Extended Abstract:

Using smart card data to analyse the disruption impacts on urban metro systems

Nan Zhang, Daniel J. Graham, Jose M. Carbo

Railway and Transport Strategy Centre, Department of Civil Engineering, Imperial College London, United Kingdom

1. Background and objective

Incidents occur regularly on urban metro systems. It is common to encounter a disruption, especially for those systems operated for over a century, such as London Underground and New York Subway. Disruptions happen for a variety of reasons that may be internal or external to the metro system, including signal failure, strike, lack of rolling stock, blockage on the track, engineering work, adverse weather conditions, natural disasters, emergencies like terrorist attacks, etc. (Mattsson et al., 2015).

The occurrence of disruptions is likely to cause delays and disorder in the punctuality and regularity of the metro operation (Melo et al., 2011), which hence reduces passengers' satisfaction and service level. This problem can be further magnified during peak hours, when commuting travel demand significantly increases. Moreover, there will be safety problems from crowding passengers stranded on the platform. Therefore, it is important to understand the mechanism of disruption occurrence and analyse their impacts on metro systems. The research results will help metro operators prepare better recovery plans for urgent service interruptions.

This study aims to use smart card data to analyse the impacts of disruptions on London Underground, including two objectives:

- We estimate the causal effects of metro disruptions on travel demand and average journey time, using Propensity Score Matching (PSM) methods. The propensity score is calculated by the probability model of disruption occurrence in metro stations, allowing for temporal correlation among disruptions.
- Second-stage regression models are built to find determinants of estimated disruption impacts. We show how our probability model can be applied to prevent future disruptions, and how to identify vulnerable stations.

2. Literature review

In an attempt to understand the incidence of disruptions, Melo et al. (2011) build a regression model for incident numbers across urban metro lines. Their result reveals that passenger demand is a major factor of disruption occurrence. In other studies, passenger behaviour, driver performance, system design and severe weather events are also found to be relevant for metro disruptions (Rjabovs & Palacin, 2017; Wan et al., 2015; Brazil et al., 2017; Gonzva et al., 2017).

Traditional studies on metro disruption impacts are generally based on survey data. Pnevmatikou & Karlaftis (2011) estimate the difference of metro users' willingness to pay for alternative modes before and after the partial closure of Athens Metro, using revealed preference questionnaires. Rubin et al. (2005) conduct a follow-up survey of psychological and behavioural reactions to the bombings in London on July 2005. The result shows that the intention of traveling by London Underground is reduced although the system went back to normal services. Other studies have used incident delay models to predict maximum delay time, average delay time and corresponding punctual rates (Weng et al., 2015).

Despite the fact that survey data are useful, the availability of smart card data allows researchers to apply continuous and high quality datasets. Silva et al. (2015) provide a framework to estimate the number of exiting passengers on disrupted metro stations, based on Oyster card data under normal conditions. But without modelling real disruption observations, this approach is subjectively built to approximate changing demand. Sun et al. (2016) use smart card data for Beijing subway to evaluate the travel time and delay for affected passengers. They apply a shortest path assumption to assign travel flow, which need further investigation to identify how well it is able to predict actual flows.

To the best of our knowledge, the proposed probability model is the first one that could integrate the aforementioned factors and predict disruption occurrence, also accounting for temporal correlations among disruptions. The estimated impacts will be more accurate because we eliminate the confounding biases.

3. Data and methodology

3.1 Data description

The research is based on data provided by Transport for London, including Oyster card data, Cupid incident data and train movement data. The smart card data contain information such as passengers' boarding and alighting stations, transaction time, fare amount, card type (e.g. student, young person or adult) and ticket type (e.g. pay as you go credit or travel card). The incident logs and train movement records can be mined to reveal the location and duration of disruptions. To avoid potential influences from seasonal travellers and public holidays, the study period is selected during 28/10/2013 to 13/12/2013. All datasets cover the same period.

3.2 Methodology

To address the biases caused by confounding factors due to non-random disruption occurrence, we use propensity score based estimators to measure the changes of ridership and average journey time. The validity of PSM method relies on conditional independence assumption and overlap assumption (Imbens & Rubin, 2015).

3.2.1 The propensity score model

Our first task is to build the propensity score model to capture the probability of experiencing a disruption for each observation unit. In the London Underground system, there are 270 stations, with 42,000 minutes service time per station. We use each minute of a specific underground station as one observation unit, thus in total 42,000 * 270 = 11,340,000 observations are constructed in a panel structure.

We consider two probability models: a logistic regression model, in which the log odds ratio is modelled in linear form, and a Generalised Additive Model (GAM), in which the log odds ratio is modelled by smoothing splines. The dependent variable W_i is defined as the binary response of disruption occurrence. The dependent variables are chosen from all possible confounding factors X_i such as time of day, weather conditions, history ridership and station design characteristics. To account for temporal correlations among disruption occurrence, the past number of disruptions will also be added into the model.

The logistic regression model is:

$$pr\ (W_i = 1|X_i = x) = p(i)$$
$$\text{logit}[p(i)] = \ln[\ p(i)\ /\ (1-p(i))] = \alpha + \boldsymbol{\beta'x} + \beta_t \cdot temporal(i), \qquad i = 1, \dots, n \tag{1}$$

where α is the intercept, $\boldsymbol{\beta'}$ is the vector of regression coefficients related to the confounding covariates $\boldsymbol{x}$. (i) denotes the number of disruptions happened before the observation i, β_t represents the associated regression coefficient. The coefficients will be estimated by maximum likelihood principle.

