CISIL DATA CHALLENGE QUESTION 1
METHODS REPORT

David Ami Wulf **Pablo Geraldo** **Sarah Sotoudeh** **Adam Rohde**

Faculty supervisors: Onyebuchi Arah & Chad Hazlett

October 25, 2022

Question 1:

1. How did the October 1, 2020 reinstatement of fares affect ridership on King County Metro?
2. Did the reinstatement of fares by King County Metro have differential effects on ridership among different socio-economic groups?

1 Background

Fares for King County Metro Transit were suspended from 3/21/2020 until 10/1/2020. Fare suspension was a measure taken to reduce the contact between bus drivers and riders as well as to reduce individuals handling cash during the early stages of the COVID-19 pandemic. During the summer of 2020, consensus developed among the King County as well as national transit community that fares could be reinstated. During this time, plexiglass barriers were also installed in buses that reduced the contact that drivers would have with riders. Fare reinstatement for King County Metro Transit was driven by the need for revenue. For accounting purposes, King County Metro Transit decided the reinstatement of fares should happen on the first of a month. Initially 9/1/2020 was considered but 10/1/2020 was ultimately chosen. Drivers were told to be lenient in their enforcement of fare collection in the time immediately after fares were reinstated; additionally, drivers are told not to argue with riders who are not willing to pay fares.[1] Further, on 9/19/2020 major service changes were implemented to many of the King County Metro Transit bus routes.[2]

Question 1 deals with how ridership changed as a result of the reinstatement of fares. The CISIL data challenge defines "ridership" as stop-level passenger boardings (getting on the bus) and alightings (getting off the bus). In what follows, we focus on boardings, since boardings and alightings track each other very closely overall (see Figure 1). Notice that, as we might expect, the day of week has a large influence on ridership, which we take into account in our analyses. With this background in mind we next consider our research design.

2 Research Design

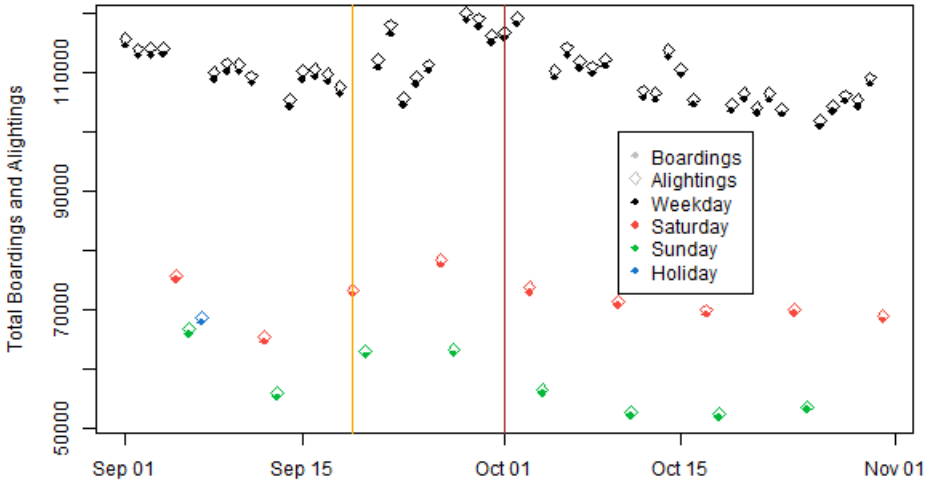
To address Question 1, we will consider fare reinstatement (i.e., passengers having to pay fares again after 10/1/2020) as our treatment variable and ridership (i.e., boardings) as our outcome variable. We are interested in the effect of fare reinstatement on ridership across the entire King County Metro Transit bus system. Since fares were reinstated for the entire King County Metro Transit system on 10/1/2020 (and no data is available for any similar transit system, like Sound Transit[3]), we have a setting in which all units (either routes, trips, stops, etc.) are untreated before 10/1/2020 and all units are treated on and after 10/1/2020. Therefore, in the post-treatment period, there is no untreated group against which we can compare the treated group. We do have data on stop level ridership across the pre-treatment month 9/1/2020 - 9/30/2020 as well as across the post-treatment month 10/1/2020 - 10/31/2020. As a result, we can use the information from the pre-treatment period, in which all units are untreated, as a basis of comparison for the

[1] Information in this paragraph provided by CISIL and KCMTD (2022).

[2] Refer to Switzer (2020) for which routes had reduced service.

[3] CISIL and KCMTD (2022)

Figure 1: Total Boardings and Alightings by Date



Note: Daily boardings and alightings track each other very closely. In our analysis, for simplicity, we focus on boardings.

same population in the post-treatment period, in which all units are treated. This type of research design is similar to one-group pretest–posttest designs that have been common since at least the 1950s (Campbell, 1957). Such a design can be approached in multiple ways. As we discuss further below, we consider simple pre-versus-post comparisons between matched units in addition to interrupted time series approaches. The main substantive assumption of these approaches is that, absent the treatment (reinstatement of fairs), the pre- and post-treatment levels would have been approximately the same. This can also be conceptualized as a claim of "stability" in how ridership behaves in the two periods, and consequently the "comparability" of outcomes just before and just after the change. We attempt to identify factors that might render such comparability less credible, i.e. reasons that the outcomes would change between time periods other than the treatment in question. When possible attempt to account for these factors in our data preparation and models. The ultimate credibility of the conclusion rests on how certain we can be that, after possible adjustments, there are not remaining unobserved factors whose over-time change is driving the result rather than the change in policy over the period of comparison.

3 Data

3.1 Stop-Level Ridership Data

There are three relevant data levels: Route, Trip, and Stop. A route ID largely corresponds to the number riders see on their bus and bus stop. For example, the number 42 line leaves several times per day in both inbound and outbound directions. A trip ID corresponds to a single one of those scheduled departures for buses on that route. For example, there is a unique trip ID for the inbound number 42 bus starting its journey from the origin stop every Tuesday at 12:42. Finally, a stop ID corresponds to a given physical bus stop that many trips from multiple routes may stop at over the course of a day. This stop-level ridership data was provided by CISIL and KCMTD (2022).[4]

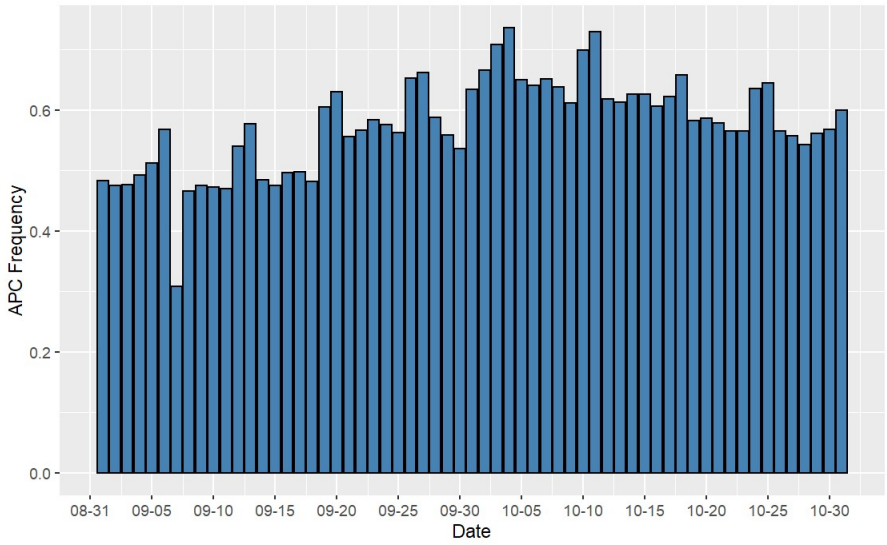
The central complication arises in what data we have access to. We would like to know how many passengers boarded at every stop. However, only some physical buses have the capability to count passengers that board and alight at each stop. These Automatic Passenger Counters (APCs) are featured on most buses, and transit authorities suggested that those buses were assigned to trips as good as randomly (i.e. those choosing which bus ran which trip on a given day did not do so with any intention to tie APC and non-APC buses to different trips). There are two ways to check this in the data. First, we could look at whether APC-enabled trips differ from non-APC trips as determined by the

[4]These data do not include labor day (9/7/2020).

trip-specific variables to which we have access. Second, we could look at whether analysis methods applied to just those trips with APCs produce different results than when applied to all trips. Any such differences might suggest that there are differences between APC trips and non-APC trips, even if we do not see systematic differences in top-line summary statistics about those trips. Note that even though individual buses may be assigned without intentionally directing APC buses to particular routes, cycles in the availability of such buses may mean that there are differences between those trips with and without APCs.

