constraints: for each cellpath $p \in \mathcal{P}$, there are n_p routes corresponding to p, such that their flow must sum up to the cellpath flow f_p:

$$Ux = f \ : \quad \sum_{r \in \mathcal{R}^p} x_r^p = f_p \quad \forall \, p \in \mathcal{P} \tag{3}$$

We note that the subsets of routes $\mathcal{R}^p$ are disjoint (each route has at most one cellpath associated to it), hence (3) along with the nonnegativity constraint in (2), together forms a block simplex constraint, which we further analyze in §3.

In general, $m \ll n$ and $q \leq n$, thus typically the Hessian A^TA of our convex quadratic objective is singular ($A^TA \in \mathbb{R}^{n \times n}$ but rank $(A^TA) \leq m \ll n$). Thus the problem might have multiple optimal solutions (underdetermined) or might have more observations than unknowns (overdetermined), depending on the number of cellpath flow constraints. Our cellpath formulation encodes more constraints than methods that consider less detailed flow measurements (e.g. OD flow), thereby constraining the solution space. Moreover, when there are uncorrelated measurement errors on the vector flow b (absence of interactions between the detection process of the link sensors), the ordinary least squares is the best unbiased estimator of the route flow.[4]

We now make the following observation, which is key for our algorithm described in §3.

Proposition 1. Problem (2) can be reduced to a least-squares problem with (separable) standard simplex constraints:

$$\begin{array}{l} \min \ \frac{1}{2}\|\tilde{A}\tilde{x} - b\|_2^2 \\ \text{s.t. } \mathbf{1}^T\tilde{x}^p = 1, \ \tilde{x}^p \geq 0, \quad \forall \, p \in \mathcal{P} \end{array} \qquad \text{where} \quad \tilde{A} \in \mathbb{R}_+^{|\mathcal{L}| \times |\mathcal{R}|} \ : \ \tilde{A}_{lr} = \begin{cases} f_p & \text{if } l \in r \in \mathcal{R}^p \\ 0 & \text{else} \end{cases} \tag{4}$$

where $\mathbf{1} = [1, \cdots, 1]^T \in \mathbb{R}^{n_p}$ and $\tilde{A}$ is a modified link-route incidence matrix containing the cellpath flows f_p.

Proof: The constraints $Ux = f$ in (2) can be written explicitly: $\sum_{r \in \mathcal{R}^p} x_r^p = f_p$, $\forall \, p \in \mathcal{P}$. With the change of variables $\tilde{x}^p := x^p / f_p$ for all p, the constraints become $\sum_{r \in \mathcal{R}^p} \tilde{x}_r^p = 1$, $\forall \, p \in \mathcal{P}$, or in matrix form: $\mathbf{1}^T\tilde{x}^p = 1$, $\forall \, p \in \mathcal{P}$. Since $f_p > 0$ for all p, then the inequalities $x^p \geq 0$ are equivalent to $\tilde{x}^p = x^p / f_p \geq 0$. Finally, the vector Ax has entries $v_l = \sum_{r \, : \, l \in r} x_r$ for $l \in \mathcal{L}$, where v_l is the flow on link l. The sum can be decomposed between the different cellpaths p: $v_l = \sum_{r \, : \, l \in r} x_r = \sum_p \left\{ \sum_{r \, : \, l \in r \in \mathcal{R}^p} x_r^p \right\} = \sum_p \left\{ \sum_{r \, : \, l \in r \in \mathcal{R}^p} f_p \tilde{x}_r^p \right\} = (\tilde{A}\tilde{x})_l$, hence the objectives are the same. □

[4] The errors must also have zero-mean and constant variance, then the result holds as link flows linearly depend on route flows: $\hat{b} = Ax + \epsilon$, from the Gauss-Markov theorem.

2.3. Compatibility of our formulation

Our formulation is related to the *traffic assignment problem* (also called the route assignment problem) used to solve traffic equilibrium problems (Wardrop & Whitehead, 1952), (Sheffi, 1985, §3), (Bell & Iida, 1997, §5), where $\mathcal{A}$ is the set of all links (arcs) in the network, $A^{\text{full}} \in [0,1]^{|\mathcal{A}|\times|\mathcal{R}|}$ is the full link-route incidence matrix, and ϕ is the Beckmann objective function (Beckmann *et al.* , 1956):

$$\min\ \phi(A^{\text{full}}x) \quad \text{s.t.} \quad Tx = d,\ x \geq 0 \tag{5}$$

This is a standard formulation in traffic assignment in which a local minimum of (5) is a Wardrop equilibrium of a congestion game (Monderer & Shapley, 1996); the relation between OD and route flows given by $Tx = d$ is equivalent to the "approach proportions" formulation of Bar-Gera (Bar-Gera, 2002). If the cellpath-route incidence matrix U is reduced to an OD-route incidence matrix (see Fig. 3), both (2) and (5) share the same constraints. Furthermore, the cellpath constraints can be added to (5) to restrict its solution space as well. The main difference lies in the minimization objective: in (2) it is the link flows measurement residual while in (5) the Beckmann objective ϕ expresses the incentives of all vehicles (or players) to take the shortest route.

In the context of game theory, each cellpath can be seen as a player who chooses a strategy or a probability distribution with weights $(\tilde{x}_r^p)_{r\in\mathcal{R}^p}$ over the n_p routes, and a set defined by $S^p := \{\tilde{x}^p \in [0,1]^{n^p} \mid \sum_{r\in\mathcal{R}^p} \tilde{x}_r^p = 1\}$ is a *strategy set* or a *probability simplex* over the routes $r \in \mathcal{R}^p$.

We observe that the traffic assignment problem (5) can also be reduced in a similar fashion, where O^2 is the set of all OD pairs:

$$\begin{array}{l}\min\ \phi(\tilde{A}^{\text{full}}\tilde{x}) \\ \text{s.t.}\ \ \mathbf{1}^T\tilde{x} = 1,\ \tilde{x}^k \geq 0,\ \forall\, k \in O^2\end{array} \qquad \text{where} \quad \tilde{A}^{\text{full}} = \begin{cases} d_k & \text{if } l \in r \in \mathcal{R}^k \\ 0 & \text{else} \end{cases} \tag{6}$$

Our formulation in (2) is also compatible with several other types of data, *e.g.* turning ratios, link capacities, OD flows. We demonstrate this by augmenting our problem formulation with turning ratios as follows. If, at some node (intersection) $j \in \mathcal{N}$, we know the flow of vehicles coming from link $a = (i,j) \in \mathcal{A}$ and turning into link $a' = (j,k)$, we denote the pair of successive links by $t = (a,a')$, denote $\mathcal{T}$ the set of monitored traffic turns (intersections), let $G \in \{0,1\}^{|\mathcal{T}|\times|\mathcal{R}|}$ be the turn-route incidence matrix, and denote h the vector of flow that passes through each monitored intersection. Then, the objective of (2) can be generalized to include turning ratios:

$$\begin{array}{l}\min\ \frac{1}{2}\|A'x - b'\|_2^2 \\ \text{s.t.}\ \ Ux = f,\ x \geq 0\end{array} \qquad \text{where} \quad A' = \begin{bmatrix} A \\ G \end{bmatrix},\quad b' = \begin{bmatrix} b \\ h \end{bmatrix} \quad \text{and} \quad G_{tr} = \begin{cases} 1 & \text{if} \quad t = (a,a') \,:\, a,a' \in r \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Similarly, the objective of (2) can be generalized to include OD flows, and we later demonstrate the the incorporation of this information in our numerical experiments:[5]

$$\min\ \frac{1}{2}\|A'x - b'\|_2^2 \quad \text{s.t.} \quad Ux = f,\ x \geq 0 \qquad \text{where} \quad A' = \begin{bmatrix} A \\ T \end{bmatrix} \quad \text{and} \quad b' = \begin{bmatrix} b \\ d \end{bmatrix} \tag{8}$$

Suppose we know the link capacities $\tilde{m}_a$, then the constraints $A^{\text{full}}x \preceq \tilde{m}$, where $\tilde{m} := (\tilde{m}_a)_{a\in A}$ is the link capacities vector, can be added to program (2). To approximate the new problem as a program with simplex constraints, we can make the added constraints implicit in the objective:

$$\min\ \frac{1}{2}\|Ax - b\|_2^2 + \sum_{a\in\mathcal{A}} \Phi(L_a^T x - \tilde{m}_a) \quad \text{s.t.} \quad Ux = f,\ x \geq 0 \tag{9}$$

where the barrier Φ is an approximation of the indicator function of R_- given as $I_-(u) = (0 \text{ if } u \leq 0 \text{ else } \infty)$, and the vectors L_a^T, $a \in \mathcal{A}$ are the rows of A^{full}. A common choice for Φ is the logarithm barrier $\Phi(u) = -\alpha\log(-u)$ where $\alpha > 0$ is a parameter that sets the accuracy of the approximation (Boyd & Vandenberghe, 2004, §11.2.1).

In summary, we pose the route flow estimation problem as a mathematical program optimizing the fit to link sensor data (and possibly other sources of data) over feasible route flow distributions, constrained to those which are consistent with measured cellpath flows in the network. Our data-driven approach is compatible with other types of data and also similar in formulation to route-based traffic assignment models.

