Modelling Disruptions and Latent Structure in the London Underground

Ricardo Silva, UCL and the Alan Turing Institute

Joint work with Nicolo Colombo, Soong Kang and Edoardo Airoldi

Computational Statistics and Machine Learning

- Solving problems of **prediction** and **latent structure estimation**.
  - "Supervised" and "unsupervised" learning.

- Judicious use of algorithms and computational resources to target entire systems.

How Can This Help a Transport Authority such as TfL?

- Estimating local loads:
  - train schedule optimization, day-to-day/within-a-day patterns, ...
- Expected travel times:
  - congestions/failures detection, personalized planning, ...
- Route preferences:
  - single-user modelling, system's behaviour under disruptions,...
- All leading to better journeys, network maintenance planning, rail-replacement buses, simpler life for TfL staff, money savings...

This Talk

- A sample of some past and on-going computational stats/machine learning research in my group, tackling passenger behavior in systems like the London Underground.

Task 1

- To provide an estimate of **passenger behaviour** when an **unplanned closure** takes 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 closure: interruption of service in lines and stations due to incidents (e.g., as in you can see in the TfL twitter account)
  - OD system: origin and destination of anonymised passenger is observed (Oyster card)

Task 2

- To reconstruct **unobservable passenger behaviour** from millions of anonymised origin-destination data using a scalable approach.

  - Unobservable behaviour: **route choice** and **location** between particular origin-destinations at particular times of the day.

Task 3 (Work in progress)

- Generalisations of the original problem of behavior under disruption, to capture more refined information.

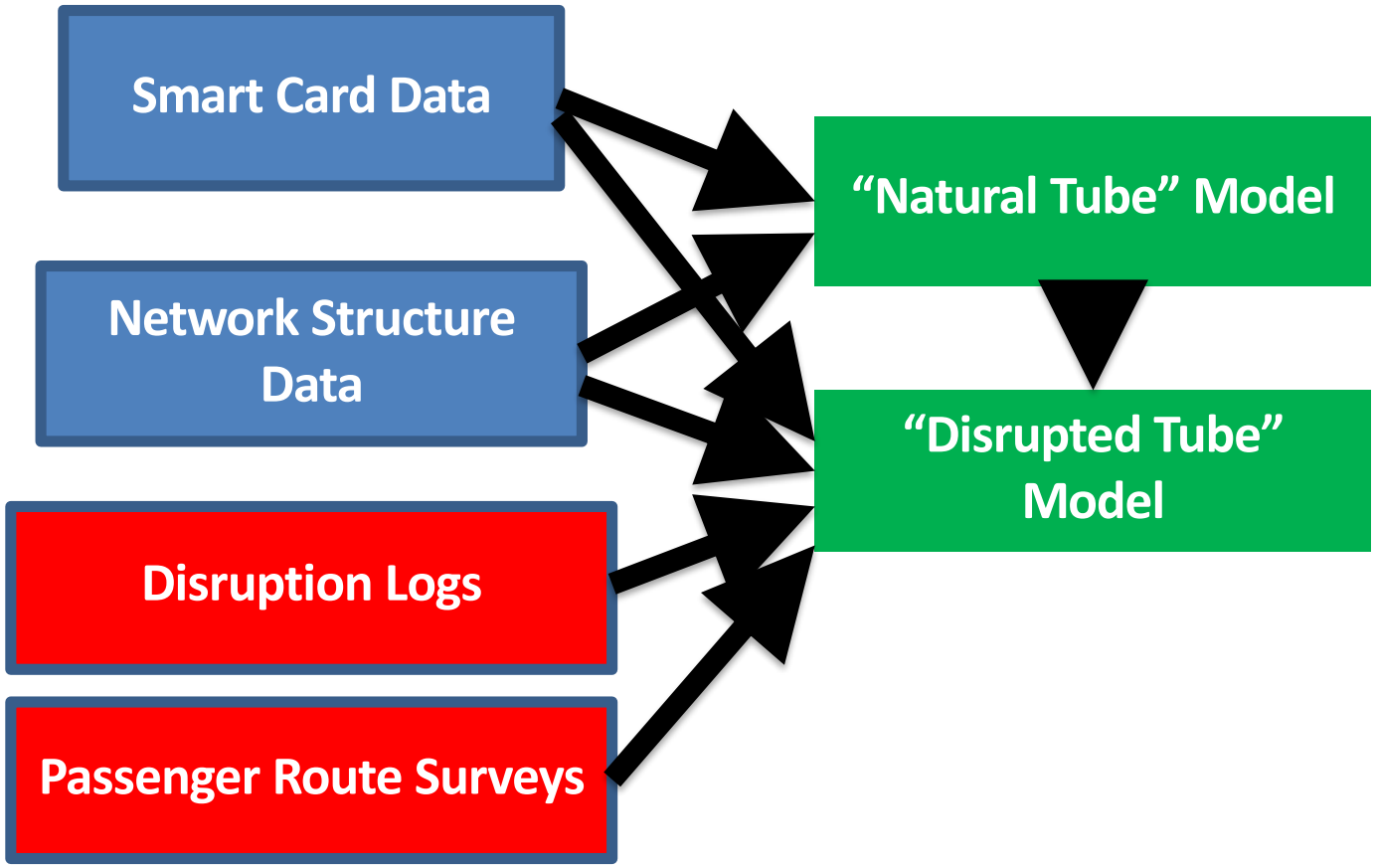
Task 1: Approach

1. Build a model for origin-destination counts for all 374^2 pairs of stations (includes Overground and DLR) and every minute of the day in the **natural regime** (that is, no disruptions)

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 factual behaviour under disruption using a linear model

