Evaluation of the causal effect of disruption on delay as a benchmarking tool for public transport network performance

Guillaume Amann

Department of Civil and Environmental Engineering
Imperial College London

Keywords: causal inference; benchmarking; synthetic control; disruption; network.

Abstract

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. While black-box models are becoming increasingly accurate for prediction problems, the desire to master and understand the relationships between factors and the dependent variables is driving the need for the development of Causal Inference, which is gaining momentum. This empirical analysis leverages large-scale publicly available data from the General Transit Feed Specification (GTFS) feed of the Washington DC metro network. The Synthetic Control approach proves to be a relevant method to generate a typical synthetic day of operation on the Orange line, which can be used as a benchmark to assess the performance and resilience of the transit system under different scenarios. Recommendations include further research on the constitution of a spatial control unit and future key performance indicators.

1. Introduction

The growth of significant transport infrastructure projects, such as the Greater Paris, the Crossrail in London, and the East Side Access in New York, has highlighted the increasing complexity and size of urban rapid transit networks. These expansions make it more challenging to operate and benchmark public transport network performances in addition to have brought to light the need for a better understanding of these network architecture (Derrible & Kennedy, 2021) (Adjetey-Bahun, *et al.*, 2016), as disruptions can have a significant yet different impacts on delay. This is particularly important in the context of increasing investments in sustainable mass transit systems and the growing awareness of the risks associated with unreasonable and inefficient performance measures. These are the reasons why it is crucial to comprehend both the networks structures and the causal effect of disruptions on delay in order to develop relevant strategies to improve their resilience and maintain reliable and efficient service, even under disruptions (Lu, 2018). In the context of urban rapid transit systems, this is synonym with minimising delay and ensuring that the system can quickly recover from disruptions.

The General Transit Feed Specification (GTFS) enables access to large scale high dimensional data on which many statistical and machine learning techniques can be applied in order to highlight such confounding effects (Graham, 2021). However, the mass quantification of resilience is a double-edged sword that enables a more accurate assessment of the impact of disruptions on delay (Zhang, *et al.*, 2021) while necessity careful design before integration to prove relevancy and scalability. Despite a significant need for tools that aid network management to assess their performance, there is still not universally accepted standard or benchmark for measuring performance in the transit industry. Instead of putting the emphasis on a granular analysis of individuals travel behaviours and on an estimate of the passenger generalised journey time, this study proposes to shift the focus and to equip managers with actionable insights on the scope and impact of their decisions.

This paper proposes to apply Synthetic Control, a proven method in economics (Abadie & Cattaneo, 2018), which helps generating a synthetic benchmark to which can be compared the outcome of a specific treated unit. This can be used to facilitate and scale up the quantification of system resilience from the operator point of view, by addressing the causal effects of service disruptions on delay in arrival time. Indeed, interventions, like disruptions, are non-randomly events occurring on the transport network. Neither their timing nor their location is free from confounders amongst the independent variables. This confounding effect can be addressed to obtain clear and accurate measurements of the true causal inference of interventions or disruptions, therefore laying the ground for future benchmarking tools to be utilised. This study highlights the capability of the Synthetic Control method in the context of transport operation to generate a relevant control group (other service days of the system that did not experience the same treatment/disruption) to which can be compared the causal effect of a disruption during the day on subsequent delays.

The remainder of this paper is organized as follows. Section 2 reviews literatures on transport network vulnerability analysis and on the different *ex-post* evaluation methods for causal inference. The methodology introducing Synthetic Control method is proposed in Section 3. Section 4 illustrates the formatting of Schedule GTFS data and the gathering of real-time information. Section 5 presents the results of an application of the methodology in Washington DC Orange line of the metro network with conclusions and limitations provided in Section 6, alongside some recommendation for future work.