Figure 2:

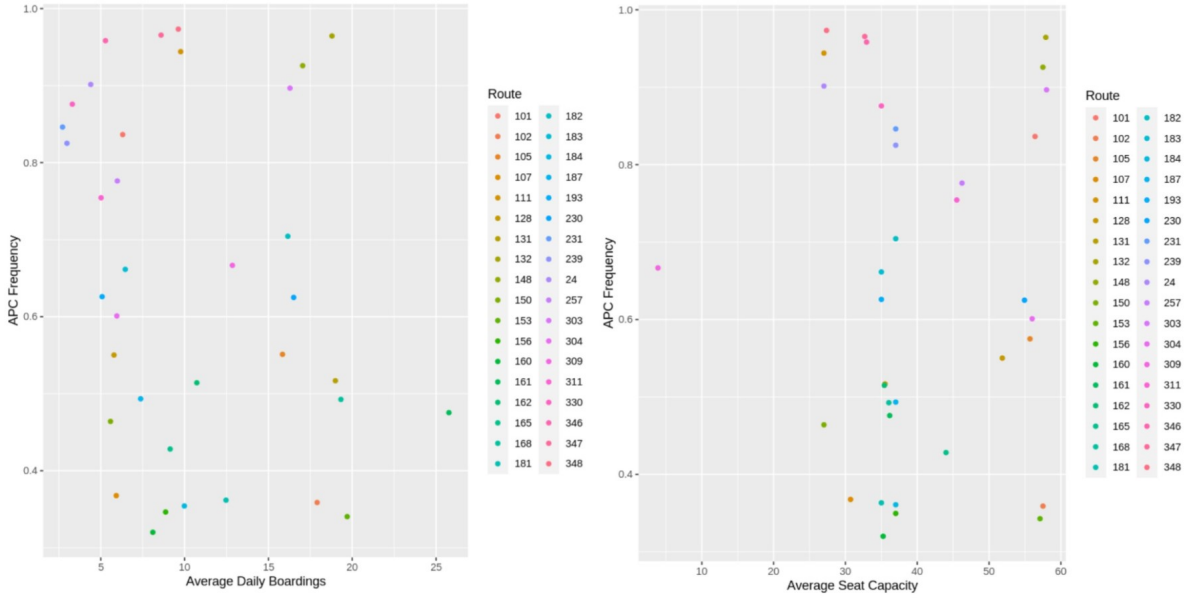
APC Trip Frequency Overtime - Full Service Routes



Note: APC Trip Frequency is consistent over time across the same day of the week; however, we do see an increase in APC Trip Frequency after the September 19 Service Change.

Figure 3:

APC Placement Patterns - Full Service Routes



Note: We see no significant relationship between APC Frequency and average daily boardings or average bus seat capacity.

Beginning with the former approach, we investigated whether we saw any patterns in how APCs were placed across routes. First, we analyzed APC frequency by route to see if there were any overwhelming problems with certain routes that we were going to be using in our analyses that we needed to address. We calculated APC frequency by tabulating the number of times each trip had an APC on it given the number of times it was listed in the APC records dataset and then divided that by the total amount of times that trip was scheduled to run. We took the APC frequency by trip calculations and determined the APC frequency by route calculations by grouping appropriately. We found that trips apart of the full service routes had a mean and median APC frequency of 0.611 and 0.555 respectively. Routes apart of the full service routes also had a mean and median APC frequency of 0.475 and 0.493 respectively. We see in Figure 2 that the APC frequency generally stays the same over time from a given day of the week to the same day of the week; however, there is a concerning rise in APC frequency after the September 19 service change. Full service routes were classified as operating at full levels, so assuming that the number of trips stayed the same for each route (as expected for a full service route) this change could indicate a higher number of APCs being distributed to more buses post September 19. Another potential explanation for this increase could be that full service routes include routes that operated at full service levels but were restructured through the Rent, Kent, and Auburn Area Mobility Plan. Therefore, this restructuring could increase the number of trips in a given route. We also investigated the relationship between APC placement and average boardings as well as APC placement and average bus seat capacity and did not see any obvious relationship for either as shown in Figure 3. The average daily boardings was calculated by finding the average boardings by trip and then weighting each trip appropriately (depending on the number of days per week it ran) and then grouping by route to find the average boardings per route. As stated earlier, the data challenge organizers have said that APCs were not intentionally distributed on certain buses over others. However, given the change in APC frequency that we observe over time, we proceed our analyses with an understanding that there could be unintended or immeasurable factors influencing APC frequency and placement.

Taking our second approach to identifying trip differences based on APC presence, we conducted a week-to-week matched analysis as discussed below in the methods section. By narrowing down the trips we look at to those with APCs present in successive weeks, we can ensure we are measuring only those week-to-week changes for which we have observed data. If APCs really were assigned randomly, this subsetting would not affect our estimates of week-to-week changes. If instead those trips with APCs were different from other trips, the trips about which we have data may differ day-by-day, and we might see such differences appear in the measured week-to-week shifts. We did in fact find differences in boarding trends, often significantly, between the two populations (all trips vs APC-assigned trips only), though such differences were not persistently in one direction or the other. This neither negates the possibility that APCs were assigned non-randomly, nor proves they were, though it suggests an analysis subsetting to trips without missing data is worthwhile. As a result, we consider the matched analysis below an important tool to remove any potential biases in APC assignment. This approach is worthwhile regardless of the underlying APC assignment mechanism; if APCs were assigned randomly, the matched analyses should give the same results as non-matched analyses, and if they were not assigned randomly, the matched analyses will be less biased.

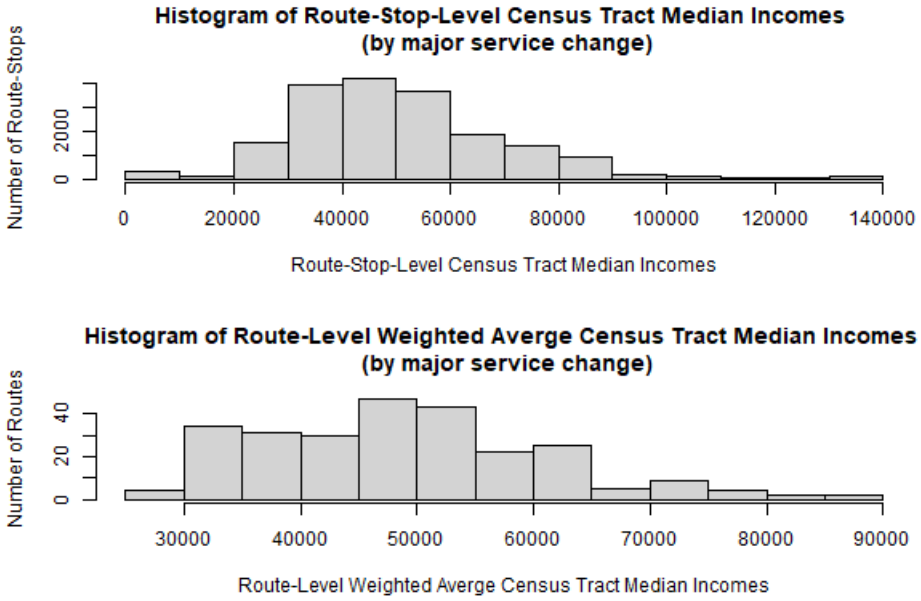
3.2 American Community Survey Data

Rather than use the 2019 American Community Survey data provided by CISIL and KCMTD (2022), we pulled 2020 data on median income by census tract.[5] We do this to ensure that we are able to properly assign census tract data to stops and do not believe that there is risk of this data being effected by the reinstatement of fares. There are of course other ways we might measure socio-economic status other than median income but we opted for this simple measure. We follow the approach suggested by CISIL and KCMTD (2022) for determining which census tract each bus stop is in. There are 200 stops that don't get assigned a census tract. We then join the 2020 ACS census tract median income data to each stop by census tract. There are multiple ways that we might use this data to measure socio-economic status. We consider a measure of the income level of the entire service area of routes, rather than try to parse or model how riders of varying socio-economic statuses might ride.[6] Our simple approach is to take a weighted average of stop-level median incomes across the entire route, where we weight by boardings. This provides a measure of the socio-economic status of the entire service area of routes, which riders traverse. We do this for routes separately before and after the 9/19/2020 service change, since the stops could change after this date. We do not consider any other minor service changes, however. For the purposes of our analyses, we look at strata of this route level weighted average median income as our income strata. We divide the routes in thirds based on this measure. This is somewhat arbitrary but allows us to compare the low income service area routes to the high income service area routes. See Figure 4 for histograms of the stop-level and route level income metrics.

[5] We use the variable `B06011_001E`. See U.S. Census Bureau (2020).

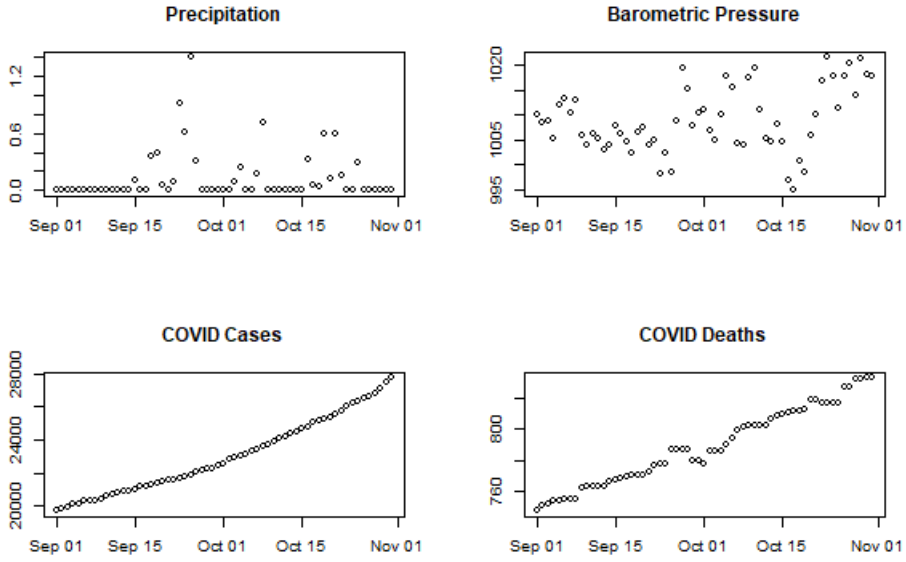
[6] We don't have any information on the riders. We don't know if people are boarding or alighting where they live or work or otherwise visit. So we are not able to cleanly make assumptions about socio-economic status of riders at the stop level.