[5] Since the inequalities $Ux = f,\ Tx = d,\ x \geq 0$ might not define simplexes, we chose formulation (8) over: $\min\ \frac{1}{2}\|Ax - b\|_2^2$ s.t. $Ux = f,\ Tx = d,\ x \geq 0$ to have the same constraints as in (2) for our algorithmic approach. Besides, with dense cellular networks, satisfying $Tx = d$ is redundant with the constraints $Ux = f$ because OD demands are included in cellular network data, hence both formulations reduce to (2).

3. Dimensionality reduction and projection via isotonic regression

In this section, we present an efficient constraint elimination technique relying on the choice of a particular nullspace, which is suitable for both the proposed route flow estimation problem (2) and the traffic assignment problem (5). The projection on the inequality constraints is performed in linear time via isotonic regression.

3.1. Exploiting the structure of the equality constraints

We consider the reduced route flow estimation problem (4) and the reduced traffic assignment problem (5):

$$\begin{aligned} \text{route flow estimation problem:} \quad & \min_x \tfrac{1}{2}\|\tilde{A}\tilde{x} - b\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^T\tilde{x}^p = 1,\ \tilde{x}^p \geq 0, \quad \forall\, p \in \mathcal{P} \\ \text{traffic assignment problem:} \quad & \min_x \phi(\tilde{A}^{\text{full}}\tilde{x}) \quad \text{s.t.} \quad \mathbf{1}^T\tilde{x}^k = 1,\ \tilde{x}^k \geq 0, \quad \forall\, k \in O^2 \end{aligned} \tag{10}$$

We consider a general objective function f and the simplices $S^p = \{\tilde{x}^p \in [0,1]^{n^p} \mid \sum_{r\in\mathcal{R}^p} \tilde{x}^p_r = 1\}$ as constraints, but the following analysis applies for both problems. We use standard linear algebra operations to eliminate the equality constraints (Boyd & Vandenberghe, 2004, §4.2.4). Since the constraints have disjoint support, we treat each one of them separately. For all $p \in \mathcal{P}$, we find a direction e^p which is a particular solution of $\mathbf{1}^T\tilde{x}^p = 1$, and a matrix N^p whose range is the *orthogonal complement* of the vector $\mathbf{1} \in \mathbb{R}^{n_p}$, denoted $\{t\,\mathbf{1} \mid t \in \mathbb{R}\}^\perp$. With the vectors $\{e^p\}_{p\in\mathcal{P}}$ stacked into an overall vector $\tilde{x}_0 := (e^p)_{p\in\mathcal{P}}$, and the matrices $\{N^p\}_{p\in\mathcal{P}}$ encoded in an overall block-diagonal matrix $N := \text{diag}((N^p)_{p\in\mathcal{P}})$, the resulting problem is:

$$\begin{aligned} &\min_z \ \tfrac{1}{2} f(\tilde{x}_0 + Nz) \\ &\text{s.t.} \ \ \tilde{x}_0 + Nz \geq 0 \end{aligned} \quad ; \qquad \text{or with the blocks made explicit:} \quad \begin{aligned} &\min_z \ f((e^p + N^p z^p)_{p\in\mathcal{P}}) \\ &\text{s.t.} \ \ e^p + N^p z^p \geq 0, \quad \forall\, p \in \mathcal{P} \end{aligned} \tag{11}$$

Vectors of the form $[\cdots, 1, -1, \cdots]^T$ are orthogonal to $\mathbf{1} \in \mathbb{R}^{n_p}$. We also choose a simple e^p solution of $\mathbf{1}^T x^p = 1$:

$$e^p := [0, \cdots, 0, 1]^T \in \mathbb{R}^{n_p}; \quad N^p = \begin{bmatrix} 1 & & \\ -1 & 1 & \\ & -1 & \ddots \\ & & \ddots \end{bmatrix} \in \mathbb{R}^{n_p \times (n_p - 1)} \quad \forall\, p \in \mathcal{P} \tag{12}$$

where the columns of N^p form a basis of $\{t\,\mathbf{1} \mid t \in \mathbb{R}\}^\perp$. These choices result in a simplification of the constraints in (11), and we can interchangeably operate on variables x^p in (2) and variables z^p in (4) since they are simply related:

$$\begin{aligned} \tilde{x}^p &= e^p + N^p z^p = [z_1^p,\ z_2^p - z_1^p,\ \cdots,\ z_{n_p}^p - z_{n_p-1}^p,\ 1 - z_{n_p}^p]^T,\ \forall\, p \in \mathcal{P} \\ z^p &= [\tilde{x}_1^p,\ \tilde{x}_1^p + \tilde{x}_2^p, \cdots,\ \textstyle\sum_{i=1}^{n-2} \tilde{x}_i^p,\ \sum_{i=1}^{n-1} \tilde{x}_i^p,]^T,\ \forall\, p \in \mathcal{P} \end{aligned} \tag{13}$$

The constraint $e^p + N^p z^p \geq 0$ becomes an ordering constraint $0 \leq z_1^p \leq \cdots \leq z_{n_p-1}^p \leq 1$. The program (11) is now:

$$\min_z \ f((e^p + N^p z^p)_{p\in\mathcal{P}}) \quad \text{s.t.} \quad 0 \leq z_1^p \leq \cdots \leq z_{n_p-1}^p \leq 1, \quad \forall\, p \in \mathcal{P} \tag{14}$$

The main advantage of this constraint elimination is the reduction of the dimension from n to $n - q$, where n is the number of routes and q the number of cellpaths (see Table 1). If each cellpath has maximum k routes, then we have $n \leq kq$, hence $n - q \leq n(1 - 1/k)$. For our target problem, we generally have $k \in [3, 50]$ hence the dimension can be reduced by as much as $1/3$.

The problem (14) can be solved quite efficiently with a simple (accelerated) first order or second order projection algorithm, or an Augmented Lagrangian method. In particular, the basic descent projection algorithm (see Algorithm 1) iteratively takes a step in a descent direction Δz (line 2) from the current point z, projects the new point $z + \Delta z$ onto the constraint set $z^+ := \Pi(z + \Delta z)$ (line 3), and performs a line search (line 4). The projection step is performed with q Euclidean projections of $z^p + \Delta z^p$ onto ordering constraints:

$$\Pi^p(y^p): \quad \min_{u^p} \|u^p - y^p\|_2^2 \quad \text{s.t.} \quad 0 \leq u_1^p \leq u_2^p \leq \cdots \leq u_{n_p-1}^p \leq 1 \quad \forall\, p \in \mathcal{P} \tag{15}$$

In line 4 of Algorithm 1, we perform a backtracking line search (Boyd & Vandenberghe, 2004, §9.2). This is an Armijo-rule based step size selection that ensures sufficient descent, it approximately minimizes the objective along the projected arc $\{z + t(z^+ - z) \mid t \in [0,1]\}$. Since the feasible set is convex, the projected arc is feasible, hence the method also ensures feasibility of the next iterate. We apply backtracking with objective $f(z) = \|A(\tilde{x}_0 + Nz)\|_2^2$ and descent direction $d = z^+ - z$.

Algorithm 1 `Proj-descent(·)` General projected descent method

Require: initial point $z = (z^p)_{p\in\mathcal{P}}$ in the feasible set $\mathcal{X}$.

while stopping criteria not met **do**
  Determine a descent direction $\Delta z = (\Delta z^p)_{p\in\mathcal{P}}$
  Projection: $(z^p)^+ := \underset{u^p}{\operatorname{argmin}} \{\|z^p + \Delta z^p - u^p\|^2 \ : \ 0 \le u_1^p \le u_2^p \le, \cdots \le u_{n_p-1}^p \le 1\}, \ \forall\, p \in \mathcal{P}$
  Line search on the projected arc: $\gamma :\approx \operatorname{argmin}\{f(z + t(z^+ - z)) \ : \ t \in [0,1]\}$
  $z := z + \gamma(z^+ - z)$
end while
return z

3.2. A simple projection using isotonic regression

The projections (15) have general form (16), given data points $y := [y_1, \cdots, y_n] \in \mathbb{R}^n$, weights $w := [w_1, \cdots, w_n] \succ 0$, and bounds $L < U$.[6] Without bounds, we have an isotonic regression problem (17) (see (Tibshirani *et al.* , 2011) and references therein).

$$\text{ISO}_{1\to n}^{[L,U]}(y,w): \quad \min_u \ \textstyle\sum_{i=1}^n w_i(y_i - u_i)^2 \quad \text{s.t.} \quad L \le u_1 \le u_2 \le \cdots \le u_n \le U \tag{16}$$

$$\text{ISO}_{1\to n}^{\mathbb{R}}(y,w): \quad \min_u \ \textstyle\sum_{i=1}^n w_i(y_i - u_i)^2 \quad \text{s.t.} \quad u_1 \le u_2 \le \cdots \le u_n \tag{17}$$

where we use the notation $\text{ISO}_{s\to t}^{I}(y,w)$ such that subscript $s \to t$ means we only consider data points with indices from s to t, and superscript I is the interval in which the variables $u_s, u_{s+1}, \cdots, u_t$ lie. Since both problems are strongly convex, they both have a unique solution. The solution to (17), denoted u^{iso}, can be computed in linear time using the *Pool Adjacent Violators* (PAV) algorithm (Best & Chakravarti, 1990, §3), so one hopes that the solution to (16), denoted $u^\star$, derives easily from u^{iso}. In fact, we prove the following result:

Proposition 2. The solution $u^\star$ to (16) is the Euclidean projection of the solution u^{iso} to (17) onto $[L,U]^n$.