2. Literature review

Incidents occur regularly on urban metro systems and the availability of GTFS data allows operators to track journey time reliability. Semiparametric regression modelling is one way to understand the underlying causes of journey time variance in urban metro systems (Singh, et al., 2020). But this mix of train location and passenger trip data decomposes total journey times. Rather than focusing on a detailed analysis of individual travel behaviour or estimating the generalized journey time for passengers, this study aims to provide managers with actionable insights into the scope and impact of their decisions. The occurrence of disruptions is likely to cause delays and disorder in the punctuality and regularity of the metro operation. The concept of resilience appeared as a useful and applicable approach to measure system's ability to both absorb perturbations (Lu, 2018) and recover from disruptions when applied in a scale free network scenario on Paris data (Adjetey-Bahun, et al., 2016). Following the same line of thought, an analysis of the London metro under the assumption of scale free network highlights "sources of structural and functional vulnerabilities" (Derrible & Kennedy, 2021). Although the traditional method used to build vulnerability indicators of metro systems is topology-based, meaning the public transport network is considered as scale free networks, it has been pointed out that this assumption is not sustainable (Serafino, et al., 2020). Because treatments and disruptions are typically non-randomly assigned on the transport network, a confounding effect needs to be addressed to obtain unambiguous measurements of the true causal inference.

Multiple evaluation methods help reckoning with this confounding effect. First, the Outcome Regression (OR) model estimates the effect of an intervention by modelling the outcome directly as a function of both treatment and covariates. This helps understanding how changes in treatment impact the outcome, providing a clear picture of the relationship between the treatment and the outcome. On the other hand, the Propensity Score (PS) model estimates the likelihood of receiving a treatment given a set of covariates, and then uses this probability to adjust for treatment selection biases. It helps in creating comparable groups for more accurate treatment effect estimation (Zhang, *et al.*, 2021). The Doubly Robust (DR) model is a combination of the previous two. It leverages the strengths of both approaches by correcting for selection biases through the propensity score and adjusting for residual confounding using outcome regression, ensuring more reliable and valid estimates (Graham, *et al.*, 2016).

Unfortunately, the validity of all the above-mentioned methods requires maintaining strong ignorability hypothesis. This supposes that, conditional on a set of observed covariates X, the treatment D assignment is independent of the potential outcomes Y. This ensures that any differences in outcomes between treatment and control groups can be attributed to the treatment itself rather than to other confounding factors. In a real-world scenario, this assumption cannot hold due to either a lack of covariates or the presence of other sources of endogeneity (Graham, 2021).

The two main alternative methods that do not require the strong ignorability assumption to hold are: Instrumental Variable (IV) and Difference-in-Differences (DID) methods. The Instrumental Variable is an approach which requires to find or to set a variable that only affects the treatment assignment influencing the outcome. IV's validity relies on two key assumptions: the instrument must be strongly correlated with the treatment and must affect the outcome only through the treatment. However, it is crucial that these two key assumptions of exogeneity and relevance are met, and in practice valid instruments are hard to find. The Difference-in-Differences approach compares the outcomes of a treated unit or group to a control (non-affected) group before and after treatment. By comparing the changes in outcomes over time between a treatment group and a control group, DID can control for confounding factors that are constant over time. This method is effective in dealing with time-invariant unobserved heterogeneity. However, DID relies on the assumption that, in the absence of treatment, the difference between the treated and control groups would have remained constant over time, which is arguably not the case when it comes to the relation between disruption and delay. Despite their robustness, these methods fall short under complex scenarios where confounding dynamics evolve over time or where suitable instruments are hard to find.

For the stable unit treatment value assumption, the Synthetic Control (SC) method proposes to estimate the causal effects of policies and interventions in the absence of randomized trials (Abadie & Cattaneo, 2018). This is where the SC becomes advantageous as it extends the DID approach by constructing a weighted combination of control units to create a synthetic control group that best approximates the characteristics and pre-treatment trends of the treated unit to improve comparability of the treatment and control groups (Graham, 2021). Unlike the above-mentioned methods that estimate the Average Treatment Effect (ATE), which involves calculating the potential outcomes for each individual under both treatment and control conditions, the Synthetic Control method focuses on estimating the treatment effect for a single treated unit. This allows for a more accurate comparison between the treated unit and a synthetic control group that is constructed to mirror the pre-treatment characteristics and trends of the treated unit.