Figure 4: Histograms of Median Incomes



3.3 Supplementary Data

Figure 5: Supplementary Data



We use two additional external sources of data. First, we use King County COVID-19 cases and deaths by date from the New York Times.[7] We specifically use the 2020 county level daily data found in the file `us-counties-2020.csv`.

[7] See The New York Times (2021).

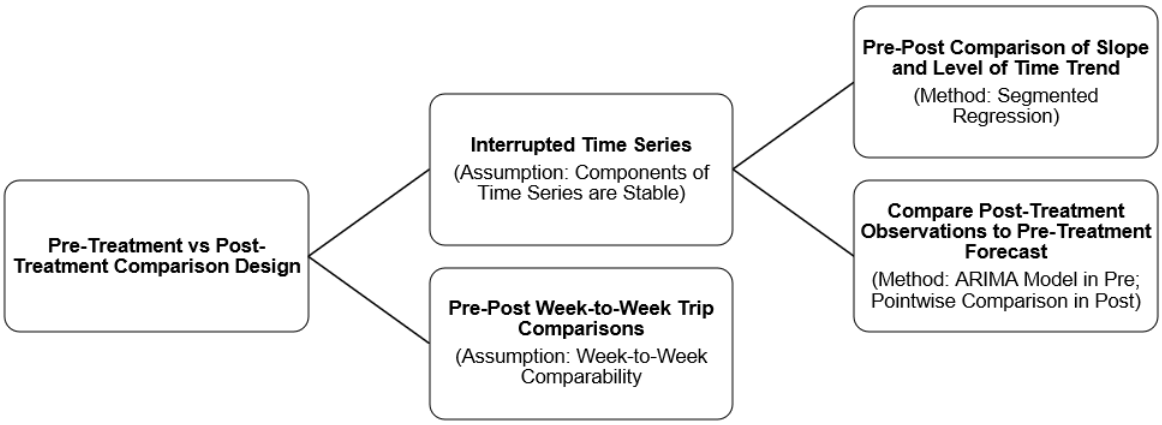
Second, we use daily precipitation and barometric pressure.[8] We choose to use just precipitation and COVID-19 cases as covariates in our analyses. See Figure 5 for these measures over time.

4 Methods

We take two approaches to evaluating the impact of fare reinstatement made in the context of the research setting described above. For our first approach, we make a simple assumption that similar trips (where similarity here means exactly matching on route, day of week, direction, and departure time of day) from one week to the next should not exhibit major changes in ridership, other than due to large changes to the system like reinstatement of fares. We generate these week-to-week matched comparisons between boarding counts on the same bus trips one week apart, assessing whether comparisons across the treatment boundary display substantively different changes than other week-to-week comparisons. For our second approach, we assume that there are components of boardings that are best treated as a time series in which observations can be correlated over time and that the components of this time series are stable over time, other than in the wake of large changes to the system like reinstatement of fares. Viewing the reinstatement of fares as an interruption to the standard time series progression, we use trends in the pre-treatment period as a basis of comparison for observations in the post-treatment period, either by measuring changes at the boundary or by making model-based predictions about a hypothetical untreated post-treatment period.

The matched comparisons should provide more precise, internally valid estimates for the subset of the data for which matches are possible. This approach narrows the focus of the analysis but avoids some of the issues surrounding APC assignment. The interrupted time series analyses provide estimates of the effect of fare reinstatement that encompass all of the available data. Such analyses may be more vulnerable to data biases but also incorporate external controls, assess longer term impacts, and allow a broader view of the threats to validity. We now describe each approach in more detail.

Figure 6: Flowchart of Methods



Note: We take two approaches to evaluating the impact of fare reinstatement. These are interrupted time series and pre-post week-to-week trip comparisons. We also use two methods for the interrupted time series approach - comparing the slope and level of the time trend in boardings and comparing post-treatment observed boardings to pre-treatment forecasts of post-treatment boardings.

4.1 Pre-Post Week-to-Week Trip Comparisons

Given that the main substantive assumption to identify the effect of fare reinstatement is that, absent such "treatment", ridership in the post-treatment period is comparable to ridership in the pre-treatment period, one approach to make this claim credible is making sure that the individual units we are comparing in each period are as similar as possible. In this analysis, we attempt to maximize the credibility of this comparability by comparing boarding counts only for identical trips, one week apart. In practice this means that we look at a given trip (a route leaving at a certain time on a certain day of the week) and the same trip exactly one week later, calculate the difference in total boardings along those two trips, and then average across all such trip pairs on the same day. Recall that not all trips include APCs, so this strategy gives us an analysis subset of fully matched data, which represents a targeted, clean set of comparisons, albeit one with less external relevance should there be unobserved differences between APC trips and non-APC trips.

[8] That is, daily accumulated precipitation (in inches) and mean barometric pressure (mbar) for water year in 2020 for King County. See King County (2020).

If ridership were steadily rising over a given period, we would expect to see increases of similar magnitude in these week-to-week comparisons, and vice versa. If week-to-week comparisons that cross the boundary represented by the treatment (fare reintroduction on 10/1/2020) display shifts that do not differ meaningfully from those comparisons made wholly before and wholly after that boundary, we cannot reject a null hypothesis that the treatment itself did not change ridership. If boundary-crossing shifts *do* differ substantially from other shifts, we interpret those shifts as being causal results of the treatment itself. The threat to validity of such a conclusion would be the presence of other changes occurring at the time of the treatment introduction that could be causing the shift in ridership.

To generate empirical assessments of the significance of any treatment effects, we conduct permutation inference by drawing randomized sets of week-to-week comparisons of the same number of days as the treatment comparison days, measuring how often we see a cumulative shift of the same or greater magnitude as the true treated comparisons. This could produce incorrect uncertainty intervals if there are day-to-day ridership correlations that could heighten the impact of our consecutively treated days, but that would not be captured in the randomly permuted treatment days. See Ernst (2004) for more on permutation inference.

4.1.1 Selecting treatment and control dates

There are several specific week-to-week comparisons that we decided to exclude from our analysis before looking at the results. First, we excluded those 7 comparisons crossing the service change boundary on 9/19/2020, as shifts in ridership between trips (and away or towards using the transit system generally) due to lowered trip availability, do not represent general ridership changes against which we want to compare the treatment-impacted week-to-week shifts. Similarly, we excluded week-to-week shifts for which the earlier or later comparison data was a holiday (the Mondays of Labor Day and Indigenous Peoples' Day).

To a lesser extent, we also considered shifts involving the weekends prior to those two holidays to represent potentially unrepresentative shifts, as transit behavior may respond to long weekends, and considered the shift leading into Halloween to represent another potentially unrepresentative shift.

In terms of treated week-to-week shifts, we considered which comparisons represent Intention To Treat (ITT) dates, and which produce "pure" treatment effects. Given the intended leniency of fare enforcement immediately after reintroduction, and the time it might take for riders to adjust their practices, we viewed the week-to-week comparisons ending on Thursday 10/1/2020, Friday 10/2/2020, and Saturday 10/3/2020 to represent intended treatments (ITT days) but less pure measures of the true treatment effect. We include three days in this category as we expect weekday and weekend ridership to be not fully overlapping populations, which may need to experience a ride before changing behavior due to the treatment. Similarly, we view the week-to-week comparisons *starting* on those three days and ending a week later to contain a messy middle-ground between representing treatments and representing controls. For the pure treatment effect measurement, then, those three days are also excluded. Effectively, the ITT analysis looks at the set of days we might have expected to see a week-to-week effect (even if watered down by delayed impact of the treatment), while the pure treatment analysis looks at those days we believe to be impacted by the full effect of the treatment.

In summary, for the ITT analysis we use the full week following 10/1/2020 as the treated units, and all other units save the service change week and the three holiday Mondays as control units. For the pure treatment analysis, we use only the 4 treated days ending between 10/4/2020 and 10/7/2020 as treatment days, and exclude from the control the service change week, the entirety of holiday weekends, and the sets of 3 days on either side of the treatment dates. This pure treatment analysis may suffer due to its limitation to only 4 treatment days, as we could miss any weekly "seasonality" of the effect. We complete this ITT and pure treatment assessment using the entirety of the route list, as well as separately within each income-stratified third of the route lists to detect any differences between the effect of the treatments on each subset.

4.2 Interrupted Time Series

We can construct a time series for boardings, across both the pre and post-treatment periods. This approach assumes that there are components of boardings that are best treated as a time series in which observations can be correlated over time and that the components of this time series are stable over time, other than in the wake of large changes to the system like reinstatement of fares. We can use these time series to evaluate how the pre-treatment outcomes compare to the post-treatment outcomes and gain some understanding of how the reinstatement of fares affected boardings. Such an analysis is called an interrupted time series[9] In what follows, we will consider daily boardings as our time series.