Task 1: From Heterogeneous Data to Models



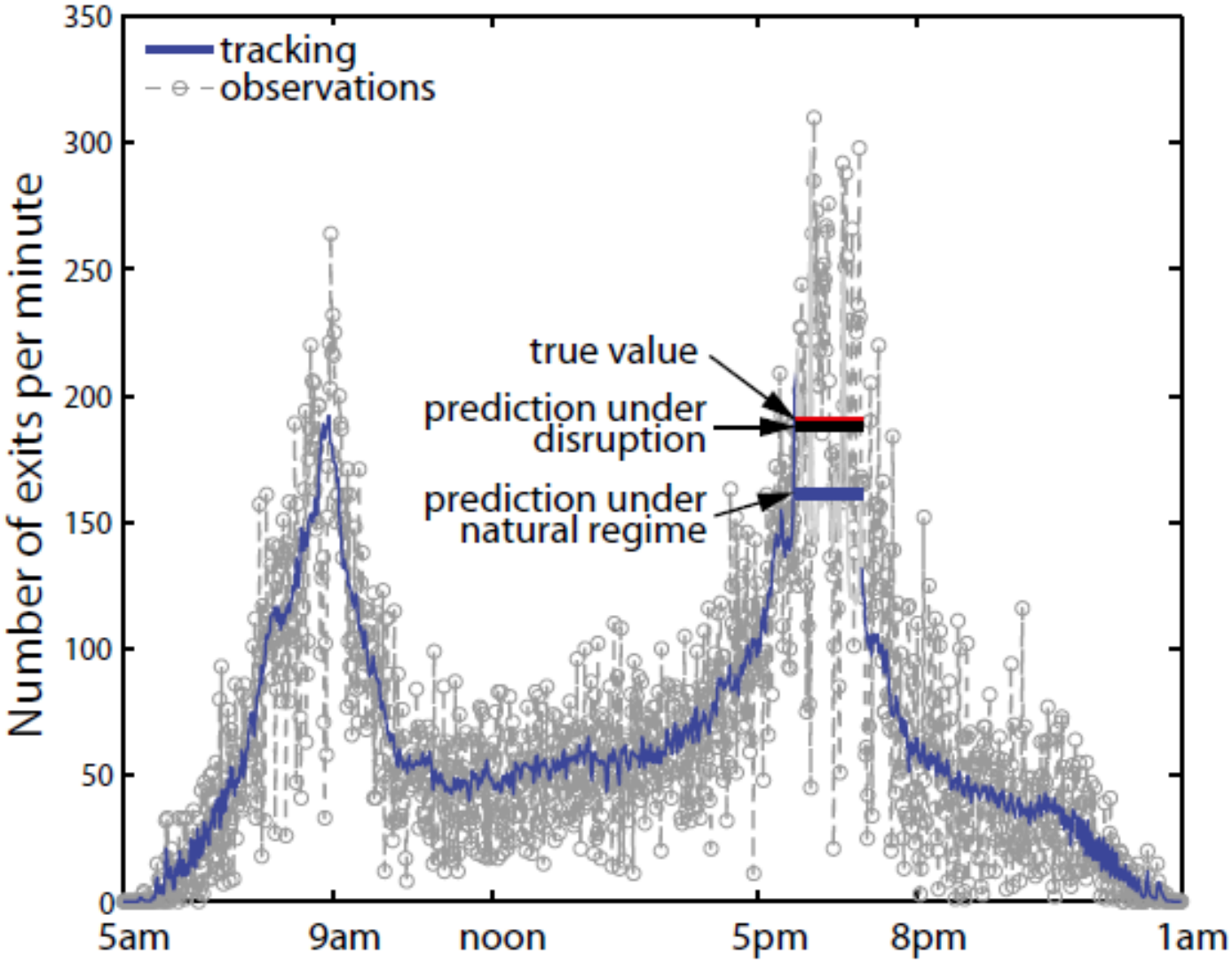
Oyster Card Data

- Anonymized IDs
- History of **taps**:
  - Event (IN/OUT, among others)
  - Location
  - Time of the day (1 minute resolution) and date
- 70 days between February 2011 and February 2012
- Some measurement errors
  - e.g., it is possible to leave the Arsenal station without tapping out
- Change of stations within connections are not usually recorded

Main Idea in a Picture

- Victoria to Brixton closes down unexpectedly around 5:30-7:30
- More people leave than usual (why?)
- In the next picture, I'll show
  - Blue: average predicted behaviour if no shock had happened
  - Red: actual observed behaviour
  - Black: estimated behaviour under shock

Victoria Station



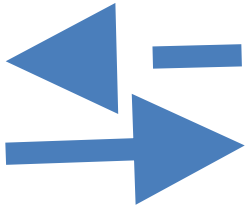
Fundamental Modelling Idea

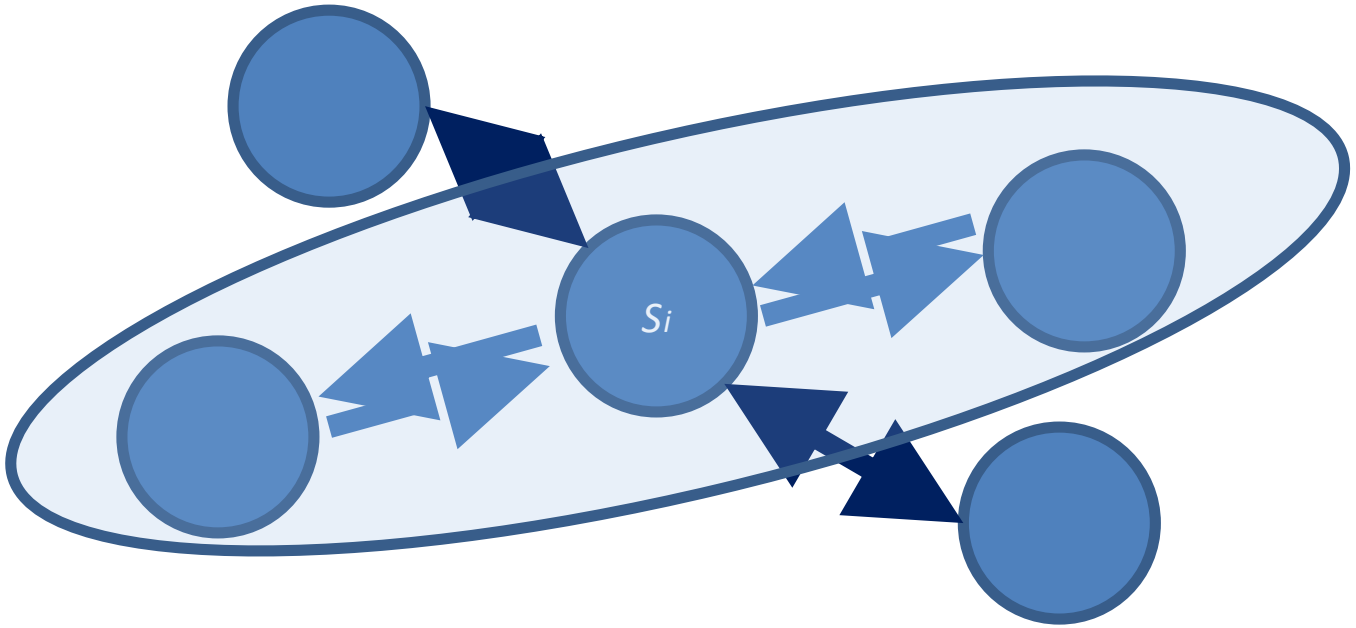
EXPECTED EXITS UNDER DISRUPTION(i) =

EXPECTED NATURAL NUMBER OF EXITS(i)

– MISSING INFLOW(i)

+ MISSING OUTFLOW(i)

