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

Ricardo Silva, Soo Moon Kang, and Edoardo M. Airoldi

Transport for London (TfL) provided us with smart-card readings covering 70 days, from February 2011 to February 2012. Smart-card readings comprise more than 90% of the total number of journeys (18). Each reading consists of a timestamp, 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 ("nEnter") or exit ("Exe") of a station. Passenger IDs are anonymized and ignored in our analysis. We discarded all_origin readings that we were unable to match at a stack, 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 11:00 AM of the next calendar day. Our analysis covers weekdays only. Weekends are assumed to be exchangeable (see Fig. S2).

Overview

We show that we can reliably predict passenger origin-destination (OD) traffic by combining several individual trip statistics, statistical models with hundreds of millions of smart-card data events. We also derive a novel model for explaining 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 close they are to being overwhelmed by passengers who have been given up at peak time.

Significance

We propose a new approach to analyzing massive transportation systems that leverages traffic information about individual travelers. The goal of the analysis is to quantify the effects of shocks to the system, such as line and station closures, and to estimate traffic volumes. We present our work at the 6th International Conference on 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 and utilizing data from validators. The method is scalable, more accurate than traditional methods, and generalizes well to other complex applications. It therefore offers important insights to inform policies on urban transportation.
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

- **Smart Card Data**
- **Network Structure Data**
- **Disruption Logs**

**Passenger Route Surveys + Network Tomography Models**

**“Natural Tube” Model**

**“Disrupted Tube” Model**
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:

$$\text{EXPECTED EXITS UNDER DISRUPTION}(i) =$$

$$\text{EXPECTED NATURAL NUMBER OF EXITS}(i) - \text{MISSING INFLOW}(i) + \text{MISSING OUTFLOW}(i)$$
Example prediction under disruption

- Tracking observations
- True value
- Prediction under disruption
- Prediction under natural regime

Number of exits per minute

- 5am
- 9am
- Noon
- 5pm
- 8pm
- 1am
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 $n > m$. 

\[
\begin{align*}
\begin{bmatrix}
y(1, t) \\
y(2, t) \\
y(3, t)
\end{bmatrix}
&= 
\begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
x(1, t) \\
x(2, t) \\
x(3, t) \\
x(4, t)
\end{bmatrix}
\end{align*}
\]
Consider \( Y_{(m \times 1)} = A_{(m \times n)} X_{(n \times 1)} \)

Given \( Y, A \) we want to find \( X \)

- Rewrite \( A = [A_1|A_2] \) with \( r(A_1) = m \), and \( x = [x_1|x_2] \)

- The posterior is \( p(x | y, \lambda) \propto p(x_2 | y, \lambda) \cdot I_{f(y,A,x_2)}(x_1) \)

- 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

\[
\begin{array}{ccc}
\begin{array}{ccc}
Q_1 & & Q_2 \\
A_1^{-1} & -A_1^{-1}A_2 & \\
\end{array}
\end{array} = \begin{array}{ccc}
\begin{array}{ccc}
B_1 & & \\
0 & 1 & \\
m & m & n-m \\
\end{array}
\end{array}
\]
Finding the first vertex (almost)

Start from $y = Ax$, with $A$ of size $m \times n$

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

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

Notice that $AQ = \begin{bmatrix} I_m & 0 \end{bmatrix}$

So $x' = \begin{bmatrix} y \mid 0 \end{bmatrix}$ 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