Although isotonic regression is generally studied in the form (17), the bounded version (16) has appeared in (Grotzinger & Witzgall, 1984). The simple connection presented in Proposition 2 is new to the best of our knowledge. This result can be written $u^\star = \Pi_{[L,U]^n}(u^{\text{iso}})$ where $\Pi_{\mathcal{K}}$ is the Euclidean projector onto space $\mathcal{K}$. When $\mathcal{K} = [L,U]^n$, the projected vector $p := \Pi_{[L,U]^n}(u)$ is obtained from $u \in \mathbb{R}^n$ by simply projecting each entry u_i onto $[L,U]$, *i.e.* $p_i = u_i$ if $u_i \in [L,U]$, $p_i = L$ if $x_i < L$, and $p_i = U$ if $x_i > U$. We first give a lemma.

Lemma 1. Given u^{iso} the solution to (17), if there exists k such that $u_k^{\text{iso}} < u_{k+1}^{\text{iso}}$ then (17) reduces to two subproblems:

$$\begin{aligned} \text{ISO}_{1\to k}^{\mathbb{R}}(y,w): &\quad \min_u \ \textstyle\sum_{i=1}^k w_i(y_i - u_i)^2 \quad \text{s.t.} \quad u_1 \le \cdots \le u_k \\ \text{ISO}_{k+1\to n}^{\mathbb{R}}(y,w): &\quad \min_u \ \textstyle\sum_{i=k+1}^n w_i(y_i - u_i)^2 \quad \text{s.t.} \quad u_{k+1} \le \cdots \le u_n \end{aligned} \tag{18}$$

such that $[u_1^{\text{iso}}, \cdots, u_k^{\text{iso}}]$ is the solution to the first one and $[u_{k+1}^{\text{iso}}, \cdots, u_n^{\text{iso}}]$ is the solution to the second one. The same result holds for (16) and $u^\star$, with resulting subproblems $\text{ISO}_{1\to k}^{[L,+\infty)}(y,w)$ and $\text{ISO}_{k+1\to n}^{(-\infty,U]}(y,w)$.

Proof: Since the constraint $u_k \le u_{k+1}$ is not active at u^{iso}, it may be removed from (17) without altering the solution. Then the resulting program can be separated into the two programs in (18) with respective solutions $[u_1^{\text{iso}}, \cdots, u_k^{\text{iso}}]$ and $[u_{k+1}^{\text{iso}}, \cdots, u_n^{\text{iso}}]$. □

Proof of Proposition 2: We start with two simple cases.

Case 1: $[u_i^{\text{iso}} \le L, \ \forall\, i]$. Suppose $\exists\, k$, $u_k^\star > L$. We choose k the smallest of such indices, then either $k = 1$ or $L = u_{k-1} < u_k$. In both cases, $[u_k^\star, \cdots, u_n^\star]$ is the unique solution to $\text{ISO}_{k\to n}^{(-\infty,U]}(y,w)$ from Lemma 1. Since $[u_k^{\text{iso}}, \cdots, u_n^{\text{iso}}]$ is also feasible for $\text{ISO}_{k\to n}^{(-\infty,U]}(y,w)$, we have $\sum_{i=k}^n w_i(y_i - u_i^{\text{iso}})^2 > \sum_{i=k}^n w_i(y_i - u_i^\star)^2$, and adding $\sum_{i=1}^{k-1} w_i(y_i - u_i^{\text{iso}})^2$ on both sides yields $\sum_{i=1}^n w_i(y_i - u_i^{\text{iso}})^2 > \sum_{i=1}^{k-1} w_i(y_i - u_i^{\text{iso}})^2 + \sum_{i=k}^n w_i(y_i - u_i^\star)^2$. Since $[u_1^{\text{iso}}, \cdots, u_{k-1}^{\text{iso}}, u_k^\star, \cdots, u_n^\star]$ is also feasible for (17) ($u_{k-1}^{\text{iso}} \le l < u_k^\star$), this contradicts the optimality of u^{iso}. Hence $u_k^\star = L$, $\forall\, k$, *i.e.* $u^\star = \Pi_{[L,U]^n}(u^{\text{iso}})$.

[6] For subsection 3.2 only, $U \in \mathbb{R}$ is the upper bound in problem (16). In the rest of the article, U is the cellpath-route incidence matrix.

Case 2: $[u_i^{\text{iso}} \geq U, \forall\, i]$. The analysis is similar to case 2. We have: $u_k^\star = U, \forall\, k$, *i.e.* $x^\star = \Pi_{[L,U]^n}(u^{\text{iso}})$.

General case: Without loss of generality, we suppose there exist two indices s, t such that: $u_1^{\text{iso}} \leq \cdots \leq u_s^{\text{iso}} \leq L < u_{s+1}^{\text{iso}} \leq \cdots \leq u_{t-1}^{\text{iso}} < U \leq x_t^{\text{iso}} \leq \cdots \leq x_n^{\text{iso}}$. From Lemma 1, $[u_1^{\text{iso}}, \cdots, u_s^{\text{iso}}]$, $[u_{s+1}^{\text{iso}}, \cdots, u_{t-1}^{\text{iso}}]$, and $[u_t^{\text{iso}}, \cdots, u_n^{\text{iso}}]$ are then solutions to $\text{ISO}^{\mathbb{R}}_{1\to s}(y, w)$, $\text{ISO}^{\mathbb{R}}_{s+1\to t-1}(y, w)$, and $\text{ISO}^{\mathbb{R}}_{t\to n}(y, w)$ respectively. From case 1, the vector $[L, \cdots, L] \in \mathbb{R}^s$ is solution to $\text{ISO}^{[L,+\infty)}_{1\to s}(y, w)$ and from case 2, the vector $[U, \cdots, U] \in \mathbb{R}^{n-t+1}$ is solution to $\text{ISO}^{(-\infty,U]}_{t\to n}(y, w)$. Then the global vector $x^* := [L, \cdots, L, u_{s+1}^{\text{iso}}, \cdots, u_{t-1}^{\text{iso}}, U, \cdots, U]$ is the solution to the global program:

$$\begin{array}{l} \min_u \sum_{i=1}^n w_i(y_i - u_i)^2 \\ \text{s.t.} \quad L \leq u_1 \leq \cdots \leq u_s, \quad u_{s+1} \leq \cdots \leq u_{t-1}, \quad u_t \leq \cdots \leq u_n \leq U \end{array} \tag{19}$$

Adding the constraints $u_s \leq u_{s+1}$ and $u_{t-1} \leq u_t$ to (19) does not alter the solution since they are inactive. Hence $[L, \cdots, L, u_{s+1}^{\text{iso}}, \cdots, u_{t-1}^{\text{iso}}, U, \cdots, U]$ is the solution to (16), *i.e.* $u^\star = \Pi_{[L,U]^n}(u^{\text{iso}})$. □

Algorithm 2 `PAV-proj(`y^p`)` Projection onto ordering constraints in line 3 of Algorithm 1

Require: vector $y^p \in \mathbb{R}^{n_p - 1}$

1: compute $y^{p,\text{iso}} := \underset{u^p}{\text{argmin}} \{\|u^p - y^p\|_2^2 : u_1^p \leq u_2^p \leq \cdots \leq u_{n_p-1}^p\}$ with the PAV algorithm (Best & Chakravarti, 1990)

2: project $y^{p,\text{iso}}$ onto $[0,1]^{n_p-1}$: $\tilde{y}_k^p = y_k^{p,\text{iso}}$ if $y_k^{p,\text{iso}} \in [0,1]$; $\tilde{y}_k^p = 0$ if $y_k^{p,\text{iso}} \leq 0$; $\tilde{y}_k^p = 1$ if $y_k^{p,\text{iso}} \geq 1$.

3: **return** return $\tilde{y}^p$

In Algorithm 2, we give an efficient algorithm to perform the projections (15) in line 3 of Algorithm 1. We note that without the constraint elimination described earlier, a projected descent method applied to (4) would require q projections onto the probability simplices $\{\tilde{x}^p \in \mathbb{R}^{n_p} \,|\, \mathbf{1}^T \tilde{x}^p = 1,\ \tilde{x}^p \geq 1\}$ at each iteration. The complexity of these projections is $\text{O}(n_p \log n_p)$ (Duchi *et al.* , 2008; Wang & Carreira-Perpin, 2013), which is less attractive than the $\text{O}(n_p)$ complexity of Algorithm 2.

Problems (4) and (6) are both convex, and can be solved efficiently with including interior point methods, augmented Lagrangian, gradient projection, and conjugate gradient. In particular, we choose the *Barzilai and Borwein* (BB) method for the accelerated gradient method, where z is the current iterate and z^- and previous iterate:

$$\Delta z = -((y^T s)/(y^T y))\Delta f(z) \quad \text{where} \quad y = \nabla f(z) - \nabla f(z^-), \quad s = z - z^- \tag{20}$$

The change of variable reduces the dimensionality, at the cost of losing some of the intuitive structure of the route choice problem. While long-standing algorithms such as the Frank-Wolfe assignment (LeBlanc *et al.* , 1975) and the Origin-based assignment (Bar-Gera, 2002) and their modifications may have diminished efficiency since the all-or-nothing assignment step is no longer available, their slow convergence is known (Ortuzar & Willumsen, 2001, §11.2.3.1). We propose that the estimation problem (4) and the traffic assignment problem (6) can be reduced to the form (11), and then be solved efficiently with quasi-Newton methods (*e.g.* L-BFGS (Nocedal & Wright, 2006)), accelerated gradient methods, or alternating direction methods. These algorithms are proven to have fast convergence, and the proposed projection step is efficient as discussed above. Due to space limitations, early numerical results on the speed up of the algorithms are not shown in the present article.

