Predicting traffic volumes and estimating the effects of shocks in large transportation systems

Edo Airoldi
Department of Statistics
Harvard University

LIDS Smart Urban Infrastructures Workshop, MIT, May 2017

Ill-posed inverse problems in London

PNAS PNAS PNAS PNAS

Predicting traffic volumes and estimating the effects of shocks in massive transportation systems

Ricardo Silva[a,1], Soong Moon Kang[b], and Edoardo M. Airoldi[c]

[a]Department of Statistical Science and Centre for Computational Statistics and Machine Learning, University College London, London WC1E 6BT, United Kingdom; [b]Department of Management Science and Innovation, University College London, London WC1E 6BT, United Kingdom; and [c]Department of Statistics, Harvard University, Cambridge, MA 02138

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved March 20, 2015 (received for review July 8, 2014)

Public transportation systems are an essential component of major cities. The widespread use of smart cards for automated fare collection in these systems offers a unique opportunity to understand passenger behavior at a massive scale. In this study, we use network-wide data obtained from smart cards in the London transport system to predict future traffic volumes, and to estimate the effects of disruptions due to unplanned closures of stations or lines. Disruptions, or shocks, force passengers to make different decisions concerning which stations to enter or exit. We describe how these changes in passenger behavior lead to possible overcrowding and model how stations will be affected by given disruptions. This information can then be used to mitigate the effects of these shocks because transport authorities may prepare in advance alternative solutions such as additional buses near the most affected stations. We describe statistical methods that leverage the large amount of smart-card data collected under the natural state of the system, where no shocks take place, as variables that are indicative of behavior under disruptions. We find that features extracted from the natural regime data can be successfully exploited to describe different disruption regimes, and that our framework can be used as a general tool for any similar complex transportation system.

smart cities | transportation | regime change | complex systems

Well-designed transportation systems are a key element in the economic welfare of major cities. Design and planning of these systems requires a quantitative understanding of traffic patterns and relies on the ability to predict the effects of disruptions to such patterns, both planned and unplanned (1).

There is a long history of analytic and modeling approaches to the study of traffic patterns (2), for example using simulated scenarios in simple transportation systems (3), and analysis of real traffic data in complex systems, either focusing on a small samples (4) or using more aggregate data (5, 6). Here we take this approach to the next level by making use of smart-card data and incident logs to (*i*) predict traffic patterns and (*ii*) estimate the effect of unplanned disruptions on these patterns. We analyzed 70 d of smart-card transactions from the London transportation network, composed of ~10 million unique IDs and 6 million transactions per day on average, resulting in one of the largest statistical analyses of transportation systems to date.

A related literature deals with various aspects of dynamics in complex networks and complex systems in general (7–9), using a variety of data sources, from emails (10) to the circulation of bank notes (11) to online experiments on Amazon Turk (12). More recently, a number of analyses have leveraged mobile phone data as proxies for mobility (4, 13–15).

However, smart-card technology allows us to obtain large

Transport for London Data

The London transportation system is composed of several connected subsystems. We focus on the Underground, Overground, and Docklands Light Rail (DLR), all of which are train services aimed at fast commuting within the Greater London area only. A map of the system is provided in Fig. S1.

Transport for London (TfL) provided us with smart-card readings covering 70 d, from February 2011 to February 2012. Smart-card readings comprise more than 80% of the total number of journeys (18). Each reading consists of a time stamp, a location code, and an event code. The location code uniquely identifies each of the 374 stations of the system that were active during the months covered by our data. The two events of our interest are generated when a passenger touches the smart-card reader at the entrance ("tap-in" event) or at the exit ("tap-out" event) of a station. Passenger IDs are anonymized and ignored in our analysis. We discarded all tap-in readings that are not matched to a tap-out, and vice-versa. Time resolution of the recorded time stamps is 1 min. Each day is composed of 1,200 min, starting at 5:00 AM until 1:00 AM of the next calendar day. Our analysis covers weekdays only. Weekdays are assumed to be exchangeable (see Fig. S2).

Overview of Analysis

We show that we can reliably predict passenger origin–destination (OD) traffic by combining around 140,000 nonparametric statistical models with hundreds of millions of smart-card data events. We also show that the same model provides features that explain behavior under a shock (or "disruption") to the system, defined as an unanticipated period during which a station or a line is (partially) closed down. The resulting model allows us to predict the outcome of disruptions and to evaluate stations by how prone to overcrowding they are given disruptions at peak time.

