**Benchmarking public transport network performance using statistical and machine learning models on large-scale high-dimensional data**

**Quantifying the determinants of public transport journey time performance using statistical and machine learning models with large-scale and high-dimensional data (A,C)**

Dr Daniel J. Graham - Dr Nan Zhang

Abstract:

In an era where black-box models are increasingly accurate in their predictions, the desire to master and understand the relationships between factors and the dependent variable is driving the need for the development of Causal Inference, which is gaining momentum. Urban mass transit systems typically generate large volumes of data on various aspects of operations. Statistical analyses can be used to summarise and present such data, drawn from within and between systems, to understand the drivers of performance. This project will leverage large-scale publicly available data from the General Transit Feed Specification (GTFS) to characterise the performance of transit systems at a sub-system level, focusing on train on-time performance and the inference of disruption occurrence on delays.

The outputs of the GTFS analysis will then be fused with data from the Community of Metros (COMET) longitudinal database to develop a high-dimensional explainable machine learning model of the determinants of the performance domains characterised via GTFS.

Data:

* General Transit Feed Specification (GTFS)
* COMET longitudinal data (TSC, Imperial College London)

Methods:

* Coding to scrape data from GTFS (Washington DC) (+cleaning)
* Large-scale data processing, summary, and dimensionality reduction methods
* Explainable machine learning algorithms
* Regression modelling for parameter estimation

Bringing something to the table = demonstrate their is a gap | results to interpret and eventually conclusion (policy)

No universally applied industry standard for punctuality measurement exists

1. Not a scale free graph because of the scale (ok for airport but not a metro)
2. Reduce from 2 dimension network analysis to 1 D line analysis
3. Instead of going too much into details (granular analysis) of an individual-level travel behaviour, provide a tool for network management so they know what to do and the reach/extend of their action.
4. There is still room for deploying that method to total journey time or even improve delay propagation using epidemiological approaches.
5. Room for improvement:
   * + - * removing ‘contamination’ from exogenous determinants [1]
         * engineering interventions often tend to be non-binary [17]





Literature Review

e.g.

Nip, K. H., Gardner, L., Davies, C. M. and Elghazouli, A. Y. (2010). Extremely low cycle fatigue tests on structural carbon steel and stainless steel. Journal of Constructional Steel Research. 66(1), 96-110.

Wadee, M. A. and Gardner, L. (2012). Cellular buckling from mode interaction in I-beams under uniform bending. Proceedings of the Royal Society A. 468(2137), 245-268.

**References**

1. J Graham, Ramandeep Singh, Model-based adjustment for conditional benchmarking, IMA Journal of Management Mathematics, Volume 33, Issue 3, July 2022, Pages 381–393, <https://doi.org/10.1093/imaman/dpab021>

*develop a simple model-based adjustment approach for comparative benchmarking by removing ‘contamination’ from exogenous determinants via regression analysis and then recovering adjusted performance metrics with reduced variance.*

*The quantitative methods to evaluate and compare performance are those concerned with efficiency frontier estimation, including stochastic frontier analysis (SFA)(parametric) and data envelopment analysis (DEA)(non-parametric). Our objective is to empirically adjust random variable Y so that it captures variance in intrinsic performance but not in exogenous influences.*

*Dependence between Pi and Xi no longer possible to obtain consistent estimates of β methods for causal inference in the presence of unobserved confounders.*

1. ~~Prateek Bansal, Daniel Hörcher, Daniel J. Graham, A Dynamic Choice Model to Estimate the User Cost of Crowding with Large-Scale Transit Data,~~ *~~Journal of the Royal Statistical Society Series A: Statistics in Society~~*~~, Volume 185, Issue 2, April 2022, Pages 615–639,~~ [~~https://doi.org/10.1111/rssa.12804~~](https://doi.org/10.1111/rssa.12804)

*~~unlike revealed preference data, Stated preference data offer complete information about the sociodemographic characteristics of decision-makers and ensures enough variation in the attributes of the alternatives faced by decision-makers.~~*

*~~Crowding cost estimate is the main input for assignment models that predict the distribution of travel flows.~~*