4. Experimental setting and validation process

We demonstrate our approach by providing numerical results on networks of varying sizes, applying different traffic assignment models and sensor configurations, all based on the I-210 highway corridor in Los Angeles. To demonstrate the versatility to the underlying data-driven approach, we investigate the following three scenarios (see Figure 4):

1. *Highway network in user equilibrium* (UE), with varying cell densities and static sensor coverage.
2. *Highway network in system optimum* (SO), with varying cell densities and static sensor coverage.
3. *Activity-based agent model on full network*, with varying cell densities, 5% static sensor coverage.

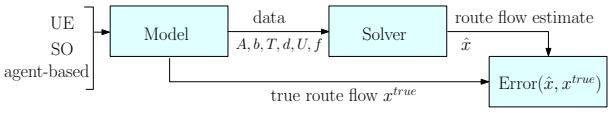

Fig. 4. Our experiment flow block diagram, where the model is comprised of a network, traffic assignment model, and sensor configuration. The solver is presented in §3. The error metric represented here is a function of the estimated and actual route flow. We may compute the percent flow error or, using additional information (e.g. network topology and actual link flow), we may also compute the link flow GEH error.

We have intentionally selected three different traffic assignment models (UE, SO, agent-based) to test the versatility of our method; we aim to demonstrate that our method is not only accurate (provided enough measurements) and appropriate for full-scale networks but also model agnostic, thereby highlighting a major advantage of our approach to those that require more rigid assumptions on agent behavior. Thus, we study networks of different sizes and complexities, different driver behavior models, and trade-offs for different sensor placements. We additionally present preliminary investigation on the effect of measurement and model error on the accuracy of the approach.

4.1. Sensor configurations

We have two main types of data: link sensors data (loop based) and cellpath sensors (cell based). We consider link sensors on a subset of the links in the network (ranging from 5% to 100% coverage). For the highway network with UE/SO flow, the subsets of links are chosen such that the *most congested* links are observed, *i.e.* links with highest traffic volumes or flows, whereas in the full large-scale network, we use locations of real highway (PeMS (Choe *et al.* , 2002)) and arterial loop sensors where the coverage is 5%.

Although the use of real cellular network data from a service provider would demonstrate even stronger applicability of our framework, its availability is restricted for privacy issues. Our team at the present time is not able to share findings based on collaborations with companies such as AT&T. Nevertheless, the use of well-designed simulators remains necessary for demonstrating how of our framework may apply to different networks and settings, and also for the ease of validation, as route flows are not yet measurable in real-world settings. Our model for cell placement is based on employee population density and locations of major roads. Most notably, many ordinances prohibit towers in residential areas but promote towers in industrial and commercial centers. For both networks, the locations $(X_i, Y_i) \in \mathbb{R}^2$ of the cell towers are randomly sampled on the plane such that the distribution models realistically represent the coverage based on region demographics. The overall sensor configuration (21,22,23) consists of $N = N^B + N^S + N^L$ total cell towers, where N^B, N^S, N^L are predetermined for each experiment and the weights of the multinomial distributions are determined by demographics and geometry. Our sensor configurations are drawn from three distribution models:

1. Within the whole region delimited by a *bounding box*: N^B cell tower locations $\{(X_i^B, Y_i^B)\}_{i=1,\cdots,N^B}$ are sampled uniformly (21).
2. Within sub-regions $\mathcal{S}$ comprising the full region: For each sub-region $s \in \mathcal{S}$, delimited by a rectangle $(X^s_{\min}, Y^s_{\min})$, $(X^s_{\max}, Y^s_{\max})$, N^s additional cell tower locations $\{(X_i^s, Y_i^s)\}_{i=1,\cdots,N^s}$ are sampled (22). The number of base stations N^s for each sub-region s is sampled from a multinomial distribution with N^S trials and weights proportional to demographic information for each region (e.g. employee population). That is, $N^S = \sum_{s\in\mathcal{S}} N^s$ is the total number of cell towers among all the sub-regions (excluding those sampled from the entire bounding box).
3. Near major links in the network: Along each link $a \in \tilde{\mathcal{A}} \subset \mathcal{A}$ (also called *arcs*), where $\tilde{\mathcal{A}}$ denotes a pre-selected subset of major links in the network, N^a cell tower locations are sampled uniformly along the link with Gaussian noise (23) where (X_s^a, Y_s^a) is the location of the start of link a, and (X_t^a, Y_t^a) is the location of the end of link a. The numbers of base stations N^a along links $a \in \mathcal{A}$ are sampled from a multinomial distribution with N^L trials and weights proportional to the length of a, where N^L is the total number of cells along links.

$$\text{Bounding box}: \quad X_i^B \sim U([X^B_{\min}, X^B_{\max}]),\ Y_i^B \sim U([Y^B_{\min}, Y^B_{\max}]), \quad \text{for} \quad i = 1, \cdots, N^B \tag{21}$$

$$\text{Sub-region } S: \quad X_i^s \sim U([X^s_{\min}, X^s_{\max}]),\ Y_i^s \sim U([Y^s_{\min}, Y^s_{\max}]), \quad \text{for} \quad i = 1, \cdots, N^s \tag{22}$$

$$\text{Link } a: \quad \begin{cases} X_i^a \sim X_s^a + t_i(X_t^a - X_s^a) + N(0,\sigma) \\ Y_i^a \sim Y_s^a + t_i(Y_t^a - Y_s^a) + N(0,\sigma) \end{cases} \quad \text{such that} \quad t_i \sim U([0,1]), \quad \text{for} \quad i = 1, \cdots, N^a \tag{23}$$

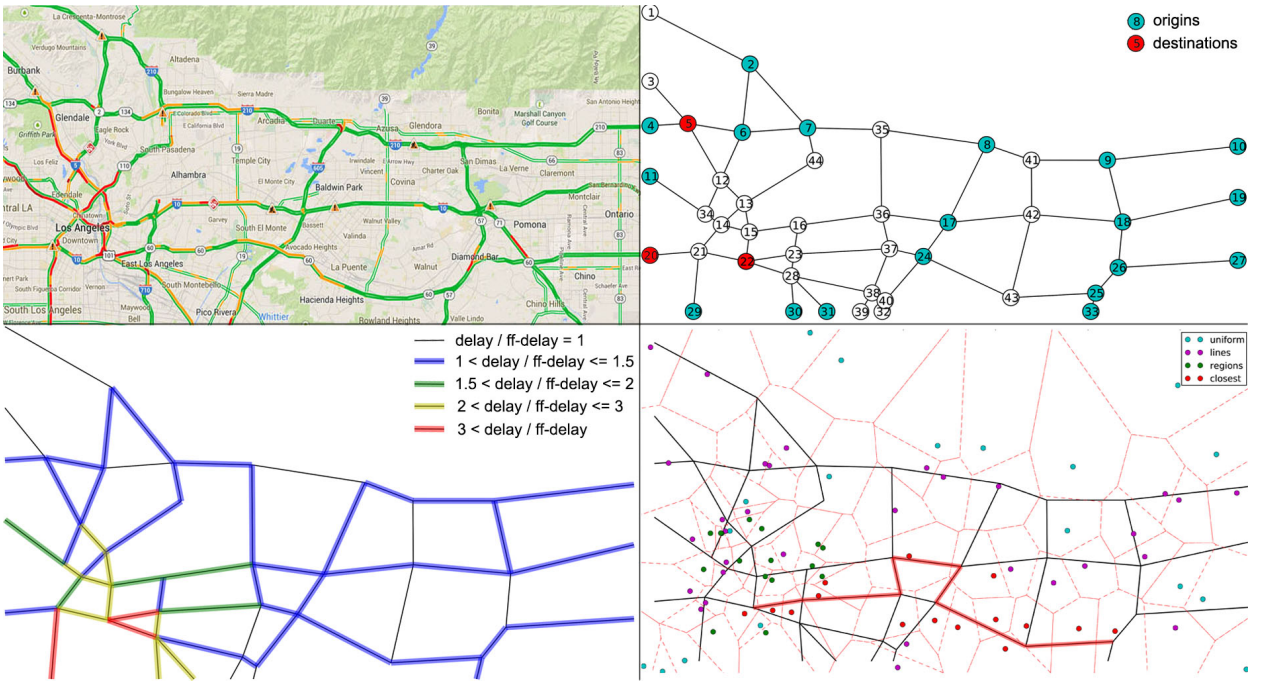

Fig. 5. Benchmark (small-scale) example used for the first numerical run: The four subfigures present the highway network of the I-210 highway corridor in L.A. county. Starting from the top left and in clockwise order: 1) The state of traffic on 2014-06-12 at 9:14 AM from Google Maps; 2) The nodes in blue and red are nodes from which positive flows emanate, nodes in red are nodes from which positive flows terminate; 3) Network with 80 sampled cells, with a higher concentration of cells near downtown. A random path from 25 to 22 is shown in red with the closest cell towers. 4) The highway network in user equilibrium with the resulting delays. Best viewed in color.