Significance

We propose a new approach to analyzing massive transportation systems that leverages traffic information about individual travelers. The goals of the analysis are to quantify the effects of shocks in the system, such as line and station closures, and to predict traffic volumes. We conduct an in-depth statistical analysis of the Transport for London railway traffic system. The proposed methodology is unique in the way that past disruptions are used to predict unseen scenarios, by relying on simple physical assumptions of passenger flow and a system-wide model for origin–destination movement. The method is scalable, more accurate than blackbox approaches, and generalizable to other complex transportation systems. It therefore offers important insights to inform policies on urban transportation.

SOCIAL SCIENCES

STATISTICS

Estimating Latent Processes on a Network From Indirect Measurements

Edoardo M. AIROLDI and Alexander W. BLOCKER

In a communication network, point-to-point traffic volumes over time are critical for designing protocols that route information efficiently and for maintaining security, whether at the scale of an Internet service provider or within a corporation. While technically feasible, the direct measurement of point-to-point traffic imposes a heavy burden on network performance and is typically not implemented. Instead, indirect aggregate traffic volumes are routinely collected. We consider the problem of estimating point-to-point traffic volumes, $\boldsymbol{x}_t$, from aggregate traffic volumes, $\boldsymbol{y}_t$, given information about the network routing protocol encoded in a matrix A. This estimation task can be reformulated as finding the solutions to a sequence of *ill-posed* linear inverse problems, $\boldsymbol{y}_t = A\,\boldsymbol{x}_t$, since the number of origin-destination routes of interest is higher than the number of aggregate measurements available.

Here, we introduce a novel multilevel state-space model (SSM) of aggregate traffic volumes with realistic features. We implement a naïve strategy for estimating unobserved point-to-point traffic volumes from indirect measurements of aggregate traffic, based on particle filtering. We then develop a more efficient two-stage inference strategy that relies on model-based regularization: a simple model is used to calibrate regularization parameters that lead to efficient/scalable inference in the multilevel SSM. We apply our methods to corporate and academic networks, where we show that the proposed inference strategy outperforms existing approaches and scales to larger networks. We also design a simulation study to explore the factors that influence the performance. Our results suggest that model-based regularization may be an efficient strategy for inference in other complex multilevel models. Supplementary materials for this article are available online.

KEY WORDS: Approximate inference; Ill-posed inverse problem; Multilevel state-space model; Multistage estimation; Network tomography; Origin-destination traffic matrix; Particle filtering; Polytope sampling; Stochastic dynamics.

1. INTRODUCTION

A pervasive challenge in multivariate time series analysis is the estimation of nonobservable time series of interest $\{\boldsymbol{x}_t : t = 1, \ldots, T\}$ from indirect noisy measurements $\{\boldsymbol{y}_t : t = 1, \ldots, T\}$, typically obtained through an aggregation or mixing process, $\boldsymbol{y}_t = \boldsymbol{a}(\boldsymbol{x}_t)\ \forall t$. The inference problem that arises in this setting is often referred to as an *inverse*, or *deconvolution*, problem (e.g., Hansen 1998; Casella and Berger 2001; Meister 2009) in the statistics and computer science literatures, and qualified as *ill-posed* because of the lower dimensionality of the measurement vectors with respect to the nonobservable estimands of interest. Ill-posed inverse problems lie at the heart of a number of modern applications, including image super-resolution and positron emission tomography where we want to combine many two-dimensional images in a three-dimensional image consistent with two-dimensional constraints (Shepp and Kruskal 1978; Vardi, Shepp, and Kaufman 1985); blind source separation where there are more sound sources than sound tracks (i.e., the measurements) available (Liu and Chen 1995; Lee et al. 1999; Parra and Sajda 2003); and inference on cell values in contingency tables where two-way or multiway margins are prespecified (Bishop, Fienberg, and Holland 1975; Dobra, Tebaldi, and West 2006).

We consider a setting in which high-dimensional multivariate time series $\boldsymbol{x}_{1:T}$ mix on a network. Individual time series correspond to traffic directed from a node to another. The aggregation process encodes the routing protocol—whether deterministic of probabilistic—that determines the path traffic from any given source follows to reach its destination. This type of mixing can be specified as a linear aggregation process A. This problem setting leads to the following sequence of ill-posed linear inverse (or deconvolution) problems,

$$\boldsymbol{y}_t = A\,\boldsymbol{x}_t, \quad \text{s.t. } \boldsymbol{y}_t, \boldsymbol{x}_t \geq 0 \quad \text{for } t = 1, \ldots, T, \qquad (1)$$

since the observed aggregate traffic time series are low dimensional, $\boldsymbol{y}_t \in \mathbb{R}^m$, while the latent point-to-point traffic time series of interest are high dimensional, $\boldsymbol{x}_t \in \mathbb{R}^n$. Thus, the matrix $A_{m\times n}$ is rank deficient, $r(A) = m < n$, in this general problem setting.