*~~SP surveys and estimate the rider’s perceived value of crowding in terms of a crowding multiplier. They elicit preferences of riders in a hypothetical binary route choice experiment. Assuming that riders choose a utility-maximising route after trading-off crowding density, travel time and cost, these studies estimate discrete choice models (DCMs) to obtain the crowding multiplier.~~*

*~~two research gaps: (1) day-to-day variation in route preferences and learning behaviour of riders are hard to capture in SP experiments and (2) a daily commuter, might not actively make a compensatory (utility-maximising) route choice. In fact, a rider can adhere to the same route until a bad experience occurs.~~*

*~~literature in psychology suggests using cumulative prospect theory (CPT for (1)) and instance-based learning theory (IBLT for (2)).~~*

*~~Thus, the off-the-shelf DCMs cannot be used to estimate riders’ valuation of transit crowding. In this study, we propose a dynamic latent class model (DLCM) which integrates IBLT-based expectation formation framework to account for learning behaviour.~~*

1. ~~Graham, D. J. (2021). Causal Inference for Ex Post Evaluation of Transport Interventions, Chapter in Encyclopaedia of Transportation.~~

*~~The objective of ex post evaluation is to quantify retrospectively the impacts that interventions (called treatment) have had on defined outcomes of interest.~~*

*~~Transport interventions are typically non-randomly assigned. This implies that locations that receive transport interventions will tend to be in some sense different from those that do not. If this is the case, confounding is said to be present and this makes it hard to separate the effect of the treatment from other influences through simple comparisons based on treatment status alone.~~*

*~~Broadly speaking there are two way of doing this. First, through model-based adjustment, in which differences between units in characteristics X, are measured and included within a model to obtain marginal causal effects. Second, by developing models which exploit sources of exogenous variance to obtain causal estimates without explicit representation of X in the model.~~*

***~~Ex Post Evaluation via Model-Based Adjustment for Confounding~~*** *~~(strong ignorability assumption)~~*

*~~To proceed we need to estimate the relevant conditional expectations in the equations above. There are three approaches that are commonly used in the literature to do this: outcome regression (OR), propensity score (PS) models, and mixed or Doubly Robust (DR) models.~~*

***~~Ex Post Evaluation Under a Non-ignorable Treatment Assignment~~***

*~~In practice, this is often untenable, either because there are insufficient measured covariates to establish conditional independent, or because other sources of endogeneity (e.g., reverse causality or measurement error) are at play inhibiting a causal interpretation of the data.~~*

*~~There are a number of popular estimators that are used in this setting to obtain causal estimates of the ATE. Some use additional variables (instruments) to extract exogenous variation in treatments, while others exploit quasi-experimental conditions for identification. Here we briefly review three of the most commonly used approaches in economic evaluation: instrumental variables (IV)(1/Find a set of instruments which are exogenous to the outcome but highly correlated with the treatment. 2/Use the instruments to enforce orthogonality between the error term and an instrument transformed design matrix), difference-in-differences (DID), and regression discontinuity designs (RDD).~~*

1. Graham, D.J. (2023) Causal inference for transportation research. Working Paper.

*According to this distinction associational inference attempts to infer the cause of some defined effect on an outcome of interest, while causal inference views the observed outcome as one potential realisation achieved through manipulation of an effect, e.g. through intervention.*

*Association implies statistical dependence between variables X and Y , but does not establish causation. In that sense, it is a necessary, but not sufficient, condition for causation. Conversely, causation does imply association as a necessary condition. Clearly, causation is a stronger claim than association, and consequently is more difficult to establish (or identify) empirically.*

*The three assumptions defined above, which are together referred to by Rosenbaum and Rubin (1983b) as strong ignorability, permit identification of APOs and ATEs for non randomly assigned treatments. This is stated formally in the following theorem:*

*Theorem 2. (Identification of the APO and ATE under strong ignorability). Under strong ignorability APOs and ATEs for non randomly assigned treatments can be identified by conditioning on X and integrating over the covariate distributions to capture the marginal causal effect.*