4.2. Scenarios 1 and 2: UE and SO on the highway network

We consider first the *highway network* of the I-210 region in Los Angeles[7]. The roads are extracted from OpenStreetMaps (OSM) and we retain only links with at least five (up to 11) lanes. This results in a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ with $|\mathcal{N}| = 44$ nodes and $|\mathcal{A}| = 122$ directed links. We calibrate the free flow delay τ_a for each link $a \in \mathcal{A}$ using the link's length and free flow speed (provided by OSM) and empirical delays values (provided by Google Maps). An illustration of the network is provided in Fig. 5.

The OD demands are based on census data and employment concentration in L.A. county, which are extracted from the Census Bureau. The OD demand model is simplified to a static morning rush hour model[8] of the region such that: i) only 21 origins have positive flows emanating from them; ii) all the trips terminate at three destinations: near Burbank at node 5, towards Santa Monica at node 20, and in Downtown L.A. at node 22; iii) we only have 42 OD pairs with positive flows ranging from 1200 veh/hour to 12,000 veh/hour.[9]

In our equilibrium-based numerical study, we consider the traffic assignment model presented in (Sheffi, 1985, §3.1) to generate route flows and cellpath flows. The delay on a given link a is assumed to be a strictly increasing function $c_a(\cdot)$ of the traffic volume (flow) v_a on that link. We choose the widely used delay function estimated by the Bureau of Public Roads, where τ_a is the free flow delay (sec) and m_a is the number of lanes on link a, and provide the Beckmann objective function ϕ^{UE} associated to the overall model (Beckmann *et al.* , 1956)):

$$\text{link delay:} \quad c_a(v_a) = \tau_a(1 + 0.15(v_a/m_a)^4), \ \forall\, a \in \mathcal{A}\ ; \qquad \text{UE objective:} \quad \phi^{\mathrm{UE}}(v) = \sum_{a\in\mathcal{A}} \int_0^{v_a} c_a(u)du \tag{24}$$

[7] The region has bounding box [-118.328299, 33.984601, -117.68132, 34.255881] in latitude and longitude coordinates.
[8] Based on observed flows on 2014-06-12 at 9:14 AM from Google Maps.
[9] Highway experiment implementation (Python) is available at `https://github.com/megacell/traffic-estimation-wardrop`.

In this section only, we use the following notation: the nodes of the network are indexed by $i \in \mathcal{N}$, the 42 OD pairs with positive OD flow are indexed by by $k \in \{1, \cdots, Q\}$, $A^{\text{full}} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{R}|}$ is the link-route incidence matrix, and $N \in \{-1, 0, 1\}^{|\mathcal{N}|\times|\mathcal{A}|}$ is the node-link incidence matrix. For each OD pair $k = (s_k, t_k)$ we define an associated vector $e^k \in \mathbb{R}^{|\mathcal{N}|}$ such that $e^k_i = -d_k$ at node $i = s_k$ (the origin), $e^k_i = d_k$ at node $i = t_k$ (the destination), and $e^k_i = 0$ otherwise. Under the assumptions of our experiment, the path-flow traffic assignment (PTA) is equivalent to the link-flow traffic assignment (LTA), *i.e.* they give the same *unique* link flow solution (Ford & Fulkerson, 1962):

$$\text{PTA}: \ \min \phi^{\text{UE}}(A^{\text{full}}x) \ \text{ s.t. } Tx = d, \ x \geq 0 \tag{25}$$

$$\text{LTA}: \quad \min \phi^{\text{UE}}(v) \quad \text{s.t. } v \in \mathcal{K} := \left\{ v \in \mathbb{R}_+^{|\mathcal{A}|} \mid \exists\, w^k \in \mathbb{R}_+^{|\mathcal{A}|},\ v = \sum_{k=1}^{Q} w^k,\ Nw^k = e^k,\ \forall\, k \in \{1, \cdots, Q\} \right\} \tag{26}$$

Since PTA is not tractable due to the computational cost of enumerating all the possible routes, we solve LTA in the first step, then perform the following steps to generate a set of routes $\mathcal{R}$ with an associated UE route flow vector $x^{\text{UE}} \in \mathbb{R}_+^{|\mathcal{R}|}$, and a set $\mathcal{P}$ of cellpaths with a feasible UE cellpath flow vector $f^{\text{UE}} \in \mathbb{R}_+^{|\mathcal{P}|}$:

1. We solve LTA and obtain the UE link flow $v^{\text{UE}} \in \mathbb{R}_+^{|\mathcal{A}|}$ and resulting link delays.
2. We find the K-shortest paths with the UE delays for each of the 42 OD pairs, using Yen's algorithm (Yen, 1971). Note that K is chosen large enough such that *at least* all used routes are extracted, *i.e.* all the routes with the (same) shortest delays as characterized by the Wardrop equilibrium. We choose $K = 5$ and extract 207 candidate routes.
3. We solve PTA with the 207 *candidate routes* starting from a random initial point. Let x^{UE} be a route flow solution (the resulting link flow $A^{\text{full}}x^{\text{UE}}$ should equal to v^{UE} since the UE link flow is unique).
4. We sample cells on the highway network following the model presented in §4.1 (see Fig. 5).
5. Among the 207 routes, we found $\left|\{r \mid x^{\text{UE}}_r > 0\}\right| = 90$ *used routes*. We compute the sequence of cells that intersects with each used route to determine the cellpath flows, given by: $f^{\text{UE}}_p = \sum_{r\in\mathcal{R}^p} x^p_r$.

On a network with SO flow, the total delay is minimized (Wardrop & Whitehead, 1952; Kelly, 1991), hence the objective function to be minimized is ϕ^{SO} in (27) subject to the constraints in (25) and (26), for the path-flow and link-flow formulations, respectively. In fact, the SO link flow corresponds to the UE link flow with the modified delay function $\tilde{c}_a(\cdot)$ in (27), called the marginal delay function (Roughgarden, 2003) (where the prime indicates the derivative function):

$$\text{link marginal delay}: \quad \tilde{c}_a = c_a(v_a) + v_a c'_a(v_a)\ ; \qquad \text{SO objective}: \quad \phi^{\text{SO}}(v) = \sum_{a\in\mathcal{A}} v_a c_a(v_a) \tag{27}$$

Steps 1 to 5 are performed for the SO objective ϕ^{SO} via the link marginal delay formulation to generate a SO route flow x^{SO} and a SO cellpath flow f^{SO} on the highway network described above, with a few minor differences:

- In step 2, we find the K-shortest paths under the marginal delays induced by the SO link flow. We choose $K = 10$ and we extract 411 candidate routes.
- In step 5, we found 164 routes with positive flow on it.

4.3. Scenario 3: activity-based agent model on the large-scale full network

We additionally consider a large *full network*, comprising of both the highway network and the arterial networks in the region. We use the OpenStreetMaps network of the greater Los Angeles area, excluding residential links. Our network comprises of 20,513 edges (links) and 10,538 nodes (intersections). We take the origins to be the Traffic Analysis Zones (TAZ) given by the US census, of which there are 778 in the region (see Fig. 6). We use a commercially available OD model for the region, called the Census Transportation Planning Products (CTPP) model.

On this large-scale network, we utilize an *activity-based agent model* for simulating the traffic assignment. MATSim is a well-known open-source traffic simulation framework (Illenberger & Nagel, 2007), which takes in a set of K agent home and work locations and outputs a set of K trajectories (time-stamped sequences of links) that each agent performed. MATSim searches for a user equilibrium in terms of utility functions defined for the agents using a

Fig. 6. Full-scale network including highway and arterial networks of the I-210 corridor used for MATSim data generation, and for the estimation problem. See Figure 2 for the Voronoi tessellation model of the cellular network and the 700 origins given by the TAZ.

co-evolutionary optimization algorithm. In our setting, we consider agent utility as a function of travel time. MATSim differs from the user equilibrium model above in that it is quasi-static, by allowing slight variation in the departure times for each agent. MATSim is suitable for performing large-scale agent simulations; we simulate the morning and evening rush-hours using 500,000 agents, as those are the most vital times to understand the state of traffic. The home and work locations for each agent are distributed randomly according to census demographics. Since these locations are selected randomly within origins and destination (as opposed to selected randomly among the region centroids), typically all of the trajectories generated are unique. Viewing the full set of trajectories as our set of possible routes lends itself to be a trivial problem in our formulation. Instead of all routes, we consider the "important" routes in the network. Therefore, we examine trajectories between each OD pair and group them by similarity as follows: 1) Find the trajectory which matches with the most other trajectories ($\geq$80% match in length). Add this trajectory to the list of routes for the OD pair; 2) Remove all trajectories that match with this route and repeat. We stop when 50 routes are selected or when there are no more trajectories. 50 routes empirically accounts for 99.4% of the 500K trajectories. This procedure yields a set of 304,695 routes.

4.4. Implementation

Our full system is available at `https://megacell.github.io/` for validation and extension, including the optimization routine, the small-scale network experiments, and the large-scale pipeline. We hope that our framework will be a benchmark for many future studies of estimation problems in transportation science. The software to run the full-scale experiments was developed mostly in `python 2.7`, using the `GEOS (v.3.4.2)` library for geometric computations. All data is managed and stored in a `PostGIS 2.1.3` database. The geometries and other data about routes, cell tower Voronoi tesselations, and the links of the road network are all stored in the database with spatial indices on all geometry columns, allowing `PostGIS` spatial queries to be performed efficiently for extracting cellpath information associated with each route. The convex optimization program[10] was developed in Python, using `scipy.sparse` and `numpy` for matrix computation. The PAV projection algorithm was written in `C`, and bindings were written so that it could be called from the Python optimization algorithm.