The application to communication networks that motivates our research is *(volume) network tomography*; an application originally introduced by Vardi (1996), which has quickly become a classic (see, e.g., Vanderbei and Iannone 1994; Tebaldi and West 1998; Cao et al. 2000; Coates et al. 2002; Medina et al. 2002; Liang and Yu 2003a; Zhang et al. 2003b; Airoldi and Faloutsos 2004; Castro et al. 2004; Lakhina et al. 2004; Lawrence et al. 2006b; Fang, Vardi, and Zhang 2007; Blocker and Airoldi 2011). An established engineering practice is at the root of the inference problem in network tomography. Briefly, the availability of point-to-point (or origin-destination (OD)) traffic volumes over time is critical for reliability analysis (e.g., predicting flows and failures), traffic engineering (e.g., minimizing congestion), capacity planning (e.g., forecasting re-

Agenda

- Transport for London (TfL)
- Sampling in extremely constrained spaces
- Concluding remarks

Main objective

- To provide an estimate of **passenger behaviour** when an **unplanned closures** take place in a **origin-destination (OD)** transportation system
  - Passenger behaviour: number of exits in a region of interest (e.g., “tap-outs” in the Tube)
  - Unplanned closures: interruptions of service in **lines** and **stations** due to incidents (e.g., as reported by the TfL twitter account)
  - OD system: origin and destination of passenger is observed (via Oyster cards)