[9]Time series have a few potential issues that must be considered in this context. These include stationarity, autocorrelation, and seasonality. To attribute causality to effect estimates, time-varying confounding must also be considered, in which variables that explain the outcome vary across time and hence act as confounders, as pre-treatment outcomes can differ from post-treatment

4.2.1 Pre-Post Comparison of Slope and Level of Time Trend

A simple way to evaluate the effect of fare reinstatement on boardings is to compare the level and slope of the time trend in boardings in the pre-treatment period to those in the post-treatment period. Such a comparison can be done using a segmented regression. Segmented regression is perhaps the simplest approach to interrupted time series analysis.[10] It requires we make an a priori assessment of the form that we believe the intervention should have on the outcome time series. For instance, do we expect the intervention to lead to a level change in the outcome series or perhaps a slope change or both a level change and a slope change. Other forms are also possible. In our setting, fare reinstatement might be considered to most likely lead to a downward level change. Some riders are willing to ride in the absence of fares but perhaps not in their presence. Though perhaps there could be an adjustment period in which ridership decreases to a new level. Interrupted time series with segmented regression, allowing for level and slope changes, takes the form

$$Y_t = \beta_0 + \beta_1 T + \beta_2 D_t + \beta_3 (T \times D_t) + \beta_4 X_t + \epsilon_t$$

where Y_t is the outcome at time t, T is the number of time periods before or after intervention (the number of periods before are negative)[11], D_t is an indicator for whether or not treatment as implemented at time t, and X_t are covariates at time t. β_1 is the time trend (slope) in the pre-treatment period. β_2 is the jump in boardings after fares were reinstated. β_3 is the change in the time trend (slope) after fares were reinstated.

4.2.2 Comparison of Post-Treatment Observations to Pre-Treatment Forecast

Rather than simply compare changes in the slope and level of the time trend in boardings, we might also consider modeling the time series using a more flexible approach in the pre-treatment period and then using that model to forecast what boardings would have looked like in the post-treatment period if fares had not been reinstated. We can then compare these forecasts with the observed post-treatment boardings to evaluate the effect of fare reinstatement. This approach takes the following steps:[12]

1. estimate an ARIMA (autoregressive integrated moving average) model in the pre-treatment period
2. forecast outcomes in the post-treatment period using the ARIMA model from the pre-treatment period
3. compare the observed outcomes and the forecasted outcomes in the post-treatment period for each/any time point; this provides an effect estimate for each time period

ARIMA methods do well modeling fluctuations in time series data, lending reliability to their use for interrupted time series analysis.[13] The key difference between ARIMA and segmented regression is that Y_t is regressed on the outcome at previous time points (i.e., $Y_{t-1}, Y_{t-2}, \dots$) and not on time itself (i.e., T). The components of an ARIMA model include[14]

- autoregressive model (AR): Y_t is predicted with lagged versions of Y_t (i.e., $Y_{t-1}, Y_{t-2}, \dots$).
- moving average model (MA): Y_t is predicted with lagged versions of the error term (ϵ_t).
- seasonal model: Y_t is predicted with lagged versions of Y_t at a regular interval.
- differencing or intergration (I): rather than directly using Y_t, we might use $Y_t - Y_{t-1}$ (and possibly additional differences) to capture the overall trends in the series.

ARIMA models are typically referred to in terms of the number of terms for the different components: ARIMA$(p, d, q) \times (P, D, Q)_S$, where p is the number of lags in the AR model, d is the number of non-seasonal differencing, q is the number of lags in the MA model, P is the number of seasonal AR terms, Q is the number of seasonal MA terms, D is the number of seasonal differencing, and S is the seasonality interval.[15]

outcomes in a systematic way that is unrelated to the treatment. In this work, we consider precipitation and COVID cases as possible time-varying confounders. All the analysis here is subject to the caveat that unobserved time-varying confounders could exist. However, given the relatively short period under analysis, we believe many such factors will not present problems (e.g., economic conditions or population trends). See Bernal et al. (2017); Schaffer et al. (2021); Menchetti et al. (2021) for details.

[10] See Bernal et al. (2017) for more details on this method. Schaffer et al. (2021) describes segmented regression as the simplest interrupted time series analysis.

[11] See Huntington-Klein (2021).

[12] See Menchetti et al. (2021) for details on this approach.

[13] See Schaffer et al. (2021); Huntington-Klein (2021) for more details on this method.

[14] This follows Schaffer et al. (2021).

[15] There are automated approaches to finding the best fitting ARIMA model. See Hyndman and Khandakar (2008) and the `auto.arima` function from the `forecast` package in R (Hyndman et al., 2022). Residual analysis can also be used to ensure autocorrelation and other issues have been eliminated. There is also potential for time varying confounders to be included in the analysis.

This approach does not require us to specify the shape of the intervention's effect (e.g., a slope change or level change or something more complicated). It also allows us to ground the analysis in the potential outcomes framework (a common, popular way to think about causal effects), elucidate the causal assumptions required to ascribe causality to any effect estimate, and clarify the particular causal estimands we might target. The formal combination of ARIMA and potential outcomes[16] is a recent approach that allows us to estimate the causal effect of a single persistent population level intervention in an time series setting. This approach is facilitated by the `CausalArima` package in R.[17] Let $Y_t(d)$ be the potential outcome for a generic unit at time t.[18] For t in the post-treatment period,

- The point causal effect is $\tau_t = Y_t(1) - Y_t(0)$.
- The cumulative causal effect is $\Delta_t = \sum_{s=t^*+1}^{t} \tau_s$.
- The temporal average causal effect is $\bar{\tau}_t = \frac{1}{t-t^*}\Delta_t$.

Causal effects are then estimated as

- $\hat{\tau}_{t^*+k} = \underbrace{Y_{t^*+k}(1)}_{\text{observed}} - \underbrace{\hat{Y}_{t^*+k}(0)}_{\text{forecast}}$
- $\hat{\Delta}_{t^*+k} = \sum_{h=1}^{k} \hat{\tau}_{t^*+h}$
- $\hat{\bar{\tau}}_{t^*+k} = \frac{1}{k}\hat{\Delta}_{t^*+k}$

It is possible that there are unobserved confounding variables that account for the change between periods but "looks like" a treatment effect. For example, it is possible that the impact of the service change on 9/19/2020 had not fully settled before 10/1/2020. If this is the case, what appears to be an effect of treatment could actually be due to lingering effects of the service change that coincide with the treatment timing. To ameliorate this particular issue, we also do our analyses limited to routes that did not have service changes. Similarly, it is possible that the weather in the pre-treatment period and post-treatment period were not the same and that the weather had an effect on ridership. A simple analysis of this setting might confound the effect of fare reinstatement with that of the weather. To address this, we run versions of our analyses that include precipitation as a variable in our models. We hope that the covariates we include and other analysis choices get us close to unconfoundedness. Our understanding is that the decision to reinstate fares on 10/1/2020, rather than some other date in September or October (other than 9/1/2020) was made for accounting purposes that are not likely to be related to other factors in the system. But other events that might have taken place around 10/1/2020 and unobserved variables that effect ridership could confound treatment effects.

See Menchetti et al. (2021) for details of additional assumptions (beyond the "no unobderved confounding" assumption above) required for attributing causality to these estimates. We believe most of these are plausible in the present setting. However, assumptions like non-anticipating potential outcomes and no temporal interference could also plausibly be violated. It is possible that riders could have changed their riding behavior in anticipation of the publicized reinstatement of fares. In our setting, units are routes or stops that do not appeal to the same riders, so we can reasonably assume that there is no temporal interference. Though this could be violated for stops that are very close or routes that cover very similar areas. But this is not likely to be common. Any violations of these assumptions threaten the validity of boardings in the post-treatment period that are forecasted using the data in the pre-period. We do not make strong claims about the validity of these assumptions, but believe that they are plausible.

We provide several versions of these analyses. We run these analyses for routes without 9/19/2020 service changes as well as for all routes (in which case, we limit to the period after 9/19/2020 for this group). We also run them for different strata of estimates of the income level of routes' service areas. Finally, for the interrupted time series approaches, we run them with and without time-varying covariates.

5 Findings

In this section, we review our results. Overall, we find evidence of a reduction in ridership after the reinstatement of fares.

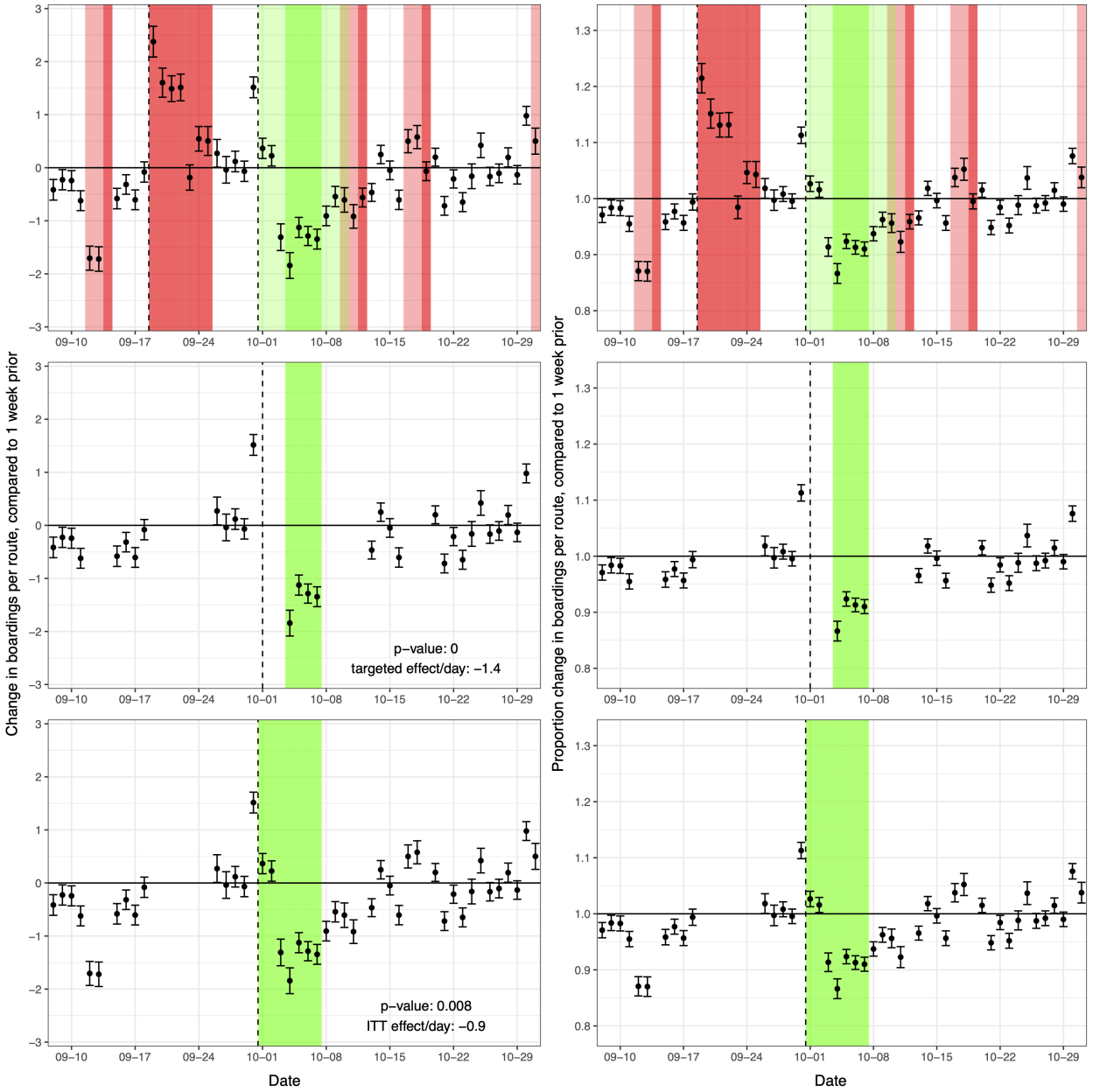
[16] See Menchetti et al. (2021) for details on this approach.