*Outcome Regression (OR) leaves fD|X (d|x) and fX (x) unspecified, and explicitly models the conditional expectation function E[Yi|Di,Xi]. Propensity Scores (PS), which measure the probability of assignment to treatment given the set of observed pre-treatment covariates.*

*The validity of the estimation methods discussed in the previous section requires us to maintain that strong ignorability holds. In practice, this is often untenable, either because there are insufficient measured covariates to defend the CIA, or because other sources of endogeneity are at play. We will cover instrumental variables (IV), difference-in-differences (DID)(only for binary D), synthetic control (SC) and regression discontinuity design (RDD).*

*The logic of IV is that since failure of the CIA implies that we have an identification problem, we will seek to directly remove this correlation from the model by introducing other observable variable(s) known as instruments, which we will denote Z.*

*The RDD method exploits this discontinuity in treatment assignment to study the conditional distribution of the outcome either side of the threshold of the forcing variable. A discontinuity in outcome is interpreted as evidence of a causal effect of the treatment. Note that the key identifying assumption is there is a discontinuous change in the probability at threshold.*

*Another major challenge for causal inference in our field relates to the requirement that the SUTVA is met, which applies to all methods covered in the paper. A key implication of the SUTVA is that the outcome for each unit must be independent of the treatment status of other units, or in other words, there should be no ‘interference’ in treatment effects between units.*

5. Singh, R., Graham, D. J., & Anderson, R. J. (2019). Characterising Journey Time Performance on Urban Metro Systems under Varying Operating Conditions. Transportation Research Record, 2673(7), 516-528.

*In the literature, seven forms of distribution functions have been identified to define journey times (log normal, gamma, log logistic, Weibull, Burr, generalised Pareto, and normal distributions). Each of these is trialed in the analysis. Assuming that journey times are identically and independently distributed, functional parameters are estimated in a unimodal form via maximum likelihood estimation. The R statistical package ‘‘fitdistrplus’’ is used for distribution fitting.*

*To assess how well the distributions fit the data, goodness-of-fit hypothesis tests and indicators are used. Three types of test statistics are calculated: Kolmogorov–Smirnov, Cramer–von Mises, and the Anderson–Darling statistics.*

*According to the literature, a mix of congested and uncongested travel conditions is likely to be occurring within these multimodal distributions. Overall, the majority (67%) of the incident-affected distributions adhere to a unimodal form.*

*Figure 3a shows mean journey times normalised by the free-flow time. The value of the measure is greater than 1 for all lines. This indicates that mean journey times are longer than the free-flow time across the entire day in all cases.*

*Three performance measures are presented to assess the mean and variance of journey times. The first indicator measures the 15min mean normalised by the free- flow time. The second and third indicators provide measures of variance via the 15 min RBT normalised by the free-flow time, and the 15min standard deviation normalised by the free-flow time.*

1. ~~Wei Sun, Yang Quan Chen, A Simulation Study of Consensus Speed over Scale-Free Networks, IFAC Proceedings Volumes, Volume 42, Issue 22, 2009.~~ [~~https://doi.org/10.3182/20091006-3-US-4006.00013~~](https://doi.org/10.3182/20091006-3-US-4006.00013)

*~~Nowadays the scale-free network is regarded to be more robust and immune to the random mutation and perturbation than purely random network.~~*

*~~~~*

*~~The algebraic connectivity is the second smallest eigenvalue λ~~~~2~~ ~~of the graph Laplacian L(G) it (α(G) = λ~~~~2~~~~) is a measure of convergence speed (or performance) of the consensus algorithm the larger the λ~~~~2~~ ~~is, the sooner the consensus could be achieved. Let x(t) indicate the state of all the agents at time t, and L is the Laplacian of graph matrix.~~*

*~~~~*

*~~Another algorithm is the one in (Wang and Guo, 2008), which is a discrete time consensus algorithm. x~~~~i~~~~(k) is the state of the i~~~~th~~ ~~node n~~~~i~~ ~~on step k. Let N~~~~i~~ ~~denotes the set of neighbours of node n~~~~i~~~~, and assume d~~~~i~~ ~~indicates the degree of the node n~~~~i~~~~.~~*