The full network dataset for the I-210 corridor contains 280x691 routes, 778 origins, 1033 sensors, and was tested with a variety of different cells, ranging in number from 250 to 8000. The incidence matrices U (roughly 250K-by-300K matrices) are generated by finding the cellpath for each route from the database by ordering the sequence of Voronoi cells that intersect with the respective route. The link-route incidence matrices A and are formed by finding all routes whose distance from the sensor locations was less than some threshold empirically selected such that the

[10] Implementation (Python, C) is open source and available at `https://github.com/megacell/block-simplex-least-squares`.

maps matched well ($\approx$ 10 meters tolerance for the PeMS loop sensor locations). All incidence matrices are saved in the `scipy.sparse` format.

5. Numerical results

We validate our approach by measuring our accuracy in terms of the route flow estimates, denoted $\hat{x}$, given different scenarios. Note that we solve the reduced problem presented in (4) according to our algorithmic approach, and the solution in z-space is converted to $\tilde{x}$-space following the simple relation in (13), and is subsequently rescaled to $\hat{x} = \text{diag}(f^T U)\tilde{x}$. We additionally present our accuracy in terms of link flow estimates, to serve as a comparison to classical approaches to link flow estimation:

- Route flow error: $\epsilon_r = \left\|x^{true} - \hat{x}\right\|_1 / \left\|x^{true}\right\|_1$, with x^{true} the true route flow and $\hat{x}$ the estimated route flow. This relative error may be thought of as the percent error of flow allocation among all routes.
- Link flow error for observed links and all links, respectively:

$$\epsilon_l^{obs} = \frac{\left|\epsilon_{GEH}(b_i^{true}, \hat{b}_i) < 5, \forall i \in \left\{1, \cdots \left|\hat{b}\right|\right\}\right|}{|b^{true}|}, \text{ with } b^{true} = Ax^{true}, \ \hat{b} = A\hat{x} \tag{28}$$

$$\epsilon_l^{full} = \frac{\left|\epsilon_{GEH}(v_i^{true}, \hat{v}_i) < 5, \forall i \in \{1, \cdots |\hat{v}|\}\right|}{|b^{true}|}, \text{ with } v^{true} = A^{full}x^{true}, \ \hat{v} = A^{full}\hat{x} \tag{29}$$

where $|\cdot|$ denotes cardinality and $\epsilon_{GEH}(y, \hat{y}) = \sqrt{\frac{(y-\hat{y})^2}{0.5(y+\hat{y})}}$.

$\epsilon_{GEH}(\cdot, \cdot)$ is called the GEH statistic, a heuristic formula commonly used to compare two sets of traffic volumes, e.g. for calibration of microsimulation models (Dowling *et al.* , 2004, §5.6) and for validating hourly traffic flows (of Transportation (WisDOT), 2013, §11-13). For an individual link, a GEH value of less than 5.0 is considered to be a good match. For a vector of links, a fraction $\epsilon_l \geq 0.85$ of good matches is considered a good match overall between modeled and observed volumes.

Note that our method always achieves the optimal link flow error $\epsilon_l^{obs} = 1$ for all networks, traffic assignment models, and sensor configurations, since our formulation minimizes the error to the observed link flows. However, we include this metric because it is a metric upon which we can validate real network settings, without relying on traffic simulators.

5.1. Highway network

Using the highway network scenarios in §4.2, we vary the link coverage from 10% to 100% and the cell density from 10 to 120 cells such that the proportions are $N^B : N^L : N^S :: 1 : 2 : 1$. We always observe the most congested links, and regions $\mathcal{S}$ contains only 1 region and is roughly downtown Los Angeles (see 4.1). We analyze how the relative error ϵ_r in route flows vary when sensors are more sparse. Since we choose random initial points in PTA (25) and in the solver (2) to generate synthetic route flows and compute the estimate respectively, and since the cellular network is sampled randomly, all the results presented in this section have been averaged over 100 trials. Figure 7 presents the numerical results when link flows and optionally OD demands are known, and cellular network data are assimilated into the model. That is, we solve and compare both the problem without OD flows in (2) and with OD flows in (8).

In the left column of Fig. 7, we consider Problem (2), where we compare the performance of route flow estimation via cellpath flow vs OD flow alone. As expected, the accuracy increases as we observe more links and/or more cells. We observe that in the regime where we have low link sensor coverage, having even very few cells outperforms OD demands. It is interesting that observing the additional links in the 40-70% region makes a significant difference in the accuracy for the OD demands. In this region (and beyond), it is possible to achieve an accuracy of 98.8% and 98.4% (for UE and SO, respectively) with a sufficient selection of cells. Finally, we note that 80 cells and 120 cells respectively achieves 96.0% and 98.7% accuracy for UE and 93.5% and 98.0% accraucy for SO, in the absence of OD demands and with only 10% links observed. This indicates that with a sufficient selection of cells, route flow estimation may be possible even without other kinds of sensor data.

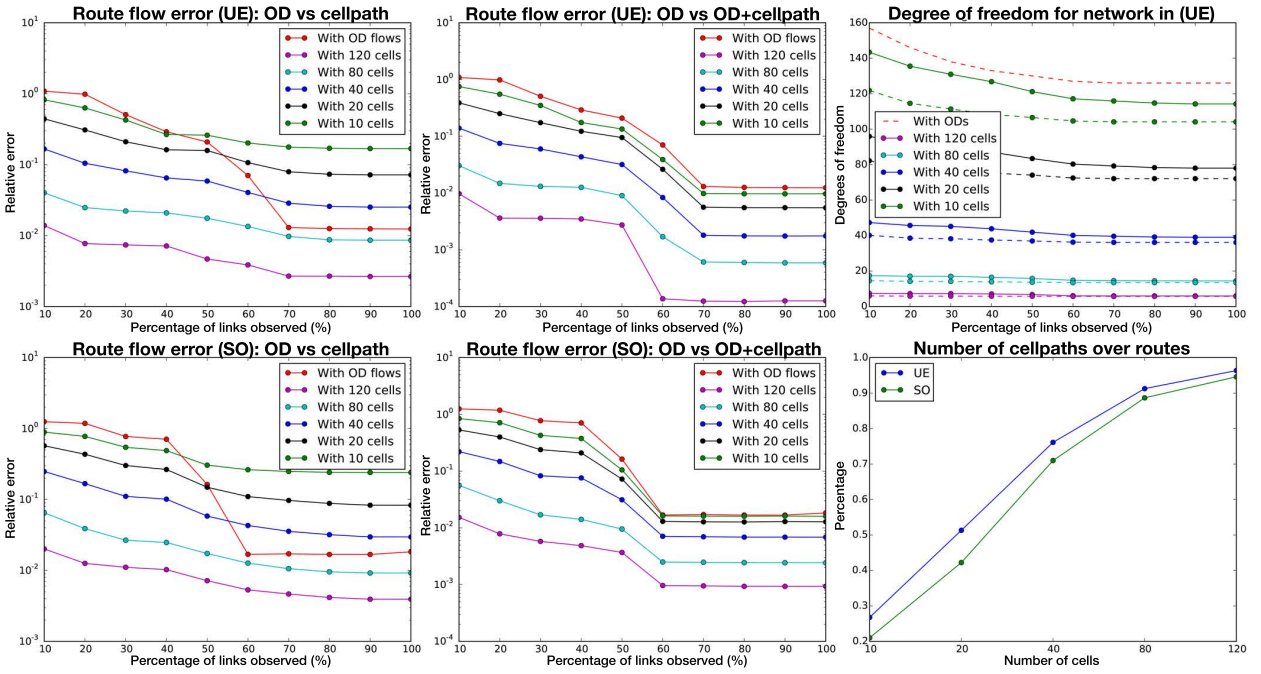

Fig. 7. The six subfigures present the numerical results for the highway network. Top row, from the left to right: 1) the route flow error ϵ_r from OD demands (red curve) and cellpath flows only (other curves) with different link coverage values and different numbers of cells for the network in UE; 2) the route flow error ϵ_r from OD demands (red curve) and OD demands & cellpath flows (other curves) with different link coverage values and different numbers of cells for the network in UE; 3) lower bound on the degree of freedom for the program with OD demands (red curve), cellpath flows only (other curves, solid), and OD demands & cellpath flows (other curves, dotted) for the network in UE; Bottom row, left to right: 4) ϵ_r from OD demands (red curve) and cellpath flows only (other curves) for different configurations of the network in SO; 5) ϵ_r from OD demands (red curve) and OD demands & cellpath flows (other curves) for different configurations of the network in SO; 6) ratio of the number of observed cellpaths to the number of candidate routes. Best viewed in color.

In the middle column of Fig. 7, we consider Problem (8), which considers cellpath flows and OD demands together for estimation. As expected, adding information from any number of cells performs strictly better than having no cells. With both types of information, the highway network can achieve beyond 99.0% accuracy (with at least 70% links observed) and 98.4% accuracy (with at least 60% links observed), for UE and SO, respectively.

For both experiments above, the accuracy in the UE settings is generally better than that of the SO settings. In the bottom right subfigure of Fig. 7, we see that the ratio of the number of cellpaths observed to the number of routes used is greater for UE than SO for all the cell counts in our experimental setup; this is due to the tendency of agents to consider more routes in SO, and thus the same number of cells provides less resolution into the route choice of the agents. This provides evidence that the SO setting is more difficult for estimation.