[17] See Cipollini et al. (2022); Menchetti et al. (2021) for details. The `CausalArima` package allows for automated approaches to finding the best fitting ARIMA model, uising the `auto.arima` function from the `forecast` package in R. See Hyndman and Khandakar (2008); Hyndman et al. (2022).

[18] Potential outcomes are the outcomes we would have observed under different, possibly counterfactual, treatments. See Menchetti et al. (2021) for more on potential outcomes.

Figure 7:

Matched trip boarding shifts, all routes.



Note: Left column shows absolute changes, right column shows relative changes (1.0 means no change). First row shows all week-to-week shifts (red represents lower reliability of data, green represent treatment reliability, as described in the text). Second row shows pure treatment subset. Third row shows ITT subset. P-values generated using permutation inference, as described in the text.

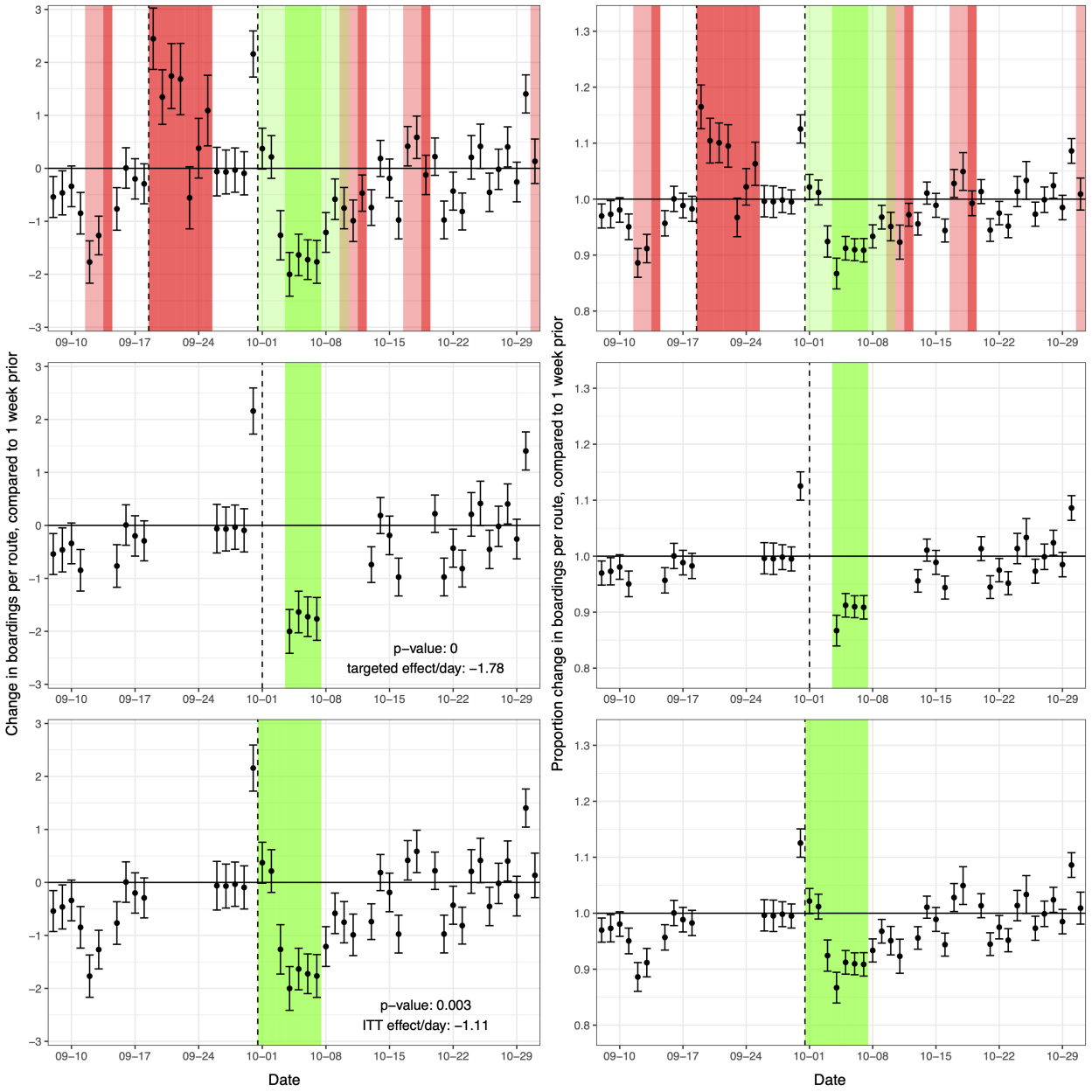
5.1 Week-to-Week Trip Comparisons

We start by looking at the matched comparisons between pre and post treatment outcomes, recalling that the ITT analysis targets what is actually under-estimate of the actual treatment effect, and the pure treatment analysis represents a targeted effect estimate that is somewhat vulnerable to day-of-week biases. In our ITT analysis for all routes, we detected an average of 0.9 fewer boardings per trip each day, with a permutation inference p-value of 0.008 (Figure 7). For the pure treatment analysis, we detected an average of 1.4 fewer boardings per trip each day, with a p-value of 0.000. Given baseline boarding counts of between 13.8 and 15.2 per trip, these shifts represent around a 10% drop in

boardings that we attribute to the reintroduction of fares among our matched subset of trips. We note that the week ending 9/30/2020 shows a noticeable increase in boardings. This is in part due to a particularly low number of boardings the week prior, but we could not identify any particular source of decreased ridership on that day, increased ridership on 9/30/2020, or subsets of routes that had particularly large impacts on those decreases and increases.

Figure 8:

Matched trip boarding shifts, routes in the lower third of the median income distribution