The SC method identifies not only a single ATE estimate, but also reveals how the ATE effect evolves over time post-treatment, making it a robust basis for integrating future resilience benchmarking tools. This is achieved by comparing the observed outcome of the treated unit with an artificial twin scenario built from a weighted average of the synthetic control group's outcome. Furthermore, Synthetic Control does not need access to post-treatment outcomes during the study's design phase meaning one can determine synthetic control weights and make all design decisions without knowing their impact on the study's conclusions, making the study fairer and the conclusions are more reliable. For all the above-mentioned reasons, the Synthetic Control approach is a good candidate to methodologically generate relevant benchmark scenarios under real-word constraints for urban rapid transit systems.

3. Methodology

Causal inference is the process of drawing conclusions about cause-and-effect relationships between variables in observational data. In research, causal inference is often used to estimate the treatment effect of an intervention, treatment, or disruption, on an outcome of interest. To assess these effects, researchers commonly use different frameworks: the *ex-post* evaluation framework to assess the impact of an intervention or policy using data collected after its implementation, or the potential outcome framework which allows researchers to estimate the Average Treatment Effect (ATE) and individual future treatment effects. Both frameworks involve comparing the outcomes of a treated group (those who received the intervention or policy) to a control group (those who did not receive the intervention or policy) using statistical methods, as in Equation 1:

$$ATE = \mathbb{E}[Y(D = 1)] - \mathbb{E}[Y(D = 0)] \quad (1)$$

where: $\mathbb{E}[Y(D = 1)]$ is the expected outcome if everyone in the population were treated and $\mathbb{E}[Y(D = 0)]$is the expected outcome if everyone in the population were not treated.

Both frameworks rely on three assumptions: the assumption of independence, the assumption of no unmeasured confounders, and the assumption of stable unit treatment value (SUTVA). The first one, the assumption of independence, assumes that any individual outcome are independent of the treatment assignment for others. This assumption actually allows the estimation of the ATE by comparing the average outcomes of the treated group to the average outcomes of the control group. The assumption of no unmeasured confounders assumes that there are no variables that affect both the treatment assignment and the outcome that are not included in the analysis. The last assumption, the stable unit treatment value assumption (SUTVA), extend the first assumption and requires that each unit has a single outcome under each treatment condition, this addresses the potential issue of multiple versions of the treatment, which the independence assumption does not cover.

For the Synthetic Control method to provide valid estimates of causal effects, it relies on two simple assumptions. The first one is that the synthetic control unit serves as a reasonable approximation of what the treated unit's outcomes would have been in the absence of treatment. The second one, known as the parallel trends assumption, posits that the treated and control units would have followed the same pattern in the absence of the intervention. *Ex-post* evaluation is integral to this analysis, as it provides a framework for assessing the impact of the intervention using data collected after its implementation. The synthetic control method is particularly well-suited for *ex-post* evaluations in observational studies like in transport, where randomised experiments are not feasible. By closely matching the treated unit's characteristics to those of the synthetic control, it reduces bias from measured confounders, though it still assumes no unmeasured confounders exist. Moreover, this method is typically applied in settings where the independence and SUTVA assumptions are likely to hold, such as large-scale interventions affecting one unit without interfering with others. Therefore, the difference in outcomes between the treated unit and the synthetic control unit after the intervention provides a credible estimate of the ATE, making synthetic control a robust tool when these key assumptions are reasonably satisfied.

Since treatments and disruptions are usually assigned non-randomly on the transport network, it is possible to observe a confounding effect that must be addressed to obtain clear and accurate measurements of the true causal inference. Throughout this research, the terms 'treatment' and 'disruption' are used interchangeably, as the focus is on disruption, and these can be considered treatment from a mathematical perspective. Whether the treatment is observed, *ex-post* evaluation framework, or is yet to apply, potential outcome framework, the Synthetic Control method is one approach to causal inference that can be used to estimate or measure the treatment effect for a single treated unit, without the need for random assignment to treatment and control groups. By assuming SUTVA, consistency, and positivity, the Synthetic Control method allows for the estimation of the treatment effect for a single unit while controlling for unobserved confounders.