The accuracy of the estimates is closely related to the degree of freedom of the solution in Problems (2) and (8). We compute an upper bound on the degree of freedom as $n - \text{rank}[A^T, U^T]$ and $n - \text{rank}[A^T, T^T, U^T]$, respectively, where n is the dimension of the problem. It is an upper bound because we do not consider how the non-negativity constraint $x \geq 0$ limits the solution space. In the top right subfigure of Figure 7, we observe that in all cases, the use of cellpath information limits the degree of freedom moreso than with only OD information. As expected, the combined information from cellpaths and ODs (the dotted lines) limits the degree of freedom more than cellpath information alone (the solid lines). As the number of cells is increased, the degree of freedom tends towards zero, at which point we can fully recover the route flow. These numerical results confirms the utility of cellular network data for addressing the traditionally highly underdetermined route flow estimation problem.

5.2. Full network, activity-based model

Using the full network scenario in §4.3, we perform experiments using the actual locations of PeMS static highway count sensors on 1033 links (about 5% coverage). We vary cell density from 250 to 8000 cells such that the proportions

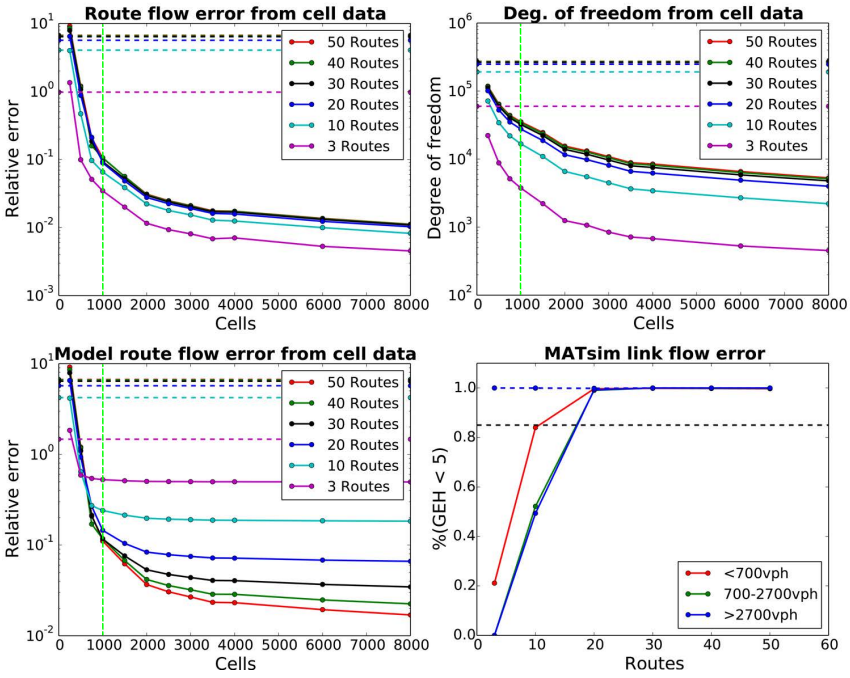

Fig. 8. Full (highway and arterial) network experiment results, corresponding to the regularized solution for the morning commute (rush hour). The top row is specific to the noiseless setting; the bottom row includes experiments with modeling error (noise). The green dotted vertical line highlights the results for 1000 cells, which is a reasonable setting for urban areas today. Top left: Route flow ϵ_r from OD demands (dotted) and cellpath flows (solid) for varying cell counts. The different curves indicate the number of routes (per OD) considered; Top right: Approximate degrees of freedom for the program with OD demands (dotted) and cellpath flows (solid) for varying cell counts. Bottom left: Including modeling error, the route flow ϵ_r from OD demands (dotted line) and cellpath flows (curves) for varying cell counts. Bottom right: Link flow error evaluated on all links ϵ_l^{full} without model error (dotted) and with model error (solid), shown for different link flow volume classes for 1000 cells. Best viewed in color.

are $N^B : N^L : N^S :: 3 : 1 : 16$.[11] The sub-regions $\mathcal{S}$ is given by the bounding boxes for the TAZ within the whole region. We analyze how the errors in route flows and link flows vary with the density of cell towers. Additionally, we study the effect of performing inference on only a subset of routes from our dataset.

Figure 8 presents the numerical results for Problem (2), where we compare the performance of route flow estimation via cellpath flow vs OD flow alone. To select a particular estimate from the solution space, we add an ℓ_2 regularization term to the objective. In our dataset, selecting the top 50 routes per OD pair was sufficient to account for 99.4% of trajectories; however, in general, the corresponding number of routes needed will vary based on the network, time of day, underlying driver behavior, etc. Thus, to represent these different settings, we present trade-off curves for accuracy when varying the number of routes from 3 up to 50. As expected, as more routes are considered by agents, the route flow accuracy ϵ_r declines, since the solution space (and its corresponding nullspace) grows. Fortunately, the accuracy increases with the number of cells. Thus, Figure 8 (top left) shows that the same level of accuracy may be attained when considering different numbers of routes (per OD pair) by varying also the number of cells. Our method performs comparably for the morning (shown in Figure 8) and evening (not shown) rush hours, achieving 89.5% and 89.9% route flow accuracy respectively and well exceeding the GEH test (with 1000 cells and 50 routes per OD), indicating the versatility of our method for diverse traffic settings. Figure 8 (bottom right) shows that we always achieve the link flow error $\epsilon_l^{full} = 1$ on all links (including those not observed) for various link volume classes, indicating that our method is effective for estimating link flows on unobserved links in noiseless settings.

Similarly to the highway network experiment, the accuracy in the estimates is closely related to the degree of freedom in Problem (2). For computational reasons, we compute an approximate measure of the degrees of freedom

[11] The I-210 region is 688mi^2 and, with cell towers spaced $\frac{1}{4}$ to 2 miles apart for suburban and urban areas, a reasonable range of cell towers for modern urban areas is 180 to 5500. We select 1000 for our baseline model.

by nullity$(AN) \geq |z| - \text{rank}(A)$, using notation from (11). Although the problem remains underdetermined (based on equality constraints in the noiseless setting), the accuracy increases substantially as the degrees of freedom decreases (Figure 8, top right). In all scenarios except the lowest cell configuration (250 cells), we observe that performing inference using cellpath flows (compared to using OD flow information) greatly improves the estimates of route flow.

However, selecting the top routes between each OD pair for a real network relies on sophisticated models and techniques. Though this article focuses on the noiseless setting, here we present preliminary results for a noisy setting, motivated by situations where not all top routes may be curated. That is, using the same flow measurements b, f, etc., we now estimate a route flow vector $\bar{x} \in \mathbb{R}^{|\overline{\mathcal{R}}|}$, where $\overline{\mathcal{R}} \subseteq \mathcal{R}$ denotes the curated routes; then, we compute the validation metrics as before, taking the corresponding entries of x^{true}. We call *modeling error* the route flow that is not modeled by the curated subset of routes. Figure 8 (bottom subfigures) shows an experiment where we consider the performance of our method where, among the top 50 routes (per OD), we are only able to curate the top 3-50 routes, and we evaluate our method in the presence of this modeling error. We observe that curating 20-50 routes (per OD) is sufficient for achieving a low ($< 15\%$) route flow error and also sufficient for performing well on the GEH metric on all links. Our preliminary results show promise for estimating route and link flows with our approach despite the challenge of selecting all routes that agents may take.

6. Conclusion

Our work demonstrates a data-driven method that is capable of estimating route-level flow accurately on a large scale network and is versatile to different vehicle behaviors. We address the traditionally highly underdetermined problem by introducing the concept of *cellpaths* for formalizing cellular network data as n-point network flows. We design a projected gradient algorithm suitable for the route flow estimation problem, as well as the traffic assignment problem. We validate our approach on several networks of varying sizes and underlying traffic assignment models, showing that the incorporation of cellular network data dramatically improves estimates over the use of traditional data sources by providing flow information on (coarse) routes.

Our methodology is highly compatible with past and present work in the transportation community. As route flows contain strictly more information than link flows and OD flows, which underlie many transportation methods, the potential for accurate route flow estimates in transportation applications is vast. Additionally, our solution method is shown to be compatible with related transportation problems, which may be combined for improved estimation.

Whereas work on traffic assignment, which models rather than estimates route flows, is critical for long-term land-use planning, strong model assumptions limit their application to short time-horizon applications. Taking a data-driven approach, our method enables new short time-horizon applications for the prediction and control of transportation such as route guidance, re-routing (e.g. minimizing effects of road closures, disasters, large events, etc.), demand prediction, and anomaly detection and analysis. Our framework aims to be widely deployable (wherever there is wide-spread cellular network coverage) and extendable, thereby providing a baseline estimator of the state of our current traffic networks, against which new controls and designs for intelligent transportation systems can compare.

The directions for future work are driven by our plans for integration with the decision support system for the I-210 corridor in California, US. We plan to analyze and improve the robustness of our model and methods in the presence of measurement error. Real loop sensors are notoriously noisy and a sizable fraction of them are offline at any given point. Since cellpath flow is not measured directly, but rather is inferred from cellular network data, it is prone to error from any inference procedure used. Fine-grained control applications will require even richer state estimates of the road network, for which we plan to extend our work to the dynamic setting. The full pipeline (summarized in Figure 1) will be implemented to perform large-scale route flow estimation using cellular network traces from AT&T and actual cell tower locations for the I-210 corridor in California, US.