The GAM is represented as:

$$pr\ (W_i = 1|X_i = x) = p(i)$$
$$\mathrm{logit}[p(i)] = \ln[\,p(i)/(1-p(i))] = \alpha + \boldsymbol{\beta}'\mathbf{f}(\boldsymbol{x}) + \beta_t \cdot \mathrm{f}_t(temporal(i)), \qquad i = 1, \ldots, n \tag{2}$$

where **f()** denotes the smoothing splines for independent variables, and $\mathrm{f}_t()$ denotes the smoothing spline for the temporal term.

After comparing the fitting results, we select propensity scores from the model with better predicting performance.

3.2.2 Matching algorithms

After obtaining the propensity score, we can use this index to match the appropriate comparison units for each treated unit. Two types of matching algorithms are used and compared:

- **Subclassification Matching:** The aim of subclassification is to form subsets, in which the distributions of all covariates are as similar as possible for treated and control groups.
- **Nearest Neighbour Matching:** This algorithm selects the K best control matches for each treated individual. We perform it with replacement in our analysis, with different values of K being tested.

3.2.3 Estimating the disruption effects

The average treatment effect for the disrupted stations can be calculated using

$$\tau_{ATT} = E(Y|W = 1, \hat{e}(\boldsymbol{X})) - E(Y|W = 0, \hat{e}(\boldsymbol{X})) \tag{3}$$

where W denotes the disruption status, $\hat{e}(\boldsymbol{X})$ denotes the estimated propensity score with given covariates $\boldsymbol{X}$. This formula represents the difference between average outcomes of matched treatment and control groups (Imbens & Rubin, 2015).

3.2.4 Second-stage regression models for identifying determinants of disruption impacts

We use linear regression models to explain the observed variation in station-level disruption effects.

$$\mathrm{Y} = \beta_0 + \boldsymbol{\beta_1} * \boldsymbol{X} + \boldsymbol{\beta_z} * \mathbf{Z} + \varepsilon \tag{4}$$

Where Y denotes the average disruption effects on each affected station, $\boldsymbol{X}$ denotes aforementioned confounding factors and **Z** demotes other relevant covariates including demographic, socio-economic, land-use and transport characteristics around metro stations. ε is the error term.

4 Current result and discussion

PSM estimation has been conducted on travel demands at both aggregate and disaggregate levels. The aggregate results indicate that there has been a small increase of exit ridership (2.3 passengers/minute) during metro disruptions. For disaggregate outcomes, the change of travel demand varies from station to station, with larger disruption impacts being spatially clustered in inner London. Via second-stage regression models, we find that stations with more surrounding bus stops, being outdoor and with larger average ridership will experience greater demand changes. Inversely, stations with more metro lines will lead passengers to transfer directly rather than leaving the system. We will also estimate corresponding disruption impacts on average journey time.

The above results have been applied to predict the disruption probability map and identify vulnerable stations to disturbance. We find that busy stations in central London, such as Oxford Circus, Victoria, Bond Street, London Bridge and Warren Street, are the most vulnerable stations in the system.

5 References

Brazil, W., White, A., Nogal, M., Caulfield, B., et al. (2017) Weather and Rail Delays: Analysis of Metropolitan Rail in Dublin William Brazil. *Journal of Transport Geography*. 59, 69–76.

Gonzva, M., Barroca, B., Gautier, P.É. & Diab, Y. (2017) Modeling disruptions causing domino effects in urban guided transport systems faced by flood hazards. *Natural Hazards*. [Online] 86 (1), 183–201. Available from: doi:10.1007/s11069-016-2680-7.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Mattsson, L. & Jenelius, E. (2015) Vulnerability and resilience of transport systems – A discussion of recent research. *Transportation Research Part A: Policy and Practice*. 81, 16–34.

Melo, P.C., Harris, N.G., Graham, D.J., Anderson, R.J., et al. (2011) Determinants of Delay Incident Occurrence in Urban Metros. *Transportation Research Record: Journal of the Transportation Research Board*. [Online] 2216, 10–18. Available from: doi:10.3141/2216-02.

Pnevmatikou, A., & Karlaftis, M. (2011). Demand changes from metro line closures. In European Transport Conference 2011.

Rjabovs, A. & Palacin, R. (2017) The influence of system design-related factors on the safety performance of metro drivers. *Proc IMechE Part F: J Rail and Rapid Transit*. [Online] 231 (3), 317–328. Available from: doi:10.1177/0954409716630007.

Rubin, G.J., Brewin, C.R., Greenberg, N., Simpson, J., et al. (2005) Psychological and behavioural reactions to the bombings in London on 7 July 2005: cross sectional survey of a representative sample of Londoners. *Bmj*. [Online] 331 (7515), 606. Available from: doi:10.1136/bmj.38583.728484.3A.

Silva, R., Kang, S.M. & Airoldi, E.M. (2015) Predicting traffic volumes and estimating the effects of shocks in massive transportation systems. *Proceedings of the National Academy of Sciences*. [Online] 112 (18), 5643–5648. Available from: doi:10.1073/pnas.1412908112.

Sun, H., Wu, J., Wu, L., Yan, X., et al. (2016) Estimating the influence of common disruptions on urban rail transit networks. *Transportation Research Part A: Policy and Practice*. [Online] 94, 62–75. Available from: doi:10.1016/j.tra.2016.09.006.

Weng, J., Zheng, Y., Qu, X. & Yan, X. (2015) Development of a maximum likelihood regression tree-based model for predicting subway incident delay. *Transportation Research Part C*. [Online] 57, 30–41. Available from: doi:10.1016/j.trc.2015.06.003.