The SC method involves several key steps and formulas. Firstly, the control group is constructed by identifying a group of control units that are as similar as possible to the treated unit in their pre-treatment characteristics. This is often done using a combination of statistical matching techniques. The goal is to create a synthetic control unit that mirrors the pre-treatment trends and characteristics of the treated unit as closely as possible. The weights are then estimated to determine how much each control unit should contribute to the synthetic control unit. These weights are calculated using a linear optimization technique that minimizes the difference between the weighted average of the control units' pre-treatment outcomes and the treated unit's pre-treatment outcome. This optimisation problem is summarised in Equation (2).

$$\underset{W}{min}(X_1 - X_0W)^T V(X_1 - X_0W) \tag{2}$$

where W is the weights vector, X_1 is a vertical vector with a length equal to the number of observed characteristics of the treated group and X_0 is a matrix containing the characteristics for every element of the control groups. The solution to this problem is a vector of optimal weights, $W^*(V)$, with V a diagonal matrix that assigns weights to each characteristic.

Once the SC group has been derived, it is possible to compare its evolution with respect to the treated group during the whole period T (the number of years/months/days/*etc.* observed after the treatment following the treatment). This method aims at highlighting the gap obtained via the Equation (3) between the treated group and its synthetic twin:

$$Y_1 - Y_0 \cdot W^*(V) \tag{3}$$

where Y_1 is a vertical vector of T observed outcomes (here delay), and Y_0 is a matrix of the dependant variable of interest in the control groups for the post treatment years.

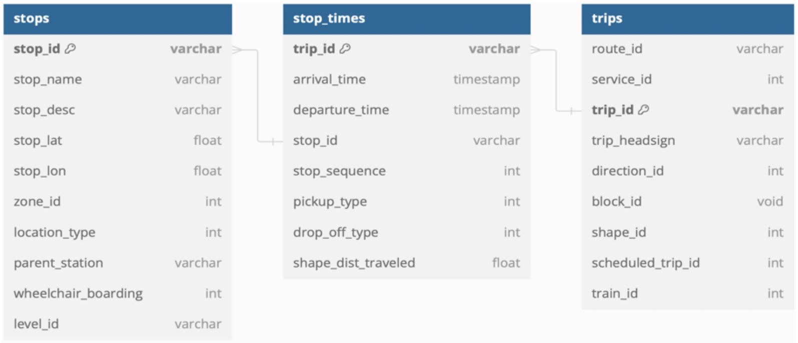
Because the Washington Metropolitan Area Transit Authority (WMATA) does not explicitly specify the station where the disruption occurs when emitting its alert messages in the GTFS feed, a spatial study of the propagation is not feasible. Hence, this paper applies SC on a treated day to compare it with a synthetic control unit based on non-disrupted days within the same service week, as well as other days from different weeks, which serve as the control group.

4. Data and data processing

This research applies the Synthetic Control method on the Orange line of the Washington DC metro network. The real-time GTFS data is gathered from January 29th, 2024, 10:00am to March 24th, 2024, 11:59pm at a 1-minute interval in order to avoid any holiday. This period is covered for two reasons: it not only offers a larger sample of control units to better assess the causal inference effect, but also leaves room for potential collection error. Before applying any statistical methods, the gathered data need to be formatted, so the real-time feed matches the schedule. It is only possible to utilise Synthetic Control technique on the data if it is structured as a table listing for all the stations of the line: the scheduled arrival times, delays, longitude, latitude, and other GTFS parameters.

To prepare the 2-month table for the Orange line, several steps are performed to transform and structure the GTFS Schedule data appropriately. First, the trips dataset needs to be filtered to extract only trips on the Orange line heading in the specified direction. The entire network can be simplified to a 1D analysis along a single line to study disruption and delay propagation. This is because, once the delay propagation is quantified, whether the stations belong to the same line has minimal impact. Any remaining effects can be accounted for by introducing a variable that represents the degree of connection between the stations. Furthermore, although GTFS Schedule proposes data for both directions of the same lane, only one is considered since this research focuses on highlighting the impact of the disruption rather than drawing any overall outlook. Finally, it is likely that any trend emerging from the analysis of one direction would just be mirrored by the opposite direction, making the whole process redundant.