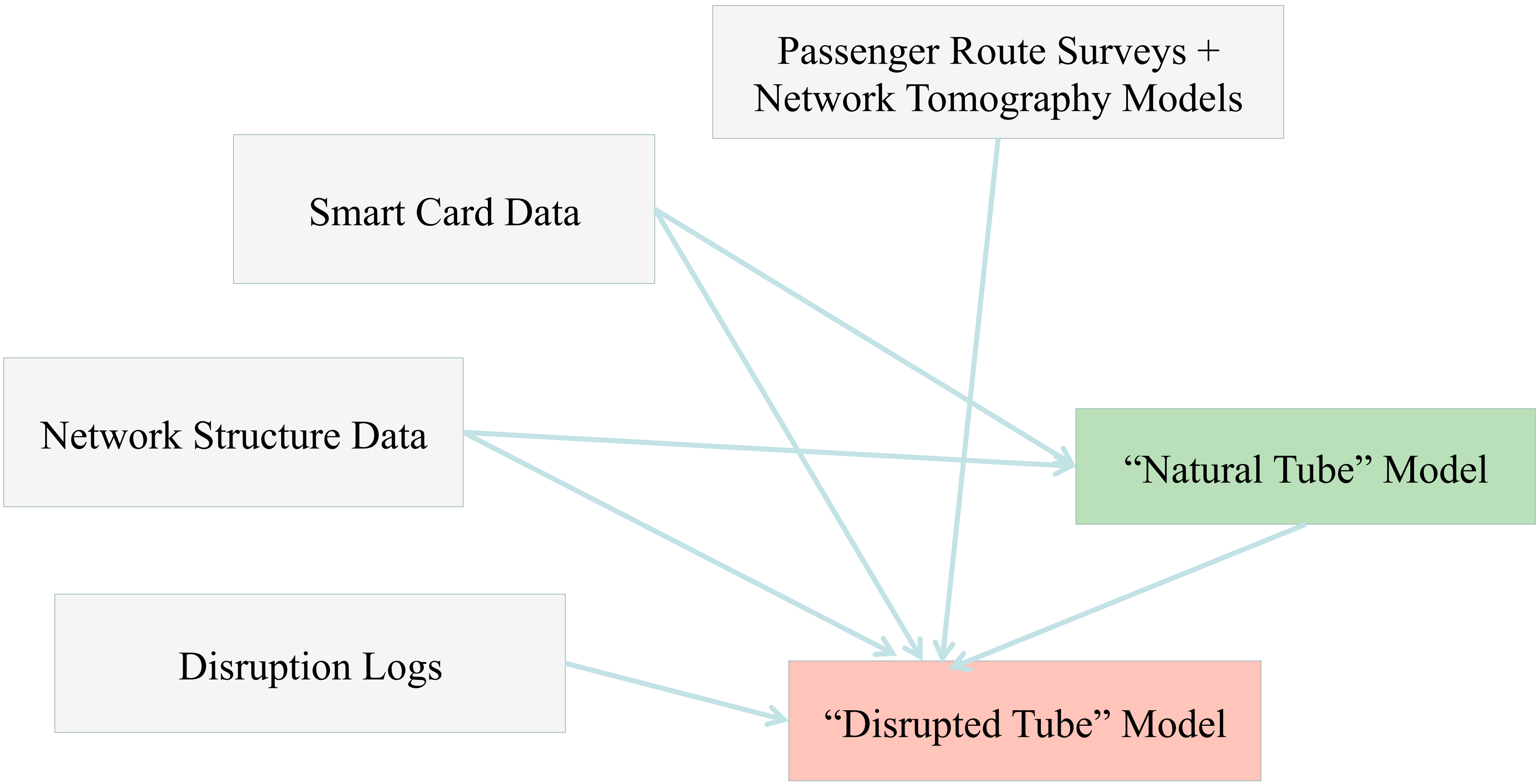
Approach

1. Build a model for origin-destination counts for all 374^2 pairs and every minute of the day in the **natural regime** (i.e., no unplanned closures)
2. Use these models to generate **counterfactual** behaviour during disruption times
   – Expected OD counts had no disruption taken place
3. Use the counterfactual behaviour as explanatory features of observed behaviour **under disruption** using a battery of linear models