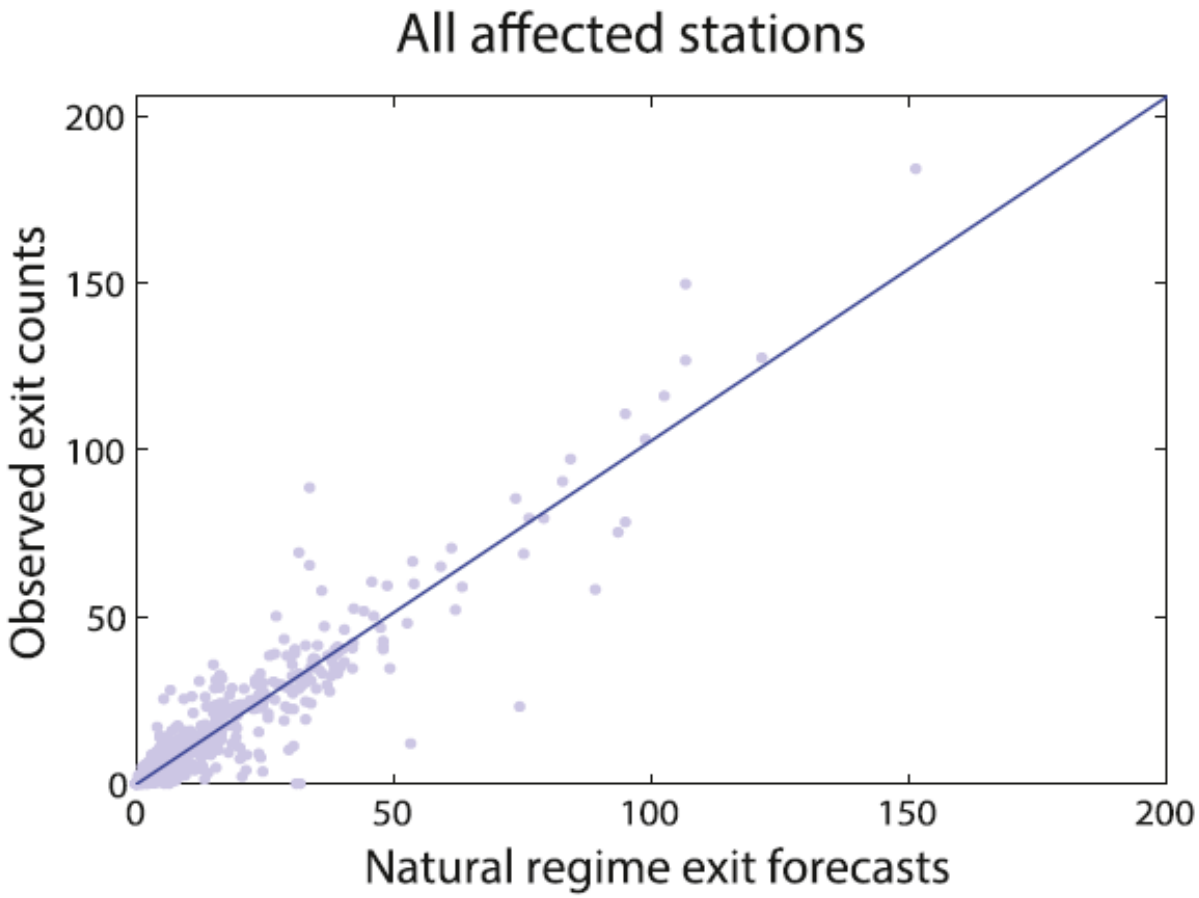




Disrupted

Fundamental Modelling Idea

- First check:

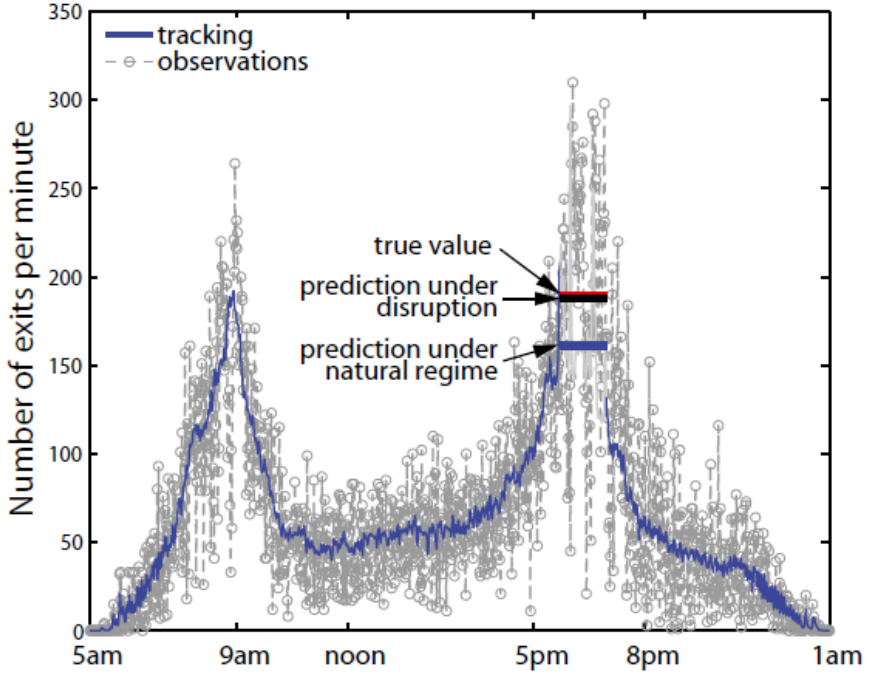


Regression Models

- Given shock event S happening on time t_1, t_2, …, t_F and for each station affected, define **outcome variable**

$$\overline{N}^{\mathcal{S}[j]}_{t_1:t_F} \equiv \sum_{t=t_1}^{t=t_F} N^{\mathcal{S}}_{jt} \Big/ F$$

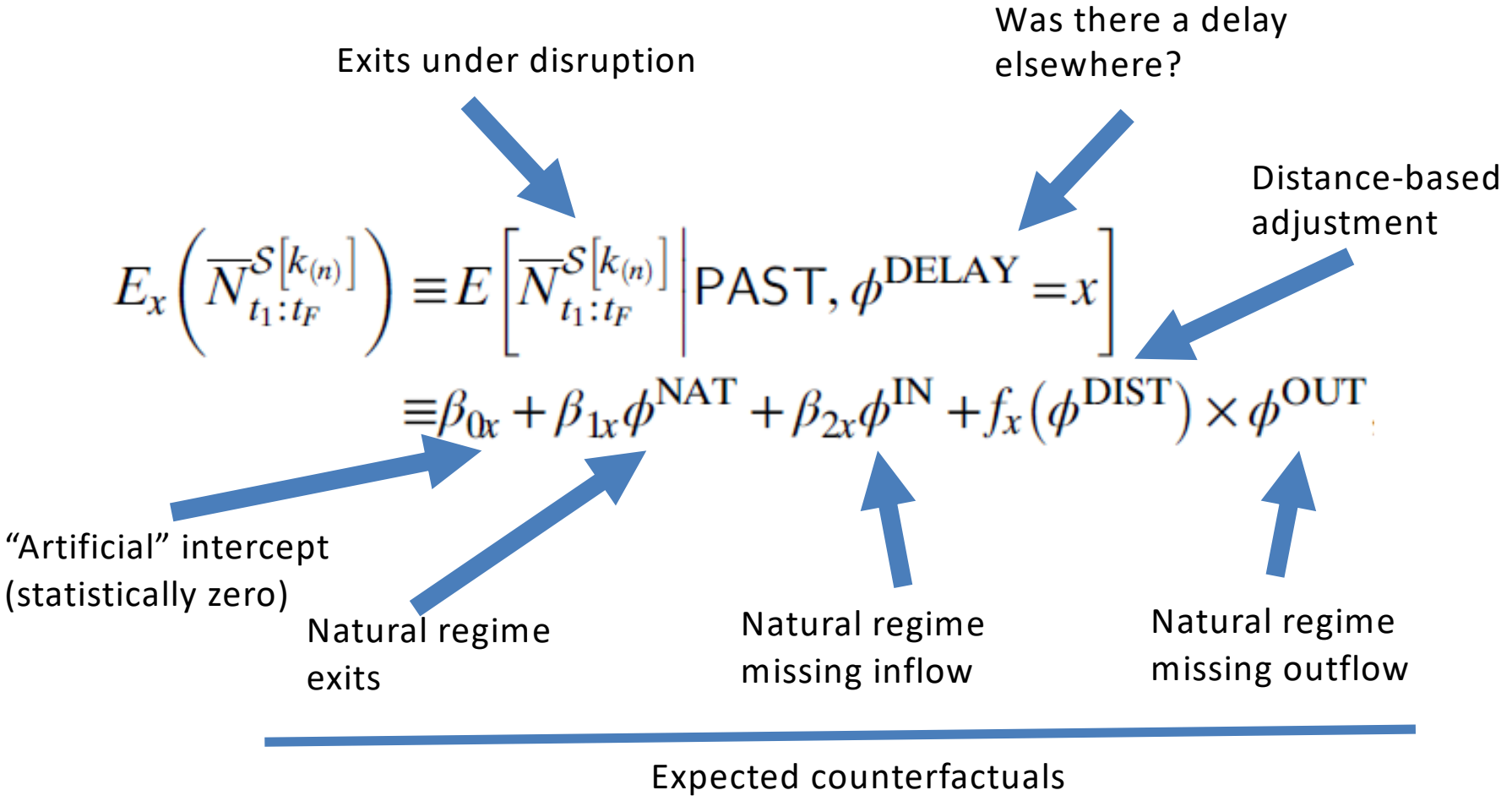
- Notice: assumption of **constant knowledge** for the passengers