Joined to the filtered trips dataset, the stop_times and stops datasets compile detailed scheduled arrival times at every stop for every trip on the line. A visual representation of the merged database is shown in Figure 1. *Nota bene*, GTFS arrival_time fields require minor adjustment as it needs to be converted from text type into a consistent 24 hours datetime object type. Additionally, overnight trips crossing midnight need to be adjusted from the 25th hour of the first day of service to the first hour of the following day. The arrival_time is eventually updated to reflect this study scenario and is set to a base date of January 29th, 2024, as the initial database does not provide any details about the date. At last, the schedule database is duplicated to extend to the observation period, providing a robust consistent sample size for the analysis. This data preparation ensure that the schedule table is in the correct format before combining it with the real-time GTFS feed since the Washington Metropolitan Area Transit Authority (WMATA) does not explicitly specify the delay in its trip update feed.



stops	
stop_id	**varchar**
stop_name	varchar
stop_desc	varchar
stop_lat	float
stop_lon	float
zone_id	int
location_type	int
parent_station	varchar
wheelchair_boarding	int
level_id	varchar

stop_times	
trip_id	**varchar**
arrival_time	timestamp
departure_time	timestamp
stop_id	varchar
stop_sequence	int
pickup_type	int
drop_off_type	int
shape_dist_traveled	float

trips	
route_id	varchar
service_id	int
trip_id	**varchar**
trip_headsign	varchar
direction_id	int
block_id	void
shape_id	int
scheduled_trip_id	int
train_id	int

Figure 1: Diagrammatic representation of the GTFS Schedule database

The stop_id and arrival_time variables serve as double identifier to connect the real-time GTFS information to the schedule table. Although most alerts messages remain, the disruption is considered as a treatment from a causal inference perspective, thus should be considered a one-off event rather than a period of time; the disruption variable is set equal to 1 only when it happens. In this paper, any GTFS alert type is considered a disruption, be they technical problem, weather, maintenance, medical emergency or other cause (except holiday). Some alerts messages are ignored as they signal that normal service has resumed.

There are two sources of information to identify the real arrival time from the real-time GTFS feed: the trip_update and vehicle_position datasets. Both gather and provide minute-by-minute update of their respective information in the form of JSON files. In spite of GTFS norm guidelines, WMATA does not provide delay in its trip_update feed. Additionally, because the schedule database is built from repeated stop sequences over time, the trip_id cannot be used to match arrival times. Hence, the vehicle_position dataset will be leveraged to the record all trains whose GPS coordinates (with a precision of 14 decimal places) match the GPS coordinates of any Orange Line station (with a precision of 8 decimal places) within a 100-meter radius, where the coordinates are rounded to 3 decimal places. The Algorithm 1 explains how to match real-time vehicle position data, extracted from JSON files, with stop locations recorded in the schedule matrix. The algorithm iterates through each row of the 1-month schedule data base. For every space-time tuple (stop coordinate and schedule arrival time) the algorithm proposes to open the corresponding JSON file and check for every train position. If one of the trains is found to be at the designated stop, it is considered on time and the file timestamp is recorded as the real arrival time. If no train is found within a 100-meter perimeter from the station, the algorithm opens the next file, looking for a train potentially matching the criteria one minute later. The maximum number of attempts is set to 15 since WMATA announces a service frequency of 15-min. If a train is found at the desired stop 15 minutes later, the later will be considered on time for the next scheduled arrival time while the current arrival time will be considered cancelled. The output table is called the real-time matrix and gathers schedule arrival times, disruption states, and real arrival times.

ALGORITHM 1: PROCESS JSON FILE TO EXTRACT MATCHING TRAIN TIMESTAMP

```
Input:
  • Scheduled Stop Data: A Data Frame containing stops' latitude (float), longitude
    (float), scheduled arrival time and file addresses (string)
  • Vehicle Position Data: A collection of JSON files containing real-time vehicle data,
    organized by timestamp in the file name
  • max_attempts: Maximum number of attempts to try opening subsequent files
    (integer)
Output: actual_AT: Timestamp of the train matching the criteria if found; otherwise,
None

1   Initialization of variables:
2       Parse json_path_base to extract the initial timestamp
3       Set attempt counter to 0
4   while (attempt < max_attempts) do
5       Try to open the JSON file of the corresponding timestamp
6       if file is opened successfully then
7           for each entity in the JSON data:
8               if vehicle information exists in the entity then
9                   Extract trip_info and position_info from vehicle
10                  if the following conditions are satisfied:
                      • trip_info['start_time'] must be earlier than the current timestamp
                      • trip_info['route_id'] == 'ORANGE'
                      • trip_info['direction_id'] == 0
                      • Rounded position_info['latitude'] equals rounded row_latitude
                      • Rounded position_info['longitude'] equals rounded row_longitude
                  then
11                      Return the timestamp from vehicle_info
12                  else:
13                      Increment the timestamp by 1 minute and repeat from Step 5
14      if no match found after maximum attempts then
15          Return None
16  end
```