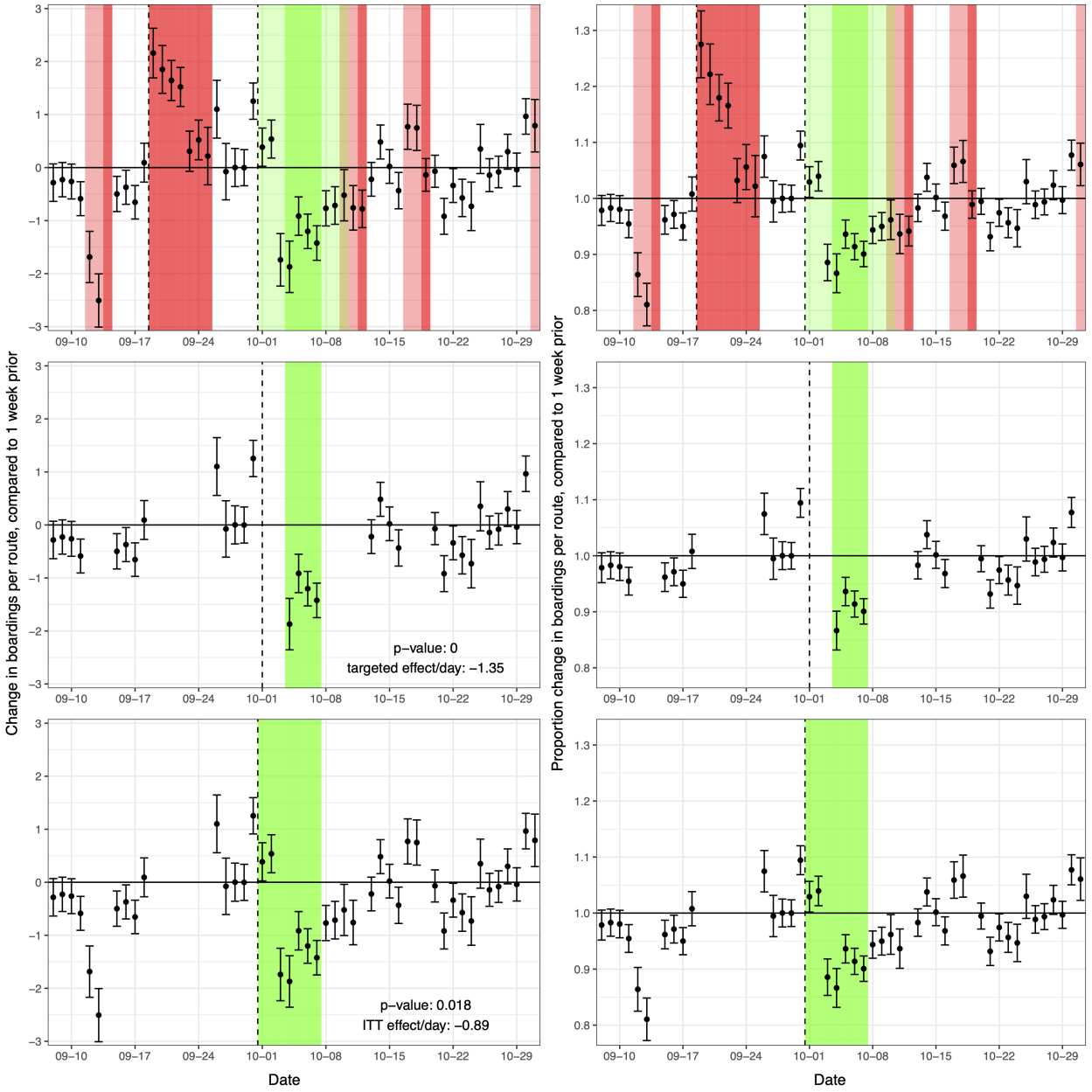


Note: Left column shows absolute changes, right column shows relative changes (1.0 means no change). First row shows all week-to-week shifts (red represents lower reliability of data, green represent treatment reliability, as described in the text). Second row shows pure treatment subset. Third row shows ITT subset. P-values generated using permutation inference, as described in the text.

We then disaggregated our analysis to detect any desperate impacts of fare reintroduction for riderships of differing income levels. The matched analysis is repeated for each third of the routes based on weighted income across their stops (Figures 8, 9, and 10). All three subsets feature statistically significant decreases in boardings at the α=0.05 level in both the ITT and pure treatment analysis. The lower income routes feature estimated magnitudes of those

Figure 9:

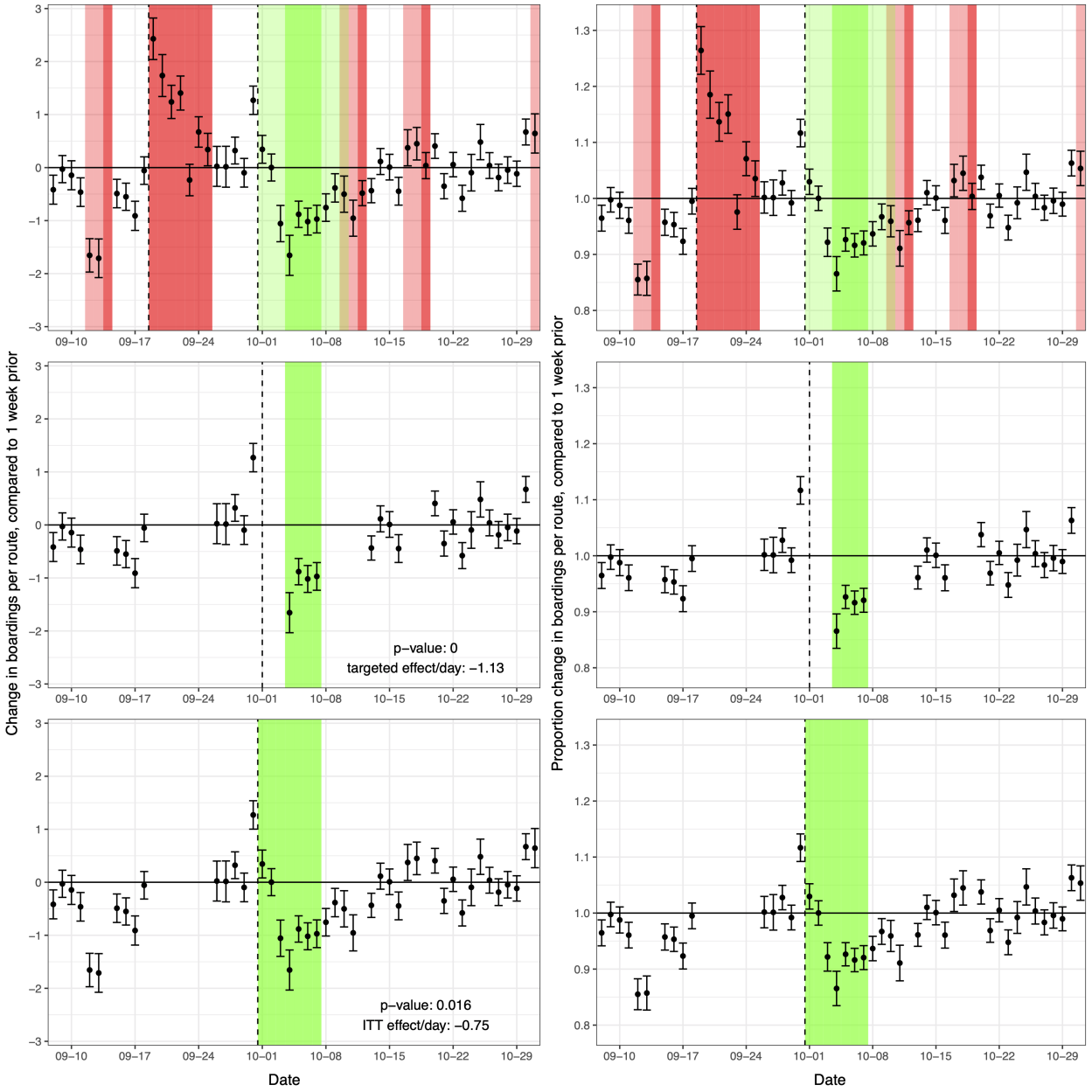
Matched trip boarding shifts, routes in the middle third of the median income distribution



Note: Left column shows absolute changes, right column shows relative changes (1.0 means no change). First row shows all week-to-week shifts (red represents lower reliability of data, green represent treatment reliability, as described in the text). Second row shows pure treatment subset. Third row shows ITT subset. P-values generated using permutation inference, as described in the text.

Figure 10:

Matched trip boarding shifts, routes in the higher third of the median income distribution



Note: Left column shows absolute changes, right column shows relative changes (1.0 means no change). First row shows all week-to-week shifts (red represents lower reliability of data, green represent treatment reliability, as described in the text). Second row shows pure treatment subset. Third row shows ITT subset. P-values generated using permutation inference, as described in the text.

decreases that are larger than in the medium and high income routes, though this analysis does not allow for detection of the statistical significance of that difference, and the percent-based decreases in ridership for all three subsets hover similarly around 90%, just like the complete route-set analysis

These results are limited in that they focus on a subset of the data that may not be generalizable, do not account for weather and covid data, and have somewhat unclear threats to validity given the analysis' construction for this setting (rather than using a well-established method). Overall, these results suggest a meaningful decrease in ridership of somewhere around 10%.

5.2 Interrupted Time Series Approaches

We now consider the time-series approaches that incorporate information about over-time changes. We first consider all full service routes, across all route income strata, where we include precipitation and COVID-19 cases as covariates. From Table 1, we see that, in this case, we are not able to tell whether either the level or slope change after fares are reinstated is statistically different from zero. Further, the level change estimate is positive while the slope change estimate is negative. We can see the regression's fit to the data in Figure 11. The blue line indicates what we predict the boardings to be, absent the fare reinstatment. We see that the blue line has a very different slope in the post-treatment period.

In Table 2, we look at all the versions of the segmented regression analysis. We see that there is inconclusive evidence about the *level change* across versions, but can see that higher income levels tend to have negative level change estimates. Looking at all routes, we see that these are statistically significant. However, when we look at *slope change* estimates, we see exclusively negative estimates and that many of these are statistically significant. In this analysis, we do not see a clear difference across income levels. However, this comparison is difficult since the coefficients in this analysis are in terms of changes in levels or slopes for the trend in overall boardings, which are not directly comparable across route sets (e.g., there are more boardings for all routes than for full service routes).

Next, we consider the comparisons of post-treatment observables to forecasts. We again start by looking at the results for all full service routes, across all route income strata, where we include precipitation and COVID-19 cases as covariates. See Table 3. We focus on the temporal average causal effects and relative effects (observed / forecasted - 1). In this instance, we estimate a 5.6% decrease in boardings on average across all post-treatment time periods. We also see that the 97.5th percentile of bootstrapped estimates is negative. That is, this estimate is statistically different from zero. We also see that the ARIMA model estimated in the pre-treatment period has three autoregressive terms and one seasonal autoregressive term with a frequency of 7, as we might expect given we know there is a weekly cycle for ridership. We can see the ARIMA fit and the forecast of the untreated outcomes in the post-treatment period in Figure 12.