Regression Models

- A model for the endpoints only, a model for all stations with a neighbour outside the line
- Covariates include counterfactual exit predictions from each origin, weighted by a “flow factor”
- Another covariate is whether the word “`delay`” appears in the textual description of the event

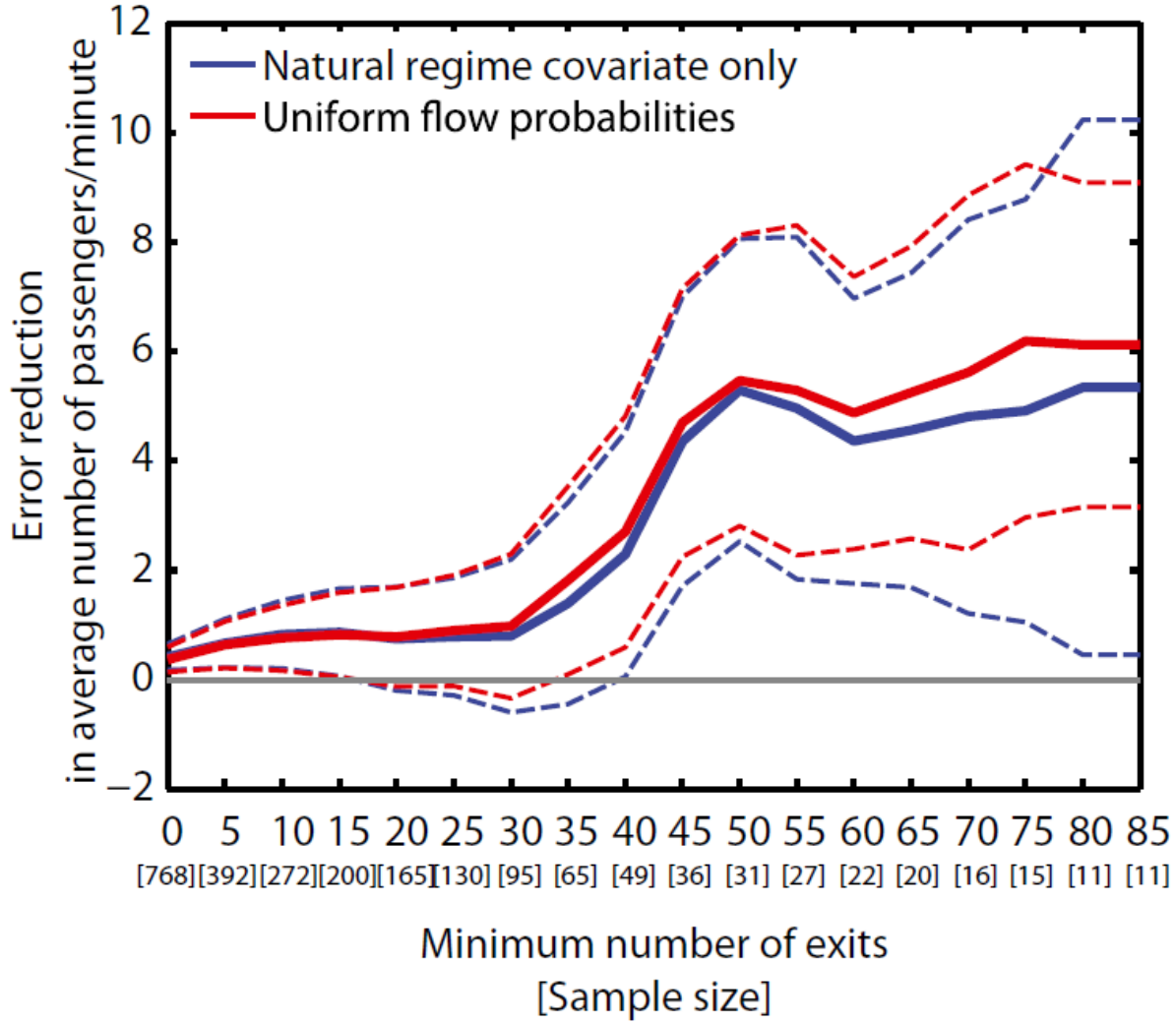
Regression Models



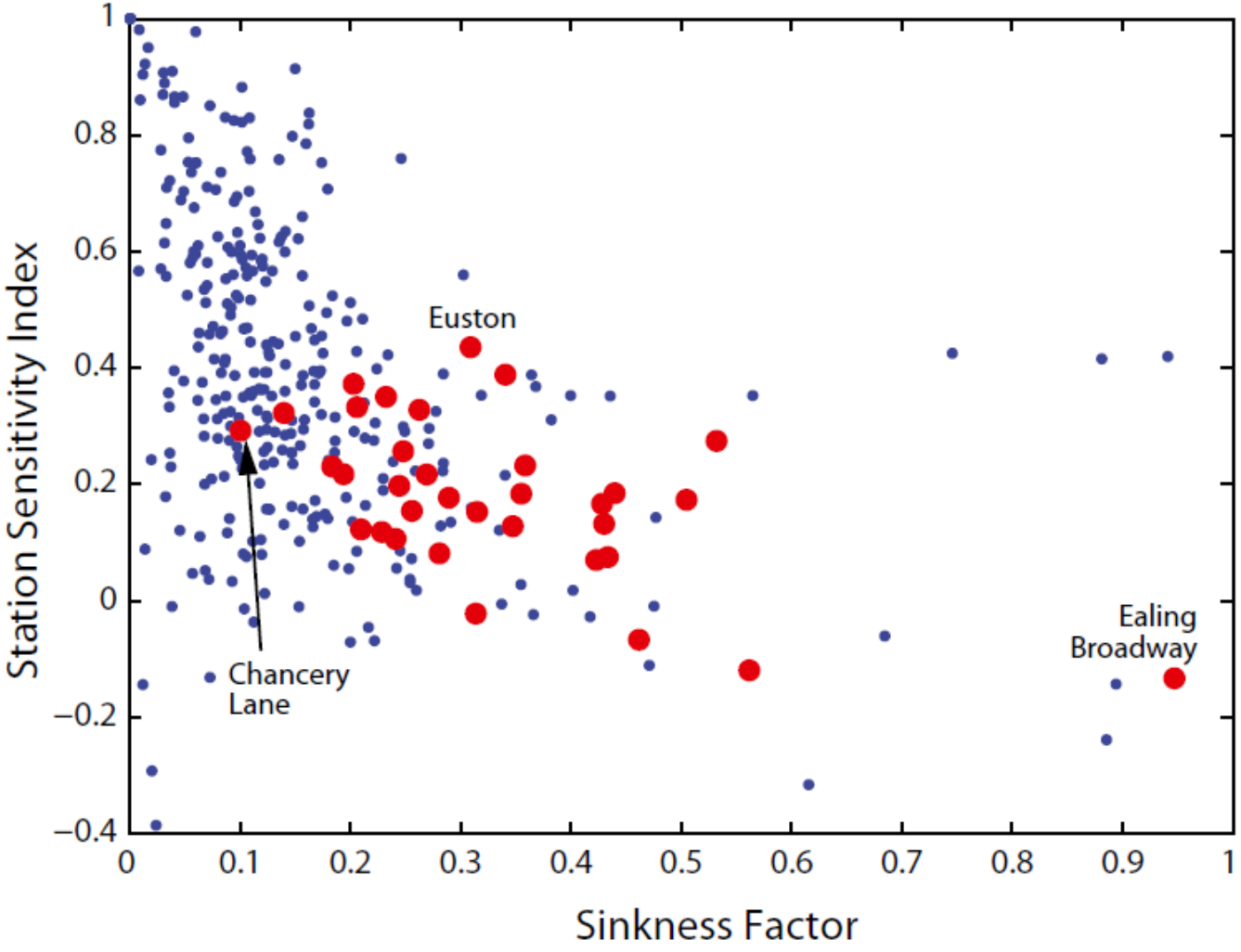
Disruption Data

- 793 data points, over our 70 (week) days
- Each data point corresponds to the outcome at a particular station, particular disruption
- One disruption → several stations of interest, several data points
- Regressions all fit by least-squares
- Define here "S_i" as the target node of a given data point

Predictions



Station Sensitivity Index vs. Sinkness Factor



Task 2: Context