Finally, in order to match the Synthetic Control requirements, the real-time matrix which is an unbalanced panel dataset needs one last adjustment. SC requires a balanced panel where each unit has consistent observations across all time periods. In other words, because some stops have more scheduled (and recorded) arrival times than others, the table need to evenly distribute the results. Consequently, the current schedule, which consists of unique pairs of stop_id and arrival_time, is hourly aggregated and summed over all stops in order to portray the line performance as a whole. This will balance the panel and prepare the dataset for the application of SC on a disrupted day to compare its hourly delay evolution with a synthetic control day.

5. Results and observations

This study applies Synthetic Control to generate a synthetic control day of operation on the Orange line. A look at the input real-time matrix provides insights into the relationship between disruption and delay and further supporting the need for this study. Before aggregation, the real-time matrix counts 115,680 rows (12,414 are missing, 10.731%), *i.e.* more than 115,000 unique couple of stop location and arrival time (either scheduled or real-time). Note that some of the observed delays are almost 15-minute-long and might actually be trains that were ahead of schedule by a couple minutes. An hourly aggregated version of the real-time matrix serves as a balanced input on which to apply Synthetic Control. Some outliers, including: the stations “PF_K01_C”, “PF_C01_C”, “PF_D13_C”, and the whole day of February 24th (for the entire line), have been excluded. All the latter corresponding values where at least 3 to 6 times lower than the average of the other stops of days, even sometimes with negative values, venturing far off the average hourly aggregated delays.

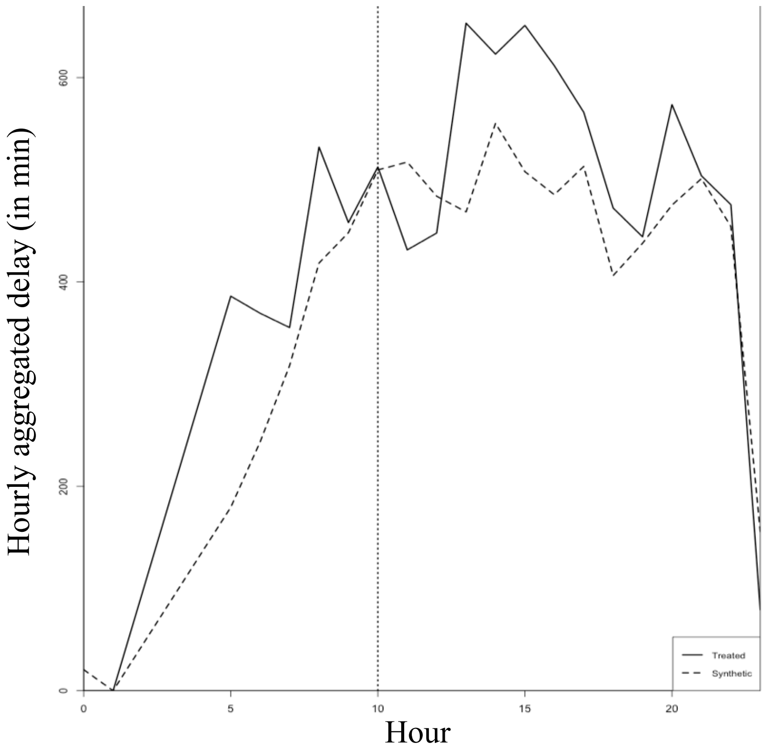


Figure 2: Hourly aggregated delay along the Orange line on a disrupted day, against a Synthetic Control day

The treated (disrupted) unit is Wednesday 20th March, while the leveraged control units are the other non-disrupted Wednesdays observed over the period of the study, and the other days of the corresponding week. The graphic output by the SC method is displayed in the Figure 2. In this graphic, both the treated unit and control group follow the same upside-down U shape. This similarity lasts up until the point of disruption occurrence when the recorded hourly aggregated delay of the treated unit starts following a different path. Peaking at 653 minutes three hours after the disruption occurrence at 1pm EST, this represents over the 23 remaining stations accounted in this study, a distributed delay of 28.4 minutes of delay for the total of scheduled trains for the 8th hour of service on that day.