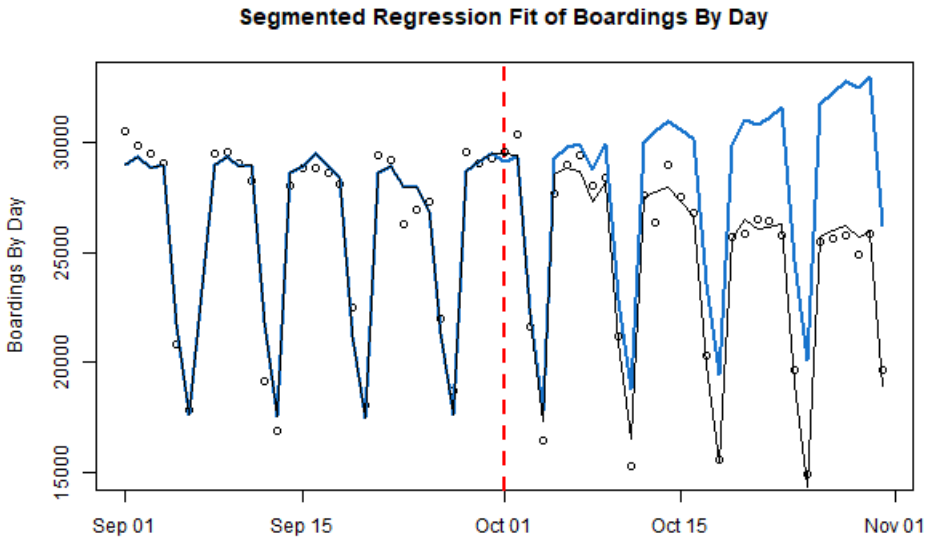
Table 1: Segmented Regression Results - Full Service Routes, All Incomes, Including Covariates - Observations: 60; F(11,48) = 146.132, p = 0.000; R^2 = 0.971; Adj. R^2 = 0.964 ; Standard errors: Robust, type = HC3

Variable	Coefficient Estimate	2.5%	97.5%	t-Value	p-Value
Intercept	-19394.860	-103069.083	64279.362	-0.466	0.643
Time (centered)	-144.610	-487.051	197.831	-0.849	0.400
Post	323.878	-667.005	1314.761	0.657	0.514
Time (centered) × Post	-253.101	-510.024	3.822	-1.981	0.053
Precipitation	-1472.188	-2750.401	-193.975	-2.316	0.025
COVID-19 Cases	1.646	-2.069	5.360	0.891	0.377
DOW = 2	11085.489	10173.020	11997.959	24.427	0.000
DOW = 3	11533.895	10591.779	12476.012	24.615	0.000
DOW = 4	11861.558	10940.136	12782.981	25.883	0.000
DOW = 5	11368.733	10544.480	12192.987	27.732	0.000
DOW = 6	11365.837	10536.185	12195.488	27.545	0.000
DOW = 7	4210.282	3053.854	5366.710	7.320	0.000

Table 4 contains results for all versions of this analysis, and we see that all temporal average causal effect estimates are negative and most are statistically significant. Since these estimates are in percentage change terms, we can compare them more easily across income levels. However, we don't see a clear pattern of larger or smaller estimates across income levels.

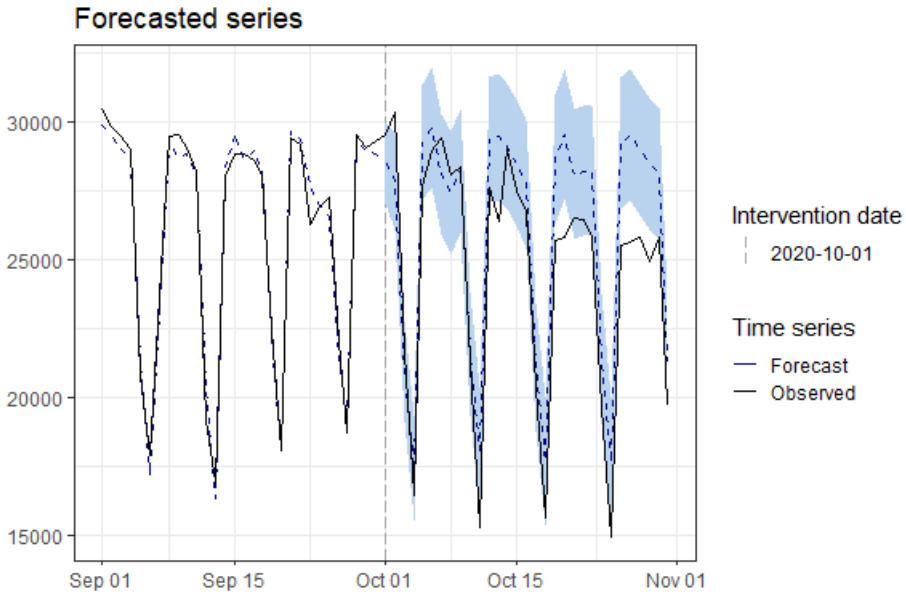
While these results suggest that fare reinstatement lead to a decrease in ridership, we do not take a strong position on the magnitude of this decrease. Further, our results are somewhat inconclusive about whether routes with lower income

Figure 11: Segmented Regression Plot - Full Service Routes, All Incomes, Including Covariates



Note: This plot shows the fit of the segmented regression model for full service routes across all income strata that includes covariates. The black line is the model fit to the data in the pre- and post-treatment periods. The blue line is the same model but assuming that the fit in the pre-treatment period continued into the post-treatment period.

Figure 12: Observed vs Forecasted Plot - Full Service Routes, All Incomes, Including Covariates



Note: This plot shows the ARIMA model fit in the pre-treatment period forecast into the post-treatment period (see the dahsed line) for full service routes across all income strata that includes covariates. The solid line is the observed boardings over time.

service areas had less or more of a decrease relative to higher income service areas. Thus, we are fairly confident that there was a negative effect of fare reinstatement on ridership, but we are much less confident about about the magnitude

Table 2: Segmented Regression Results - Standard errors: Robust, type = HC3

Level Change					
Version	Level Change Estimate	2.5%	97.5%	t-Value	p-Value
Full Service | All Incomes | Covariates	323.878	-667.005	1314.761	0.657	0.514
Full Service | All Incomes | No Covariates	685.310	-540.908	1911.529	1.123	0.267
Full Service | Low Income | Covariates	619.038	-207.191	1445.266	1.506	0.139
Full Service | Low Income | No Covariates	969.662	-55.943	1995.267	1.899	0.063
Full Service | Med. Income | Covariates	-153.988	-411.603	103.626	-1.202	0.235
Full Service | Med. Income | No Covariates	-175.155	-401.624	51.314	-1.553	0.127
Full Service | High Income | Covariates	-141.171	-375.835	93.492	-1.210	0.232
Full Service | High Income | No Covariates	-109.197	-353.999	135.606	-0.896	0.375
All Routes | All Incomes | Covariates	-4432.232	-9709.982	845.518	-1.713	0.097
All Routes | All Incomes | No Covariates	-4668.318	-10309.066	972.431	-1.684	0.102
All Routes | Low Income | Covariates	360.559	-3745.973	4467.091	0.179	0.859
All Routes | Low Income | No Covariates	168.875	-3638.207	3975.957	0.090	0.929
All Routes | Med. Income | Covariates	-2836.913	-4273.851	-1399.975	-4.027	0.000
All Routes | Med. Income | No Covariates	-2797.269	-4596.468	-998.070	-3.163	0.003
All Routes | High Income | Covariates	-1955.878	-3132.938	-778.818	-3.389	0.002
All Routes | High Income | No Covariates	-2039.924	-3388.149	-691.699	-3.078	0.004
Slope Change					
Version	Slope Change Estimate	2.5%	97.5%	t-Value	p-Value
Full Service | All Incomes | Covariates	-253.101	-510.024	3.822	-1.981	0.053
Full Service | All Incomes | No Covariates	-110.241	-174.305	-46.177	-3.456	0.001
Full Service | Low Income | Covariates	-154.273	-335.778	27.232	-1.709	0.094
Full Service | Low Income | No Covariates	-30.519	-85.316	24.278	-1.119	0.269
Full Service | Med. Income | Covariates	-58.058	-111.238	-4.879	-2.195	0.033
Full Service | Med. Income | No Covariates	-51.319	-66.329	-36.309	-6.867	0.000
Full Service | High Income | Covariates	-40.770	-107.570	26.031	-1.227	0.226
Full Service | High Income | No Covariates	-28.404	-43.064	-13.744	-3.892	0.000
All Routes | All Incomes | Covariates	-1462.528	-2239.865	-685.191	-3.837	0.001
All Routes | All Incomes | No Covariates	-941.761	-1694.945	-188.577	-2.544	0.016
All Routes | Low Income | Covariates	-950.626	-1486.584	-414.668	-3.617	0.001
All Routes | Low Income | No Covariates	-604.390	-1017.431	-191.348	-2.977	0.005
All Routes | Med. Income | Covariates	-185.085	-443.995	73.825	-1.458	0.155
All Routes | Med. Income | No Covariates	-167.767	-419.193	83.660	-1.358	0.184
All Routes | High Income | Covariates	-326.818	-556.589	-97.046	-2.901	0.007
All Routes | High Income | No Covariates	-169.605	-382.037	42.828	-1.624	0.114

of the effect and the differential impact across socio-economic strata. We believe that this effect is likely causal, while recognizing there is potential for unobserved confounding to bias our conclusions.

6 Recommendations

Results from each of the above approaches suggest that fare reinstatement led to a decrease in ridership. The varying methods, however, produce somewhat varying estimates of the effect magnitude, ranging from less than a 5% decrease to as much as a 20% decrease in ridership. Some evidence suggests that the effect was immediate (ridership dropped within a few days), and does not seem to have rebounded in the month following the fare reintroduction (and potentially decreased further).

While some analyses show signs that routes serving lower-income neighborhoods experienced a larger decrease in ridership due to fare reinstatement, such claims are not defensible with confidence and do not show up across analyses. We attribute our lack of confident conclusions for heterogeneity by socioeconomic strata to the limitations of the data and assumptions involved, not as evidence for a lack of heterogeneity.