- Network tomography:

  “Tomography is imaging by sections or sectioning, through the use of any kind of penetrating wave." (Wikipedia)

  - Medicine: reconstruct the interior of a body from external signals
  - Network science: detect network properties from partial endpoint/link observations
- Tomography of the Underground: passenger routes and implied **loads** in the links

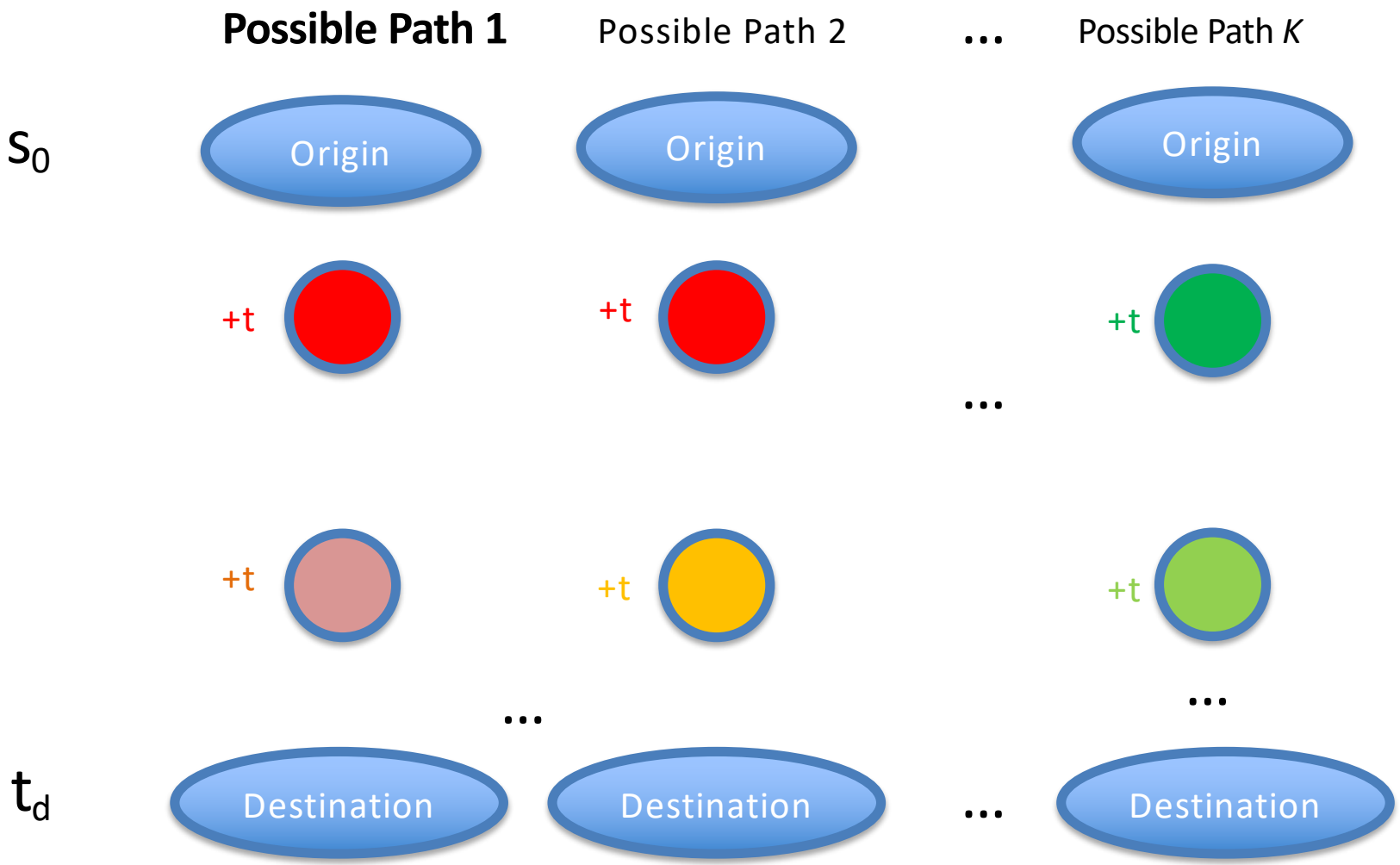
Network Tomography: Dual Formulations

- Two different but related problems:
  - Local-to-global: reconstruct OD distributions from link observations
  - Global-to-local: reconstruct link features from OD observations
- For i.i.d. data, a method for one problem can be easily converted to solve the other.
- Traffic within a day is however is highly non-stationary and methods need to take that into account.