The more unit are added to the control group the more accurate the SC. This first estimate of a synthetic control group, can be refined for another disrupted day, say Tuesday, by feeding more observed Tuesdays in the SC. The weights previously assigned to the Wednesdays results might be readjusted to eave more room for the relevant days of comparison. Ultimately, the Synthetic Control method could benefit from big data by aggregating an increasing amount of data for a more pertinent control group.

In adapting Synthetic Control to use time as the treatment unit and control group, this research proposes a novel approach where the traditional distinction between pre-treatment and post-treatment periods is redefined. Here, the pre-treatment period encompasses all time points leading up to the disruption on 20/03/2024, 10am EST. Uniquely, the post-treatment period is constrained to this single day, with subsequent time points being reclassified as part of the pre-treatment period for the purposes of the analysis. This approach effectively creates a loop where time periods following the treatment can be utilised to simulate counterfactual scenarios, thereby providing a robust mechanism to estimate what would have occurred in the absence of the treatment. The remaining periods are repurposed as a variant control unit to strengthen the pre-treatment model, ultimately enhancing the accuracy and reliability of the already robust Synthetic Control method.

6. Conclusions and recommendations

The research employs the Synthetic Control method, which is a robust approach to estimate the causal impacts of interventions. This is particularly useful in scenarios where randomised trials are not feasible, like in transport. This approach is applied to the Orange line of the Washington DC metro system to create a synthetic benchmark day of operation, allowing for a more accurate measurement of the system's resilience to delay. In adapting Synthetic Control to use time as the treatment unit and control group, this research proposes a novel approach where the traditional distinction between pre-treatment and post-treatment periods is redefined. Here, the pre-treatment period encompasses all time points leading up to the disruption (on 20/03/2024, 10am EST).

The findings highlight the method's ability to generate suitable synthetic day of operation, which are crucial for developing effective benchmarking. However, while the method proves robust in this context, its application is limited by the constant availability of high-quality data. Future work could focus on expanding the application of this method to other transit networks and expand it for spatial units benchmarking. As mentioned before, the alert message does not carry any spatial data, blocking the development of a synthetic station performance. The spatial approach of benchmarking might challenge the SUTVA assumption since connected stops are interacting units.

Despite these limitations, the research makes a significant contribution to the field of urban transit management by offering a novel approach to performance benchmarking that prioritises robustness and scalability. One way to enhance the delay propagation estimation is to reckon with the influence of the delay in a station *i* (Y_i) or the disruption state of the same station (D_i) as a confounder of the disruption state in another station *i'* ($D_{i'}$). Another case would be the influence of the delay at a time *t* over the following time period *t+1*. Such scenarios are summarised in Figure 3 where *Y* is the delay, *i* the spatial index of the metro station, *t* the time index, X_k the *k* covariates for a given station at a given time and *D* the disruption state.

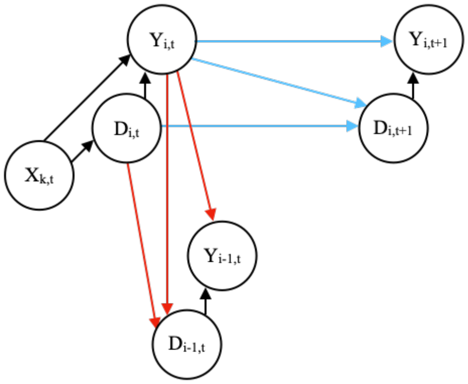


Figure 3: Summary of possible confounding bias to study (space propagation in red, time propagation in blue)