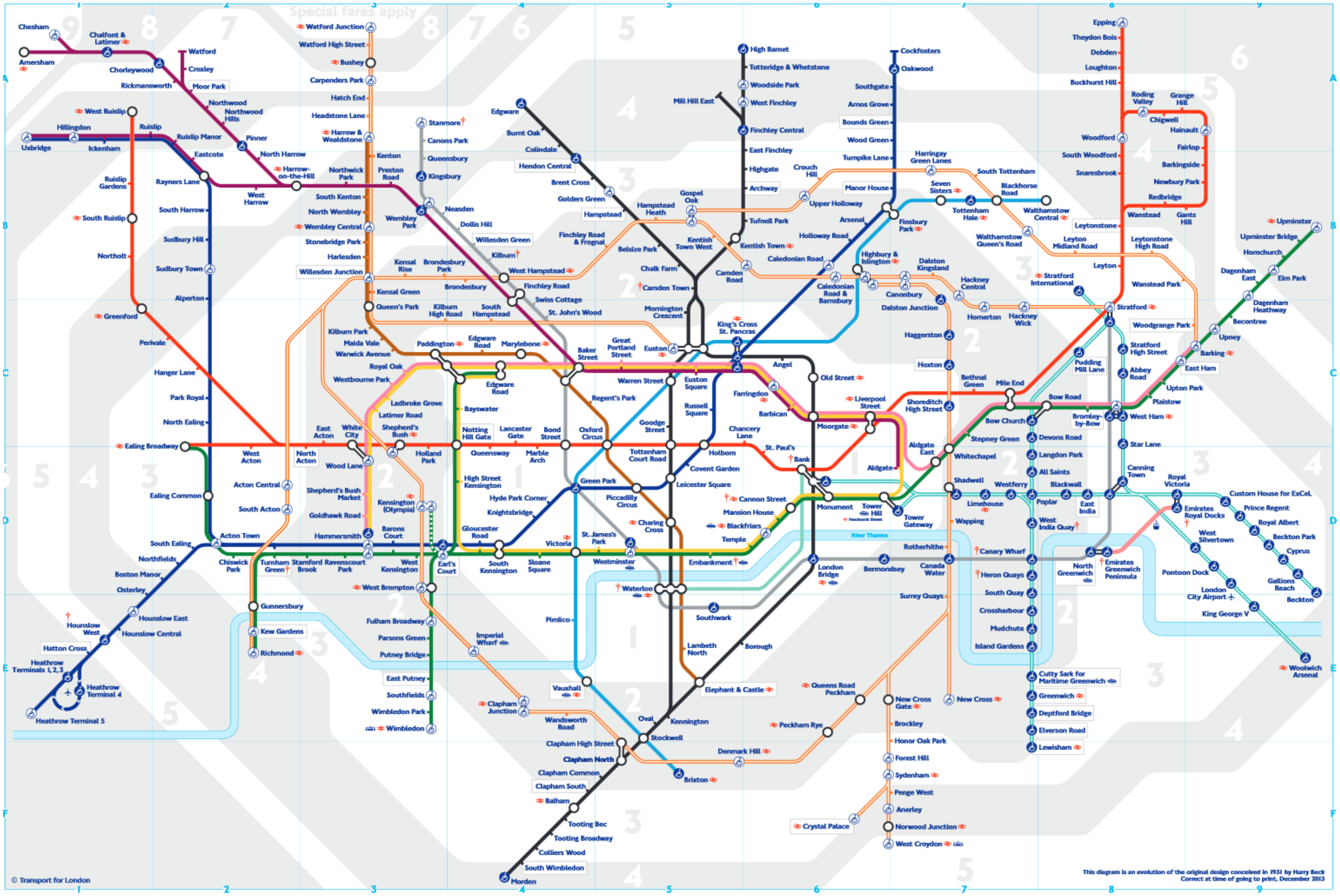
Findings

- A hierarchical model for origin-destination-time can be built with computationally and statistically simple building blocks which is robust for prediction
  - No hidden states, combination of 100,000+ nonparametric building blocks fit to 300,000,000+ smart card tap events
- Behaviour under the natural regime, plus features derived from **flow measures** (i.e., solutions to IPIP) explain much of the behaviour under disruption

Overview of data and models

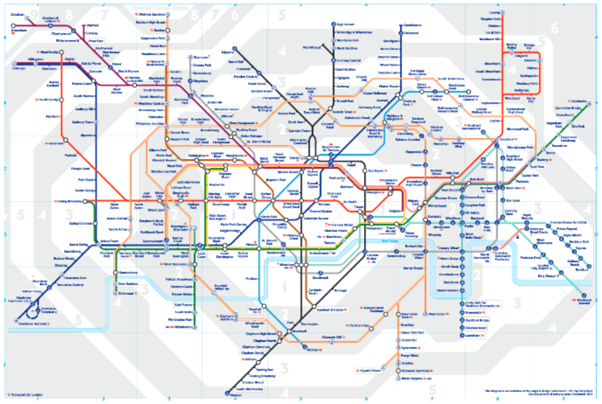


The Tube map



http://www.tfl.gov.uk/assets/downloads/standard-tube-map.pdf

Structural data



- There are **lines** and **stations**
  - Underground lines, Overground lines, and DLR
- Stations can belong to **multiple lines**
  - When there is a change of system (e.g., Stratford Underground vs Stratford DLR vs Stratford Overground)
  - Physically disjoint stations may have the same name (e.g., Edgeware Road, Hammersmith)
- Code as a **directed network**, with stations as nodes, different nodes for stations with multiple Oyster IDs
- Stations are given physical locations too

Smart Card data

- Individual, anonymized, Oyster card IDs
- History of **taps**:
  - Event (IN/OUT, among others)
  - Location, date, time of the day (1-min resolution)
- Some measurement errors
  - Staff cards included, but not labeled
  - It is possible to leave some stations without tapping out
- Change of stations within connections are not usually recorded

Disruption logs

- Official problem reports in the Tube, mostly **free text**

  E.g. "No service Finchley Road to Waterloo due to a faulty train at Baker Street. MINOR DELAYS on the rest of the line."

- **Line ID** provided, plus **starting/ending time** (seconds)
- Directionality:

  "No service West Hampstead to Stanmore **northbound** only due to a fire alert at Willesden Green. MINOR DELAYS on the rest of the line."

Disruption data

- 793 data points, over 70 week days to avoid complication due to seasonality
- Each data point corresponds to the outcome at a particular station / particular disruption
- If one disruption involves several stations of interest, it provides us with several data points
- We extract indicator variable for minor and major delays from disruption logs