Acknowledgement

The authors would like to thank Jason Du, Andrew Campbell, and Cathal Coffey for help in our experiments and visualizations. We are also grateful for insightful conversations with Professor Alexander Skabardonis, Professor Suvrit Sra, and Dr. Alexander Kurzhanskiy. We would like to thank our collaborators at AT&T: Dr. Jean-Francois Paiement, Dr. Jeffrey Pang, and Dr. Chris Volinsky. Finally, this work was supported in part by FORCES (Foundations Of Resilient CybEr-Physical Systems), which receives support from the National Science Foundation (NSF award numbers CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166).

References

Abrahamsson, T. 1998. Estimation of origin-destination matrices using traffic counts - a literature survey. Interim Report IR-98-021, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Baert, A.-E., & Seme, D. 2004. Voronoi mobile cellular networks: topological properties. In Third International Symposium on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, 29–35.

Bar-Gera. 2002. Origin-based Algorithm for the Traffic Assignment Problem. Transportation Science.

Beckmann, M., McGuire, C. B., & Winsten, C. B. 1956. Studies in the Economics of Transportation. Cowles Commission Monograph.

Bell, M., Shield, C., Busch, F., & Kruse, G. 1997. A stochastic user equilibrium path flow estimator. Transportation Research Part C: Emerging Technologies, **5**(34), 197 – 210.

Bell, M. G. H., & Iida, Y. 1997. Transportation Network Analysis. Wiley, West Sussex, United Kingdom.

Best, M. J., & Chakravarti, N. 1990. Active set algorithms for isotonic regression; a unifying framework. Math. Programming, **47**, 425–439.

Blandin, S., Ghaoui, L. E., & Bayen, A. 2009. Kernel regression for travel time estimation via convex optimization. IEEE Conference on Decision and Control.

Boyd, S., & Vandenberghe, L. 2004. Convex Optimization. Cambridge University Press.

Caceres, N., Wideberg, J. P., & Benitez, F. G. 2007. Deriving origin-destination data from a mobile phone network. IET Intell. Transp. Syst, **1**, 15–26.

Calabrese, F., Lorenzo, G. Di, Liang, L., & Ratti, C. 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing, 36–44.

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A.-L. 2008a. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical, **41**.

Candia, J., Gonzalez, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabasi, A-L. 2008b. Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical.

Castillo, E., Menendez, J. M., & Jimenez, P. 2008. Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. Transportation Research Part B: Methodological, **42**, 455–481.

Castillo, E., Gallego, I., Menendez, J. M., & Rivas, A. 2010. Optimal Use of Plate-Scanning Resources for Route Flow Estimation in Traffic Networks. IEEE Transactions on Intelligent Transportation Systems, **11**, 380–391.

Choe, T., Skabardonis, A., & Varaiya, P. 2002. Freeway performance measurement system (PeMS): an operation tool. 81st Annual Meeting Transportation Research Board, Washington, DC.

Daganzo, C. F., & Sheffi, Y. 1977. On stochastic models of traffic assignment. Transportation Science, **11**, 253–274.

Dowling, R., Skabardonis, A., & Alexiadis, V. 2004. Traffic analysis toolbox volume III: guidelines for applying traffic microsimulation modeling software. Tech. rept.

Duchi, J., Gould, S., & Koller, D. 2008. Projected subgradient methods for learning sparse gaussians. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence.

Fisk, C. 1980. Some developments in equilibrium traffic assignment. Transportation Research Part B, **14**, 243–255.

Ford, L. R., & Fulkerson, D. R. 1962. Flows in Networks. Princeton University Press, Princeton, NJ.

Grotzinger, S. J., & Witzgall, C. 1984. Projection onto Order Simplexes. Applied Mathematics and Optimization, **12**, 247–270.

Hato, E., Taniguchi, M., Sugie, Y., Kuwahara, M., & Morita, H. 1999. Incorporating an information acquisition process into a route choice model with multiple information sources. Transportation Research Part C, **7**, 109–129.

Herrera, J.-C., Work, D. B., Herring, R., Ban, J., Jacobson, Q., & Bayen, A. M. 2009. Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century experiment. Transportation Research Part C, **18**, 568–583.

Hunter, T., Herring, R., Abbeel, P., & Bayen, A. 2009. Path and travel time inference from GPS probe vehicle data. NIPS Analyzing Networks and Learning with Graphs.

Illenberger, J., G. Fltterd, & Nagel, K. 2007. Enhancing MATSim with capabilities of within-day re-planning. IEEE Intelligent Transportation Systems Conference.

J. Duchi, S. Shalev-Shwartz, Y. Singer T. Chandra. 2008. Efficient Projections onto the l1-Ball for Learning in High Dimensions. Proceedings of the 25th International Conference on Machine Learning.

Janecek, A., Hummel, A. A., Valerio, D., Ricciato, F., & Hlavacs, H. 2012. Cellular data meet vehicular traffic theory: location area updates and cell transitions for travel time estimation. Pages 361–370 of: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM.

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. Page 2 of: Proc. of the 2nd ACM SIGKDD International Workshop on Urban Computing. ACM.

Kelly, F. P. 1991. Network routing. Philosophical Transactions: Physical Sciences and Engineering, **337**, 343–367.

LeBlanc, L. J., Morlok, E. K., & Pierskalla, W. P. 1975. An efficient approach to solving the road network equilibrium traffic assignment problem. Transportation Research, **9**, 309–318.

Maher, M. J., & Hughes, P. C. 1997. A probit-based stochastic user equilibrium assignment model. Transportation Research, **31**, 341–355.

Mardani, M., & Giannakis, G. B. 2013. Robust network traffic estimation via sparsity and low rank. IEEE International Conference on Acoustics, Speech and Signal Processing.

Mathew, J., & Xavier, P.M. A SURVEY ON USING WIRELESS SIGNALS FOR ROAD TRAFFIC DETECTION.

Monderer, D., & Shapley, L. S. 1996. Potential Games. Games and Economic Behavior, **14**, 124–143.

Nocedal, J., & Wright, S. 2006. Numerical Optmization. Springer, 2nd edition.

of Transportation (WisDOT), Wisconsin Department. 2013. Unofficial WI Traffic Analysis Guidelines, Draft. `http://www.wisdot.info/microsimulation/index.php?title=Model_Calibration`. [Online; accessed 2014-08-30].

Ortuzar, J. de D., & Willumsen, L.G. 2001. Modelling Transport. 3rd, Edition, Wiley, West Sussex, United Kingdom.

Patire, A., Wright, M., Prodhomme, B., & Bayen, A. 2013. How much GPS data do we need? Transportation Research Part C.

Rahmani, M., & Koutsopoulos, H. N. 2013. Path inference from sparse floating car data. Transportation Research Part C: Emerging Technologies, **30**, 41–54.

Roughgarden, T. 2003. The price of anarchy is independent of the network topology. Journal of Computer and System Sciences, **67**, 341–364.

Sheffi, Y. 1985. Urban Transportation Networks. Prentice-Hall, Englewood Cliffs, NJ.

Shen, W., & Wynter, L. 2012. A new one-level convex optimization approach for estimating origindestination demand. Transportation Research Part B: Methodological, **46**, 1535–1555.

Tettamanti, T., Demeter, H., & Varga, I. 2012. Route Choice Estimation Based on Cellular Signaling Data. Acta Polytechnica Hungarica, **9**(4), 207–220.

Tibshirani, R. J., Hoefling, H., & Tibshirani, R. 2011. Nearly-Isotonic Regression. Technometrics, **53**.

Toole, JL., Ulm, M., González, M.C., & Bauer, D. 2012. Inferring land use from mobile phone activity. Pages 1–8 of: Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

Veeraraghavan, H., Masoud, O., & Papanikolopoulos, N. 2003. Computer Vision Algorithms for Intersection Monitoring. IEEE Transactions on Intelligent Transportation Systems, **4**, 78–89.

Volinsky, C., Becker, R., Caceres, R., Hanson, K., Loh, J., Urbanek, S., & Varshavsky, A. 2011a. Clustering Anonymized Mobile Call Detail Records to Find Usage Groups. 1st Workshop on Pervasive Urban Applications (PURBA).

Volinsky, C., Varshavsky, A., Becker, R., Loh, J., Urbanek, S., Caceres, R., & Hanson, K. 2011b. Route Classification using Cellular Handoff Patterns. 13th ACM International Conference on Ubiquitous Computing.

Wang, W., & Carreira-Perpin, M. . 2013. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. CoRR.

Wardrop, J. G., & Whitehead, J. I. 1952. Correspondence. Some Theoretical Aspects of Road Traffic Research. ICE Proc: Engineering Divisions 1.

White, J., & Wells, I. 2002. Extracting origin destination information from mobile phone data. 11th Int. Conf. on Road Transport Information and Control, London, 30–34.

Work, D., Tossavainen, O.-P., Blandin, S., Bayen, A., Iwuchukwu, T., & Tracton, K. 2008. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. 47th IEEE Conference on Decision and Control.

Yadlowsky, S., Thai, J., Wu, C., Pozdnukhov, A., & Bayen, A. 2014. Link Density Inference from Cellular Infrastructure. Transportation Research Board (TRB) 94th Annual Meeting.

Yen, J. Y. 1971. Finding the k Shortest Loopless Paths in a Network. Management Science, **17**, 712–716.