To proceed further analysis and slightly drift away from causal inference, the previously set equation (3) can be used to establish a time-to-recovery metric, denoted as τ, to reach a certain arbitrary threshold ξ, as shown by the inequality (4). It is then possible to repurpose the penalty term of a smoothing spline (James, et al., 2021) to conceive a new type of Key Performance Indicator (KPI), shifting towards a prediction problem.

$$Y_{1,x}(\tau) - Y_{0,x} \cdot W^*(V) < \xi_t \tag{4}$$

To be more precise, the regression outlined in (5.1) not only offer to display τ on a map of the network's stops and at different time of the day or the week, for better visualisation, but also provide an estimate that can be used to refine the penalty term under a new from displayed in equation (5.2).

$$\tau = \hat{f}(x, t, \dots) \tag{5.1}$$

$$\int_t \int_x \hat{f}''(x,t)^2 \, dx \, dt \tag{5.2}$$

From a mathematical point of view, the second derivative of the regression will serve as a measure of time-to-recovery "erraticism". A wiggly regression function demonstrates an uneven network, with a time-to-recovery inconsistent over time and space. The two integrals measure the overall variability of τ over time and space, thus yielding a comprehensive overview of the network. This new KPI could guide operators in benchmarking and comparing two networks, one often being a little late and another rarely but heavily disrupted.

Finally, notwithstanding this paper outcome on Synthetic Control relevancy in urban transit network benchmarking, one can argue operator interventions often tend to be non-binary, thus limiting the range of possible analysis with the existing methods. This need for mathematical tools is coupled with an increasing need to assess the causal inference of operators' interventions of continuous values.

References

Abadie, A. et Cattaneo, M. D., (2018). Econometric Methods for Program Evaluation. Annual Review of Economics. 10(1), 465–503. Available from: doi: 10.1146/annurev-economics-080217-053402

Adjetey-Bahun, K., Birregah, B., Châtelet, E. and Planchet, J.-L., (2016). A model to quantify the resilience of mass railway transportation systems. Reliability Engineering & System Safety. 153, pp. 1–14. Available from: doi: 10.1016/j.ress.2016.03.015

Derrible, S., Kennedy, C., (2010). The complexity and robustness of metro networks. Physica A: Statistical Mechanics and its Applications. 389(17), 3678–3691. Available from: doi: 10.1016/j.physa.2010.04.008

Graham, D. J., (2021). Causal inference for ex post evaluation of transport interventions. In: International encyclopedia of transportation. Elsevier. pp. 283–290.

Graham, D. J., McCoy, E. J. et Stephens, D. A., (2016). Approximate Bayesian Inference for Doubly Robust Estimation. Bayesian Analysis. 11(1), 47–69. Available from: doi: 10.1214/14-ba928

James, G., Tibshirani, R., Hastie, T. and Witten, D., (2021). Introduction to statistical learning: with applications in R. 2nd Edition. Springer.

Lu, Q.-C., (2018). Modelling network resilience of rail transit under operational incidents. Transportation Research Part A: Policy and Practice.117, 227–237. Available from: doi: 10.1016/j.tra.2018.08.015

Serafino, M., Cimini, G., Maritan, A., Rinaldo, A., Suweis, S., Banavar, J. R. and Caldarelli, G., (2020). True scale-free networks hidden by finite size effects. Proceedings of the National Academy of Sciences. 118(2), article no: e2013825118. Available from: doi: 10.1073/pnas.2013825118

Singh, R., Graham, D. J. and Anderson, R. J., (2019). Characterizing journey time performance on urban metro systems under varying operating conditions. Transportation Research Record: Journal of the Transportation Research Board. 2673(7), 516–528. Available from: doi: 10.1177/0361198119848415

Singh, R., Hörcher, D., Graham, D. J. and Anderson, R. J., (2020). Decomposing journey times on urban metro systems via semiparametric mixed methods. Transportation Research Part C: Emerging Technologies. 114, 140–163. Available from: doi: 10.1016/j.trc.2020.01.022

Zhang, N., Graham, D. J., Hörcher, D. and Bansal, P., (2021). A causal inference approach to measure the vulnerability of urban metro systems. Transportation. Available from: doi: 10.1007/s11116-020-10152-6