Rolling origin-destination survey

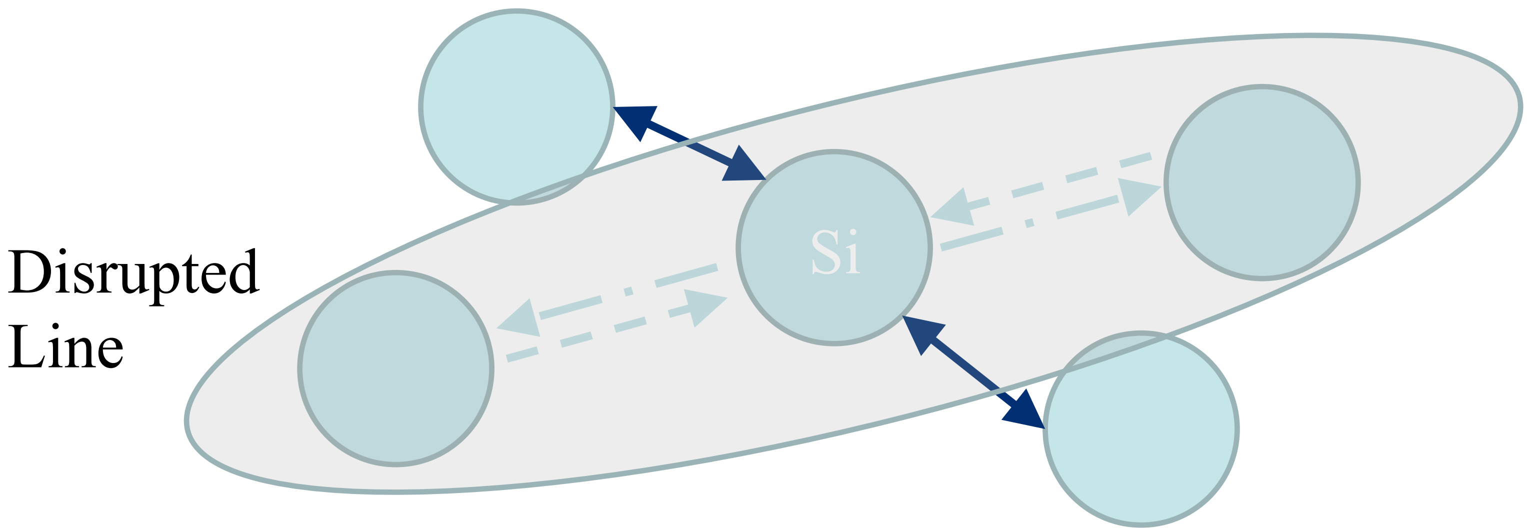
- TfL surveys with passengers regularly, who indicate frequency of routes chosen
  - Recall Oyster cards record origin-destination only
- 2012-2013 records, around 100,000 counts
- Often surveys do not indicate full route, just points of change (implies 2 types of prior information)

Basic modeling idea

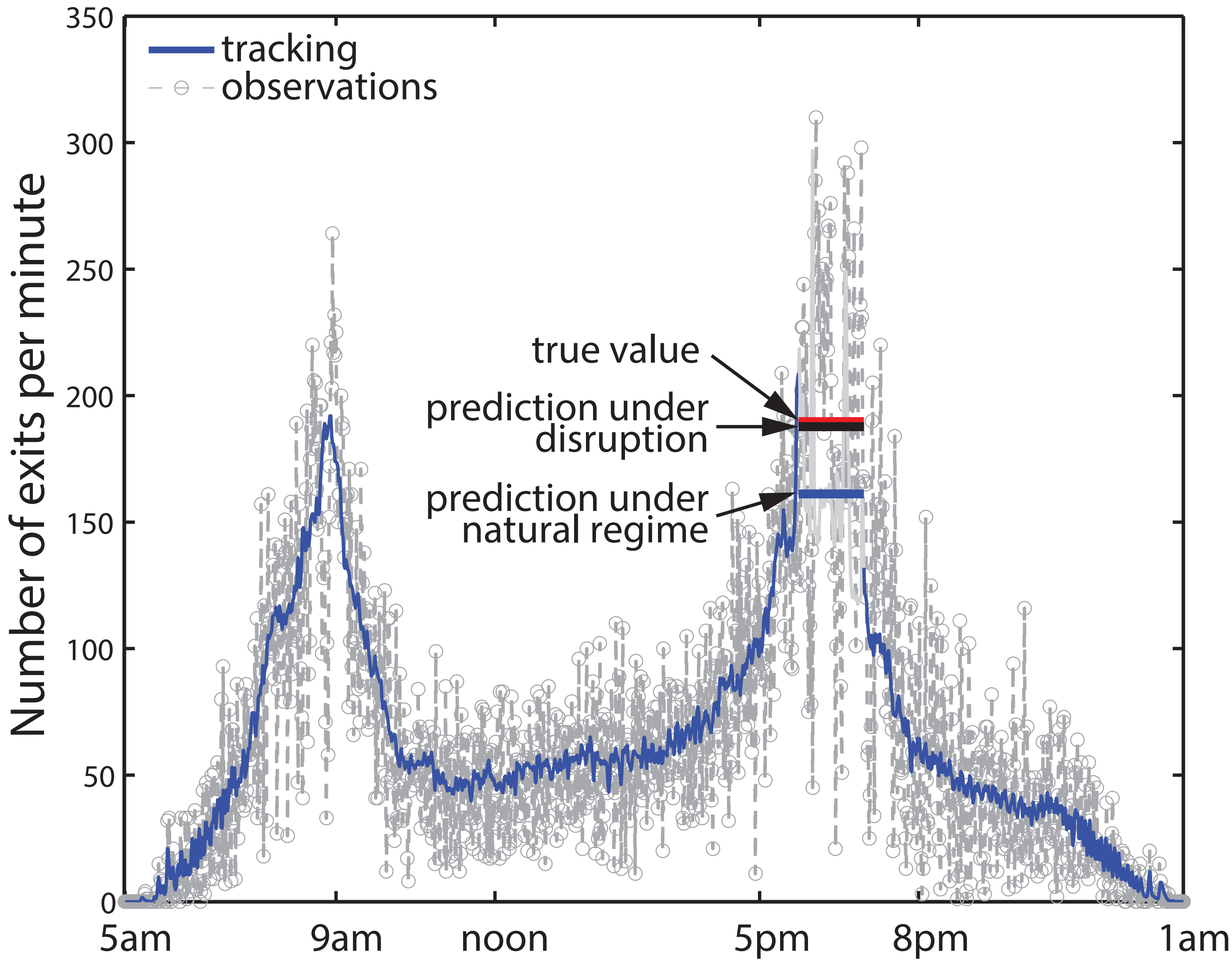
- Let S_i be a station of interest within disrupted line:

EXPECTED EXITS UNDER DISRUPTION(i) =

EXPECTED NATURAL NUMBER OF EXITS(i)
– MISSING INFLOW(i)
+ MISSING OUTFLOW(i)



Example prediction under disruption

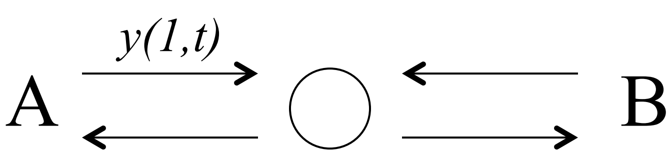


Agenda

- Transport for London (TfL)
- Sampling in extremely constrained spaces
- Concluding remarks

Mathematical formulation

Given a collection of tap-ins and tap-outs $Y_{(m \times t)}$ and a probabilistic routing matrix $A_{(m \times n)}$, infer latent OD counts, $X_{(n \times t)}$, such that $Y = A \cdot X$, where $n > m$.



$$\begin{bmatrix} y(1,t) \\ y(2,t) \\ y(3,t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{array}{cl} \begin{bmatrix} x(1,t) \\ x(2,t) \\ x(3,t) \\ x(4,t) \end{bmatrix} & \begin{array}{l} \text{AA} \\ \text{AB} \\ \\ \\ \end{array} \end{array}$$

Geometry of ill-posed inverse problems

Consider $\mathrm{Y}_{(\mathrm{m}\times 1)} = \mathrm{A}_{(\mathrm{m}\times \mathrm{n})}\,\mathrm{X}_{(\mathrm{n}\times 1)}$

Given Y,A we want to find X

- Rewrite $\mathrm{A}=[\mathrm{A}_1|\mathrm{A}_2]$ with $\mathrm{r}(\mathrm{A}_1)=\mathrm{m}$, and $\mathrm{x}=[\mathrm{x}_1|\mathrm{x}_2]$
- The posterior is $p(x \mid y,\lambda) \propto p(x_2 \mid y,\lambda) \cdot I_{f(y,A,x_2)}(x_1)$
- Part of the estimand is $x_1 = A_1^{-1}(y - A_2 x_2)$

Solutions x_2 lie in the intersection of a linear space of dim. n-m with the positive orthant: *a convex polytope*

New idea: Polytope samplers

Strategy

1. Leverage HNF to find first vertex
2. Greedily move along the edges to find all vertices (via HNF pivoting)
3. Place a distribution on the polytope; we develop three strategies to do this using Dirichlet pdf

Polytope samplers provide a new exact sampling strategy for inference in ill-posed inverse problems

Hermite normal form

- Hermite Normal Form of integer matrix A: B=AQ
- Columns of Q_2 generates null-space

$$m\underset{n}{\begin{bmatrix} A_1 & A_2 \end{bmatrix}} \overset{\begin{matrix} Q_1 & \qquad Q_2 \end{matrix}}{\underset{\begin{matrix} m & \qquad n-m \end{matrix}}{\begin{bmatrix} A_1^{-1} & -A_1^{-1}A_2 \\ 0 & I \end{bmatrix}}} = \overset{B_1}{\underset{\begin{matrix} m & \qquad n-m \end{matrix}}{\begin{bmatrix} I & 0 \end{bmatrix}}}$$

Finding the first vertex (almost)

Start from $y = Ax$, with A of size mxn

Define $x' = Q^{-1}x$

Then rewrite $y = AQQ^{-1}x = AQx'$

Notice that $AQ = [I_m \mid 0\ ]$

So $x' = [y \mid 0]$ is a solution!