*~~The node number N, the power-law parameters γ and dmin are the only parameters needed to determine a scale- free network.~~*

*~~As the algebraic connectivity can be seen as a measure of convergence speed, the investigation on λ2 should be done first. For robustness on time delay, λn should be studied.~~*

*~~~~*

*~~The first task of our investigation is the relationship between consensus time tc and γ, which is the most important parameter of a power-law distribution. The time cost tc to reach consensus grow dramatically as the γ increases. Meanwhile, the network reaches consensus more quickly when the minimum degree dmin is larger.~~*

1. Soodeh Hosseini, Mohammad Abdollahi Azgomi, A model for malware propagation in scale-free networks based on rumor spreading process, Computer Networks, Volume 108, 2016, Pages 97-107, ISSN 1389-1286, <https://doi.org/10.1016/j.comnet.2016.08.010>
2. Serafino M., Cimini G., Maritan A., Rinaldo, A., Suweis, S., Banavar, J. R., & Caldarelli, G. (2020, December 30). True scale-free networks hidden by finite size effects. Proceedings of the National Academy of Sciences, 118(2). <https://doi.org/10.1073/pnas.2013825118>

*We analyse about 200 naturally occurring networks with dis- tinct dynamical origins to formally test whether the commonly assumed hypothesis of an underlying scale-free structure is generally viable. This has recently been questioned on the basis of statistical testing of the validity of power law distributions of network degrees. Deviations from pure power law behaviour are permitted in the small degree regime. Our approach sorts out when we may reject the hypothesis that the inherent structure of networks is scale invariant.*

*The FSS ansatz has been developed precisely to infer the singular behaviour (i.e., the exponents determining the universality classes) of the physical properties of a system in the thermodynamic limit, having only information on the system properties at finite sizes.*

*For a network with a finite number of nodes, the degree distribution does not follow a pure power law but is modified by the function f (ref. 27 also has a discussion of finiteness in the context of growing network models).*

*To sum up, two independent statistical tests of the scale-free attributes of a network explained in FSS of Networks and Ratio of Moments Test are the quality of the collapse S (i.e., the reduced χ2 between data and master curve) and the compatibility of d and dE (measured through their Z score). We use these tests to define a classification for the degree distribution of empirical networks:*

* *SSF (strong scale free)if S≤1 and ZddE ≤1,*
* *WSF (weak scale free) if S ≤3 and ZddE ≤3,*
* *NSF (non scale free) otherwise (or when n*∗*<lnN for the original network or any of its subnetworks)*

*On the other hand, infrastructure networks (i.e., road and flights network) are rarely scale free (with the notable exception of air traffic control systems), possibly because of the heavy cost of establishing a connection.*

1. Singh, R., Graham, D. J., Horcher, D., & Anderson, R. J. (2020) Decomposing journey time variance on urban metro systems via semiparametric mixed methods, Transportation Research Part C: Emerging Technologies, 114, 140-163.

*In contrast, semiparametric regression methods are able to model flexible relationships generated from individual data points, resulting in parameter estimates with a potentially higher degree of fidelity to the data compared to the conventional linear models.*

*Multivariate regression methods are more widely applied to determine the causes of variance of total journey times.*

*The studies model static service supply factors including those related to the route, such as route length and the number and spacing of crossings, and those related to the rolling stock factors, namely rolling stock type. The dynamic service supply factors reflect the time-varying operating conditions, and include: headway, speed, the delay at the start of the route run, and train movements including meets, passes, and overtakes. Early formulations specify dwell time as a function of the number of passengers boarding, alighting, and on-board trains. In subsequent empirical work, studies identify additional factors beyond passenger demand that influence dwell times. Rolling stock-specific characteristics including the rolling stock model type, the number of train doors, door width, the distance between train doors, …*

*The coefficient of variation of headway (COV headway) is calculated over a 15 min period for each OD pair. This covariate is included to capture the variation in train frequencies which otherwise skew the model results, particularly at the beginning and end of service early and late in the day, and at the transition periods between peak and off-peak times.*