Data and Model

- Structure of the data: quadruplets, containing origin station, destination station, time of entrance, time of exit.
- For scalability, we model it as a **latent mixture regression** problem:
  - Covariates: origin, destination, time of entrance
  - Output: time of exit
  - Latent variables: route taken, time spent at each position

Main Modelling Idea



Network Delay Model

- Sum over candidate paths with a given probability, each link *k* traversed in (unobserved) r_k minutes.

$$t_{k+1} = t_k + r_{k+1} \qquad r_{k+1} \sim \text{Poisson}(a(k, \gamma, s_o + t_k))$$

The probability of completing the journey in t_d time is

$$p(t_d|s_o) = \sum_r \sum_{\gamma \in \Gamma} p(\gamma) \delta\left(\sum_{i=2}^{\ell} r_i - t_d\right) \prod_{j=1}^{\ell(\gamma)-1} \text{poisson}\left(r_{j+1}; a\left(j, \gamma, s_o + \sum_{k=2}^{j} r_k\right)\right)$$

Origin information

Path probability

Total delays add up to t_d

Mean delay function

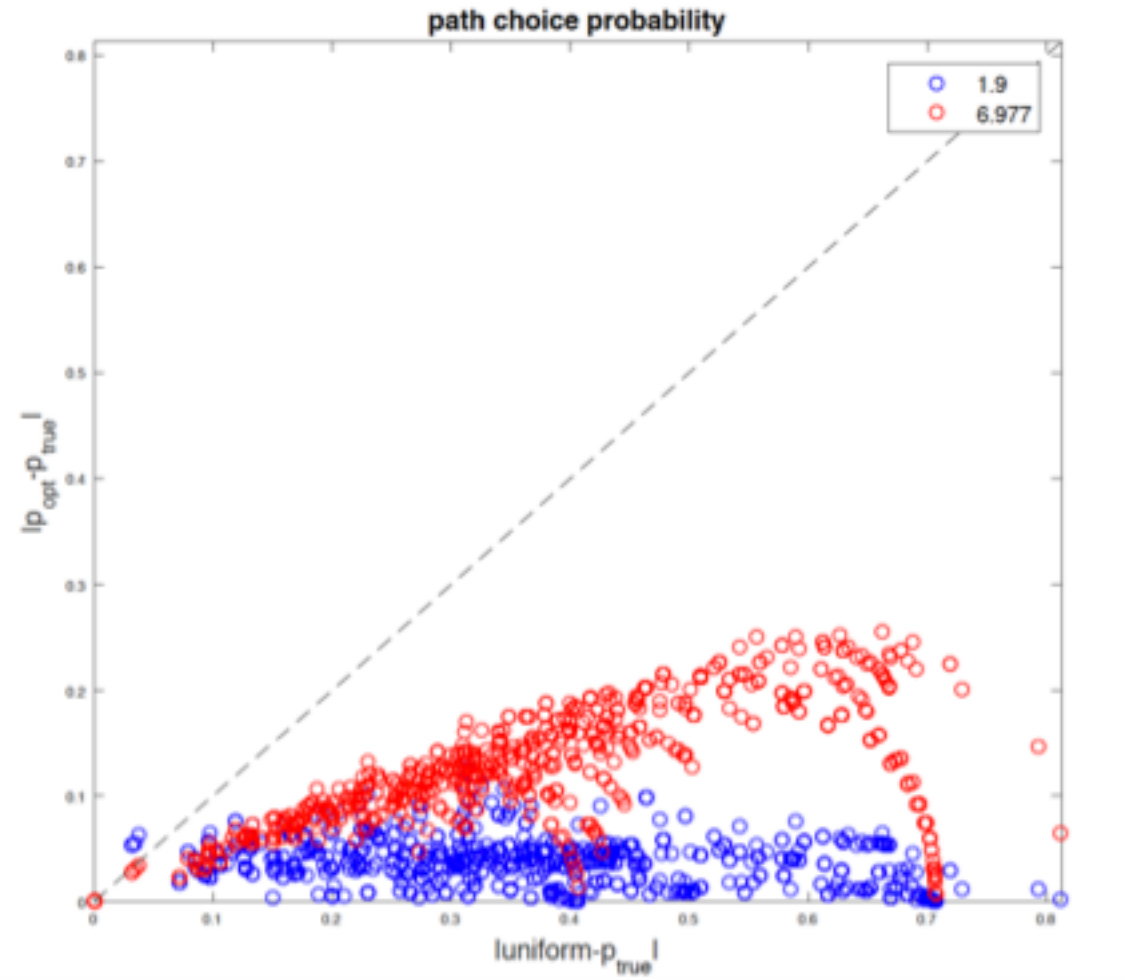
Computation

- Two approaches:
  - exact fit of an approximate model
  - approximate fit of the given “exact” model
- Stochastic gradient descent with uniform sampling
- Computation of gradient requires dynamic programming
  - Neural nets people call that “backpropagation.”