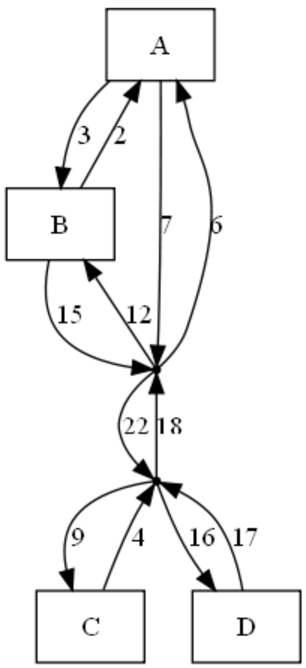
Caveats apply, but it turns out this is a good start

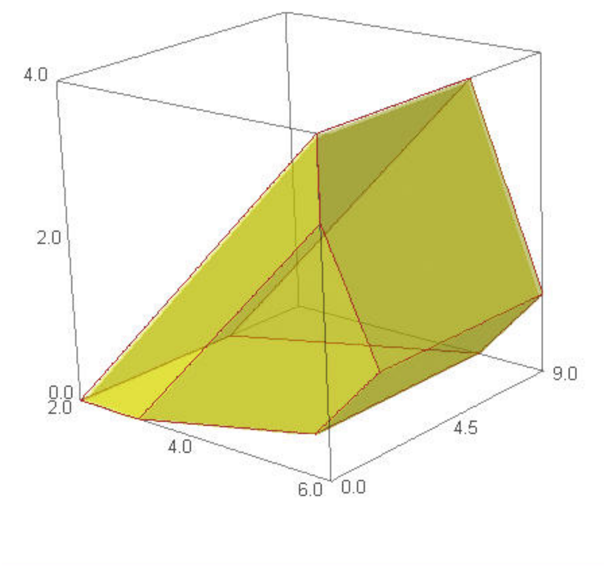
Distributions on a convex polytope

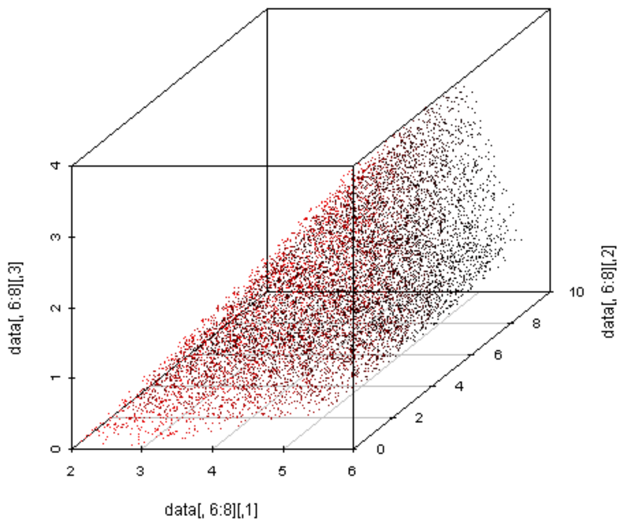
1. Lift the polytope into a higher dimensional simplex, posit a Dirichlet, project back
2. Triangulate the convex polytope into simplices, posit a collection of Dirichlet distributions weighted by their volumes
3. Direct generalization of Dirichlet that leverages moment map and projective geometry

Illustrative example

Matrix A is 9x12 and leads to a 3D solution space







Agenda

- Transport for London (TfL)
- Sampling in extremely constrained spaces
- Concluding remarks

Take home points

- Massive data on individual passengers available
- Lots of opportunities for impact
  - Assessing/predicting overcrowding, monitoring/routing
  - Planning minimally-disrupting closures for safety
  - Validating standard assumptions (congestion models, …)
- Ill-posed inverse problem
  - Samplers based on geometry of polytopes, and much more (e.g., Fienberg-Fréchet sharp bounds on OD flows)

Thanks