*The test result suggests that there is a degree of correlation between the egress times and headways, however, the value of the correlation coefficient is −0.05. The result shows that there is a very weak negative correlation, indicating that egress times tend to decrease as headways increase to a minimal degree.*

*A fixed effects structure is more appropriate if the variable has been drawn from a finite population, where inferences regarding the effect of the variable are confined to the categories of the variable included the model.*

*Non-parametric models (model form 1) perform best across the criteria that represent the degree to which the models capture variation in the data.*

*For the on-train time model, the covariates capturing headway, headway variance, and all passenger demand related covariates are insignificant at the lowest significance bound of 90%. Train speed and static route characteristics are the primary determinants of variance in train running times between stations.*

1. ~~Zhang N, Graham DJ, Carbo Martinez JM, Using smart card data to analyse the disruption impact on urban metro systems, Transportation Research Board 98th Annual Meeting~~

*~~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 average journey time.~~*

*~~The dependent variable W~~~~i~~ ~~is defined as the binary response of disruption occurrence.~~*

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

~~𝜏𝐴𝑇𝑇~~ *~~=~~* ~~𝐸~~*~~(~~*~~𝑌~~*~~|~~*~~𝑊~~ *~~= 1,~~* ~~𝑒̂~~*~~(~~*~~𝑋~~*~~)) −~~* ~~𝐸~~*~~(~~*~~𝑌~~*~~|~~*~~𝑊~~ *~~= 0,~~* ~~𝑒̂~~*~~(~~*~~𝑋~~*~~))~~*

*~~Where W denotes the disruption status, ê(X) denotes the estimated propensity score with given covariates X. This formula represents the difference between average outcomes of matched treatment and control groups. The above results have been applied to predict the disruption probability map and identify vulnerable stations to disturbance.~~*

1. ~~Singh, R., Graham, D. J., & Anderson, R. J. (2020) Quantifying the effects of passenger-level heterogeneity on transit journey times, Data-Centric Engineering, 1, 2632-6736.~~

*~~(A bit of a copy of « Singh, R., Graham, D. J., Horcher, D., & Anderson, R. J. (2020) Decomposing journey time variance on urban metro systems via semiparametric mixed methods, Transportation Research Part C: Emerging Technologies, 114, 140-163. »)~~*

1. Zhang N, Graham DJ, Hörcher D, et al., 2021, A causal inference approach to measure the vulnerability of urban metro systems, Transportation, Vol:48, ISSN:0049-4488, Pages:3269-3300

*Unbiased estimates of disruption impact are obtained by adopting a propensity score matching method which adjusts for the confounding biases caused by non-random occurrence of disruptions.*

*Incidents occur frequently in urban metro systems, mainly due to supply-side failures (e.g., signal failures), sudden increase in travel demand (e.g., public concert or football matches)*

*vulnerability a measure of susceptibility of the transport system to incidents. This paper proposes a novel alternative methodology to quantify vulnerability, by empirically estimating the causal impact of service disruptions on travel demand, average travel speed and passenger flow distribution at station-level.*

*There are two traditional methods used to build vulnerability indicators of metro systems— topology-based and system-performance-based analysis. The topological methods rely on complex network theory to convert the metro network.*

*From a methodological point of view, our empirical approach has three stages: first, we apply a causal inference method to estimate the impact of disruptions on station-level travel demand and travel speed . Then, we construct vulnerability metrics based on the disruption impact estimated in the first stage. Finally, the third stage imputes missing vulnerability metrics for non-disrupted stations using machine learning algorithms.*

***Propensity score matching (PSM) methods***

*The system-level impact, which averages the impact of all disruptions occurred within the metro system, is too generic to represent network vulnerability. Thus, we focus instead on estimating station-level disruption impacts. To predict probability of encountering disruptions (Prop. Score) at a metro station (on the baseline confounding characteristics) we use the logistic regression model with a linear link function.*

*In 2013, the London Underground (LU) had 270 stations and 11 lines, […] For connectivity among stations, LU has 56 stations connecting 2 lines, 16 stations connecting 3 lines and 8 stations connecting more than 4 lines. The role of propensity score models is to establish a comprehensive index to represent all confounding factors, rather than predicting treatment assignment.*