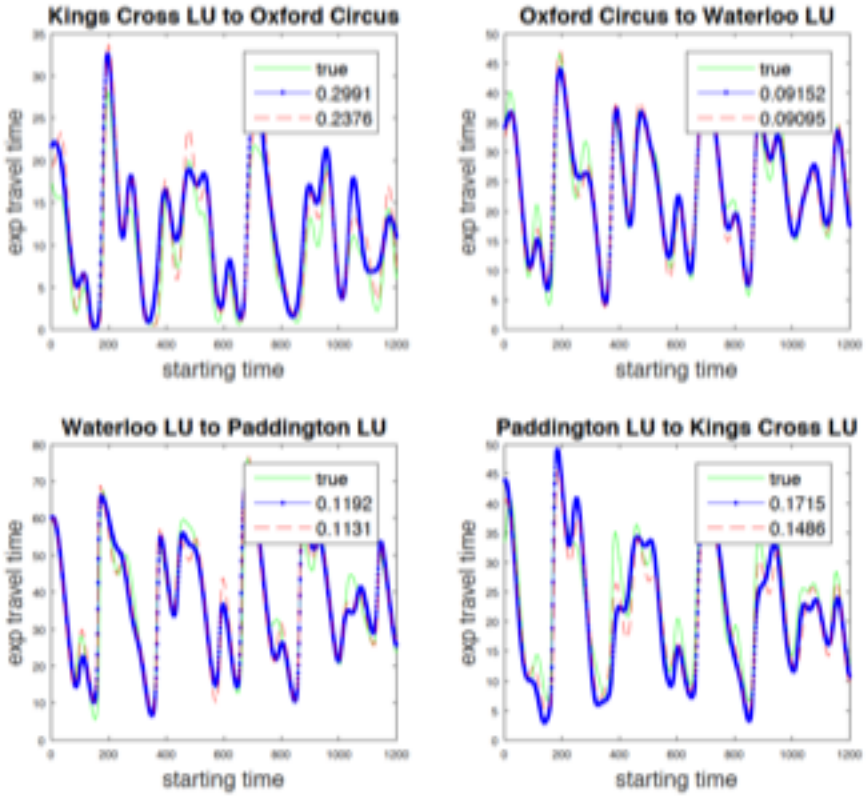
London Underground Experiments

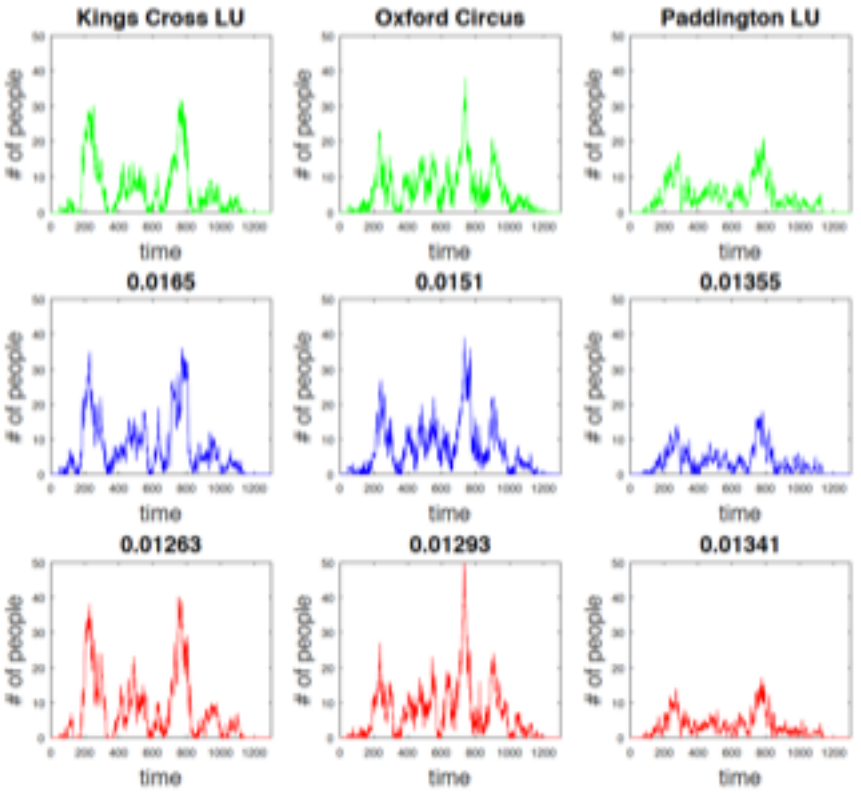
- Synthetic experiments (ground truth known) + anonymised Oyster data from a day in October 2013.
  - All 131 stations of zones 1 and 2
  - Surrogate data for loads: (actual) train schedule with respective weights on some parts of the Underground

Route Choice (Synthetic Data)

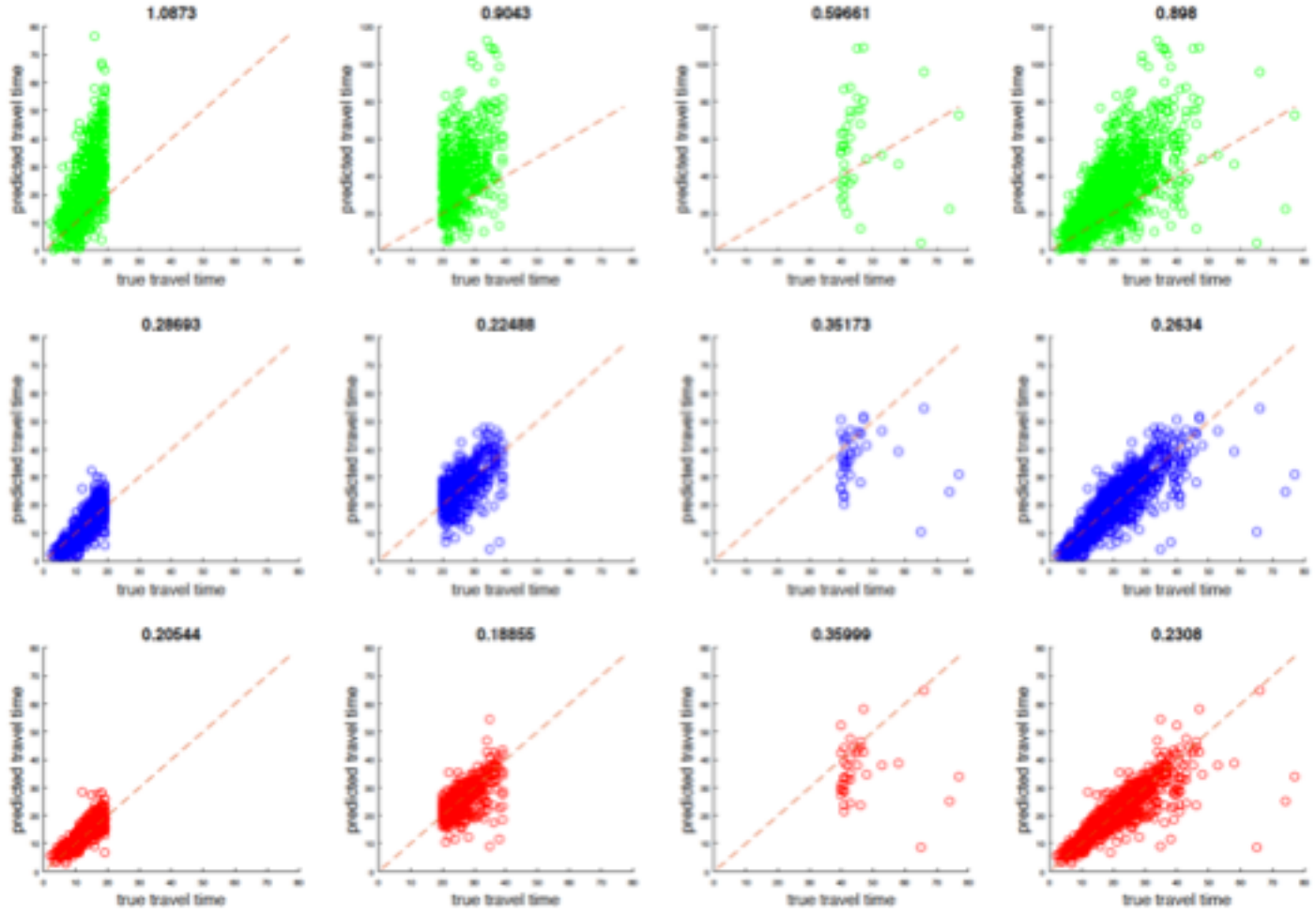


Expected Link Time (Synthetic Data)