*The histograms display apparent overlap between the treatment and control groups, even for large propensity scores. There is no treated unit outside the range of common support, which means we do not need to discard any observations. We thus conclude that the overlap assumption is tenable in our empirical study.*

1. ~~Carbo, J. M. et al. (2019) ‘Evaluating the causal economic impacts of transport investments: evidence from the Madrid–Barcelona high speed rail corridor’, Journal of Applied Statistics, 46(9), pp. 1714–1723. doi: 10.1080/02664763.2018.1558188~~

*~~Good read about SCM and DID~~*

1. ~~Daniel J. Graham. Emma J. McCoy. David A. Stephens. "Approximate Bayesian Inference for Doubly Robust Estimation." Bayesian Anal. 11 (1) 47 - 69, March 2016.~~ [~~https://doi.org/10.1214/14-BA928~~](https://doi.org/10.1214/14-BA928)

*~~Good read about DR~~*

*~~OR model is correctly specified for E[Yi|Xi,Di]~~*

1. Yu Luo, Daniel J. Graham, Emma J. McCoy, Semiparametric Bayesian doubly robust causal estimation, Journal of Statistical Planning and Inference, Volume 225, 2023, Pages 171-187, ISSN 0378-3758, <https://doi.org/10.1016/j.jspi.2022.12.005>

*Frequentist semiparametric theory has been used extensively to develop doubly robust (DR) causal estimation. DR estimation combines outcome regression (OR) and propensity score (PS) models in such a way that correct specification of just one of two models is enough to obtain consistent parameter estimation. An equivalent Bayesian solution, however, is not straightforward as there is no obvious distributional framework to the joint OR and PS model, and the DR approach involves a semiparametric estimating equation framework without a fully specified likelihood.*

*Bayesian approach offers natural solutions of quantifying uncertainty with multiple component models, predicting complex causal quantities in terms of probabilistic statements and incorporating prior information when expert knowledge becomes available. However, such approaches utilise flexible and non-parametric modelling to avoid the misspecification in the OR, which is not the primary focus of this paper.*

1. K. Adjetey-Bahun, B. Birregah, E. Châtelet, J-L. Planchet, A model to quantify the resilience of mass railway transportation systems, Reliability Engineering & System Safety, Volume 153, 2016, Pages 1-14, ISSN 0951-8320, <https://doi.org/10.1016/j.ress.2016.03.015>

*The concept of resilience has been introduced to measure not only the system's ability to absorb perturbations, but also its ability to rapidly recover from perturbations. In this work, we propose a simulation-based model for quantifying resilience in mass railway transportation systems by quantifying passenger delay.*

*Bruneau et al. where the authors studied the resilience of hospitals in a region in the aftermath of an earthquake. They define resilience as the ability of a system to reduce (1) its failure probabilities; (2) the consequences from failures, in terms of casualties, damage, and negative economic and social consequences; (3) its recovery time.*

*By using component importance measures, they classify a system's components according to the sensitivity of the system to their failures.*

*The lack of a system of systems approach to quantify the resilience of transportation system. The system of systems approach and components' interdependencies when studying the resilience of complex systems, are not taken into account sufficiently in the literature.*

*Delay : *

1. Daniel J. Graham, Causal inference for data centric engineering

*Causal inference aims to quantify effects that occur due to explicit intervention (or ‘treatment’) in non- experimental settings, typically for non-randomly assigned treatments*

*In contrast to exiting reviews, emphasis is placed on models and methods for multivalued and continuous treatments, since engineering interventions often tend to be non-binary.*

*We provide in-depth coverage of models and methods for multivalued and continuous treatments because engineering interventions often tend to be non-binary, being characterised by length, volume, number, capacity and the like.*

A graph of a graph with blue dots

Description automatically generated with medium confidence

Arrows pointing arrows on a black background

Description automatically generated

A diagram of a process flow

Description automatically generated

A math equations on a white background

Description automatically generatedA diagram of a flowchart

Description automatically generated