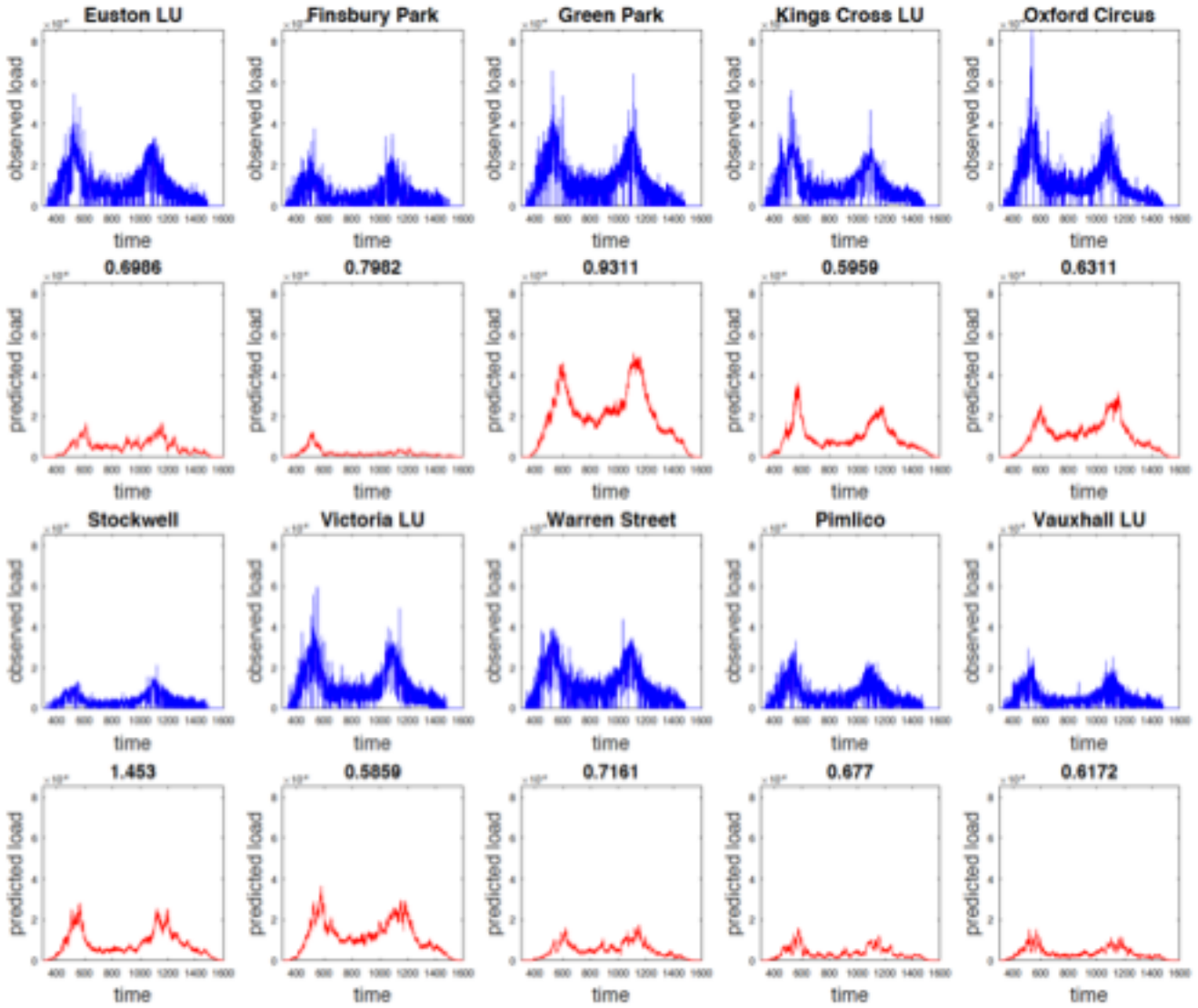




Predicted Travel Time (Real Data)



Loads (Surrogate Real Data)



Task 3 (On-going Work)

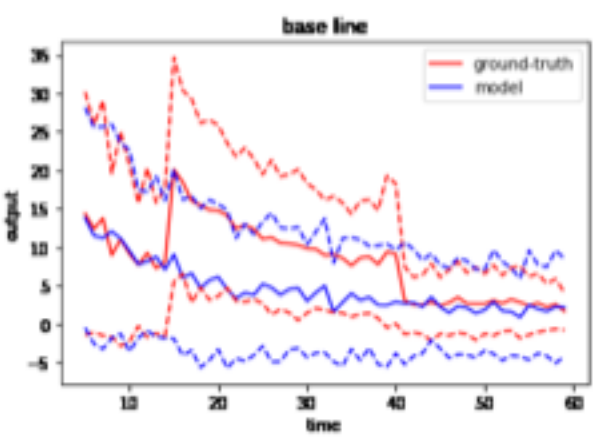
- Direction 1: refining the passenger behavior model at an individual level
  - At the moment of a disruption, assess
    - Where a given passenger is
      - Borrows from tomography model
    - Where it would go had no disruption taken place
      - Use of some modified classifiers
    - How long he/she would take to get there
      - Borrows from tomography model

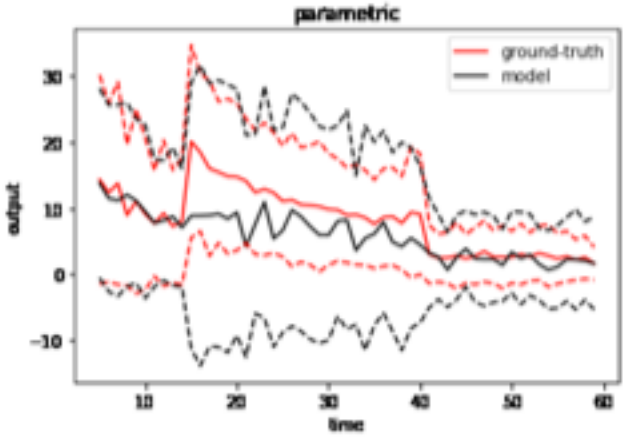
Task 3 (On-going Work)

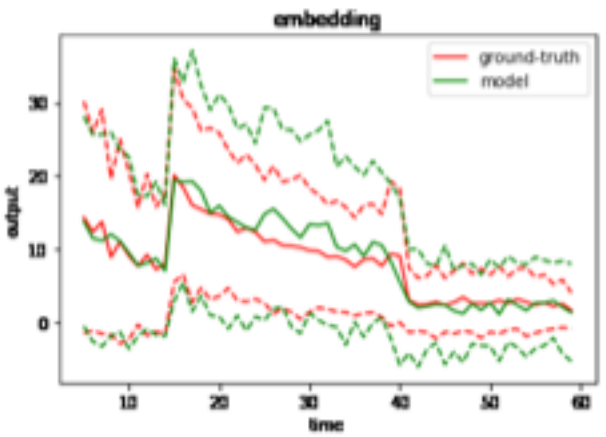
- Direction 1: refining the passenger behavior model at an individual level
  - Use data under disruption to get feedback for
    - How much interference was there?
      - Change of destination/length of travel time as a function of where they were at the beginning
    - How much this varies by ”type” of passenger?
      - Latent types based on commuter vs non-commuter, and how “fast” they typically are compared to other passengers

Task 3 (On-going Work)

- Direction 2: getting probabilistic estimates of aggregated behaviour







Conclusion

- Transport modelling as probabilistic predictive modelling.
- Many opportunities as a sandbox for large scale algorithms and causal impact assessments.
- Many opportunities for real-world impact.

References

- “Predicting traffic volumes and estimating the effects of shocks in massive transportation systems”. Ricardo Silva, Soong M Kang and Edoardo Airoldi. *PNAS, 2015*

- “Tomography of the London Underground: a Scalable Model for Origin-Destination Data”. Nicolo Colombo, Ricardo Silva, Soong M Kang. *NIPS 2017.*

- *“*Counterfactual distribution regression for structured inference”. Draft, available upon request.

The Alan Turing Institute

Lloyd's Register Foundation

Thank You

This work has been partially funded by the EPSRC grant EP/N020723/1 and the Lloyds Register Foundation.

Thanks to Transport for London for data access and feedback.