Based on the analyses presented, we want to emphasize two sets of recommendations: the first are related to the fare reinstatement policy, while the second are related to the design of impact evaluations of future policies.

Table 3: Observed vs Forecasted Results - Full Service Routes, All Incomes, Including Covariates

Temporal Average Causal Effect - Bootstrapped Standard Errors

	Estimates	2.5%	97.5%	SD
Observed	24583.742	NA	NA	NA
Forecasted	26042.792	25871.096	26200.323	85.272
Absolute Effect (Observed - Forecasted)	-1459.050	-1616.581	-1287.354	85.272
Relative Effect ([Observed / Forecasted] - 1)	-0.056	-0.062	-0.049	0.003

Cumulative Causal Effect - Bootstrapped Standard Errors

	Estimates	2.5%	97.5%	SD
Observed	762096.000	NA	NA	NA
Forecasted	807326.546	802003.984	812210.014	2643.432
Absolute Effect (Observed - Forecasted)	-45230.546	-50114.014	-39907.984	2643.432
Relative Effect ([Observed / Forecasted] - 1)	-0.056	-0.062	-0.049	0.003

ARIMA Model

	p	d	q	P	D	Q	Frequency
ARIMA Order	3	0	0	1	0	0	7

	Coefficient	SE	t-Value
Intercept	17642.6794055	4073.8131174	4.3307533
Precipitation	-988.4349384	565.2178505	-1.7487681
COVID-19 Cases	0.0062531	0.1938122	0.0322637
AR1	0.6828268	0.1781178	3.8335677
AR2	-0.1575200	0.2321681	-0.6784739
AR3	-0.4377519	0.2053643	-2.1315871
SAR1	-0.6202931	0.1640536	-3.7810386
DOW = 2	11418.0534469	341.2683398	33.4576992
DOW = 3	11739.9638302	478.0879969	24.5560732
DOW = 4	11043.2733109	603.0024898	18.3138105
DOW = 5	10641.4214327	655.8212280	16.2261009
DOW = 6	10379.5445990	613.0516579	16.9309461
DOW = 7	3526.2733767	360.9851648	9.7684717

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training Set	60.49124	553.428	452.7084	0.1913077	1.855145	0.2241333	-0.0113572

	loglik	aic	bic	aicc
Metrics	-226.4612	480.9224	500.0646	510.9224

Table 4: Observed vs Forecasted Results - Temporal Average Causal Effect - Relative Effect ([Observed / Forecasted] - 1) - Bootstrapped Standard Errors

Version	Temporal ACE Estimate	2.5%	97.5%	SD
Full Service | All Incomes | Covariates	-0.056	-0.062	-0.049	0.003
Full Service | All Incomes | No Covariates	-0.054	-0.060	-0.046	0.004
Full Service | Low Incomes | Covariates	-0.002	-0.007	0.004	0.003
Full Service | Low Incomes | No Covariates	-0.013	-0.049	0.022	0.018
Full Service | Med. Incomes | Covariates	-0.337	-0.369	-0.307	0.016
Full Service | Med. Incomes | No Covariates	-0.266	-0.413	-0.099	0.080
Full Service | High Incomes | Covariates	-0.165	-0.193	-0.135	0.015
Full Service | High Incomes | No Covariates	-0.062	-0.112	-0.014	0.025
All Routes | All Incomes | Covariates	-0.190	-0.190	-0.190	0.000
All Routes | All Incomes | No Covariates	-0.201	-0.245	-0.153	0.023
All Routes | Low Incomes | Covariates	-0.269	-0.269	-0.268	0.000
All Routes | Low Incomes | No Covariates	-0.203	-0.203	-0.203	0.000
All Routes | Med. Incomes | Covariates	-0.123	-0.124	-0.123	0.000
All Routes | Med. Incomes | No Covariates	-0.092	-0.109	-0.078	0.008
All Routes | High Incomes | Covariates	-0.114	-0.114	-0.113	0.000
All Routes | High Incomes | No Covariates	-0.078	-0.090	-0.067	0.006

6.1 Policy Recommendations

We emphasize first that our findings should be regarded as tentative and "the best guess we can muster", rather than as definitive. Any causal claims rest on assumptions that, while plausible here, are not certain. In addition, the variation in effect magnitudes across methods, while not unreasonable, suggests skepticism.

Although certain methods showed larger decreases in ridership among trips serving lower income locations, we cannot be certain that distinction was causally linked to income differences themselves. Given our expectations for differential impacts of fare reinstatement, we attribute our lack of confident conclusions to the weakness of the available data for causal analysis rather than the possibility of uniform impact across income strata.

With this skepticism in mind, the best estimates we can produce are supportive of policymakers' prior beliefs about the impacts of fare reinstatement: reintroducing fares lowered ridership a moderate amount, and that impact may have been larger among lower income ridership. Our policy recommendations given available information are:

1. Transit system and local government authorities should **expect a noticeable drop in ridership** in response to movement from free transit to paid transit
2. Services located in **lower income neighborhoods may be more likely to experience such impacts** to a greater degree
3. These **results are limited in their generalizability**: We cannot use this data to assess any impacts of the fare reinstatement beyond the first month, cannot predict the impact of different payment policies, and cannot confidently transfer learnings outside of the unique, early pandemic context of this policy change

The transit agency faces a trade-off between serving the community through increasing ridership, and increasing revenue coming from ridership. Free ridership during the high of the COVID pandemic most likely served the initial purpose of preventing transmission by avoiding contact between drivers and users during boarding, while also serving the community by helping the people most economically impacted by the pandemic. While the first purpose is probably outdated (given our increased understanding of fomite transmission), the second is still relevant for current decisions. Therefore, we believe policymakers should view these analyses as supportive of efforts to create more accessible and affordable transit options, without expecting such programs to deliver easily predictable results. Such efforts may take the form of strengthening support for existing financial aid programs, but may also drive increased focus on identifying populations who are more likely to change their ridership decisions due to fare considerations (more on this below).

6.2 Design Recommendations

Based on our experience analyzing these data, we have several design recommendation that we can divide in two groups: those regarding policy roll-outs, and those regarding measurement.

First, with respect to policy roll-outs, we recommend the following to aid future studies:

1. **Comparison group:** A first improvement to the design of future policy assessments would be to consider withholding the intervention for a subset of the affected users/routes when possible. This would need to occur for at least long enough to observe the expected impacts of the change. This provides a cleaner comparison and it makes easier to attribute impact to the changes under study.
2. **Staggered adoption:** A variation on the prior point that would provide further information for the future analysts would be to roll out the policy change in a staggered fashion, such that some groups enter to the new policy before than others. This facilitates the analysis of potential time dynamics and the cumulative effect of the change being introduced.
3. **Over-time data:** The more data made available to analysts, the more reliable our estimates and the more concrete our recommendations can be. This is particularly important for data over time, as a month of data before and after the policy change was not sufficient to be confident in our models and conclusions, especially in establishing the lasting impacts of the policy change on rider behavior. Having a month or two of additional pre- and/or post-policy change data would have been helpful, as would data from the same months in previous years.
4. **Isolating policy changes in time:** A central difficulty in the interpretation of the data was the presence of a major service change that occurred less than 2 weeks away from the fare reintroduction. When at all possible, policies and changes that will require an analysis should be implemented after a longer period of time during which riders can settle into a stable behavior equilibrium.

Second, with respect to the measurement of variables of interest, we have the following recommendations:

6. **Consistent Measurement:** Ideally, the central data collection measurements should be consistent across the transit system. Acknowledging that there were not enough buses with Automatic Passenger Counters

(APCs), however, care should be taken to establish any inconsistencies in such measurement. For example, we were informed simply that APC busses were assigned to trips without intentionally directing them to certain routes, and while more detail about that assignment process may not seem important, it can go a long way in mitigating the inconsistent measurements in the APC-based data. Related to both this point and to the point above about isolated policy timing, we note that a nearly concurrent change in trip frequencies brought changes to what proportion of trips generated measurements, which added inconsistency across the generated data as well.

7. **Differential impacts:** With a focus on fairness or equity assessments of future proposed and implemented policy changes, it would be advisable to make a consistent, localized effort in defining, identifying, and measuring different segments of the population, especially those that may be at higher risk of harm due to a particular policy change. In our example, we attempted to run analyses within different socioeconomic strata, but used very rough proxies drawn from census data, given the lack of better measurements.
8. **Domain experts and local knowledge:** Related to the above, we imagine that analysts more familiar with the King County Metro and the geography and neighborhoods of King County would have been better able to design and interpret relevant analyses. While our understanding of the causal designs available to us was critical, we imagine a lack of local domain knowledge limited our ability to provide the best analysis possible.

References

Bernal, J. L., Cummins, S., and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1):348–355.

Campbell, D. T. (1957). Factors relevant to validity of experiments in social settings. *Psychological Bulletin*, 54(4):297–312.

Cipollini, F., Menchetti, F., and Palmieri, E. (2022). *CausalArima: Causal effect of an intervention using ARIMA models*. R package version 0.1.0.

CISIL and KCMTD (2022). Causal Inference for Social Impact Lab Data Challenge and King County Metro Transit Department.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685.

Huntington-Klein, N. (2021). *The Effect: An Introduction to Research Design and Causality*. Chapman and Hall/CRC.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2022). *forecast: Forecasting functions for time series and linear models*. R package version 8.16.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

King County (2020). Watersheds and Rivers: Annual Summary Data: Daily accumulated precipitation and Daily mean barometric pressure. `https://green2.kingcounty.gov/hydrology/SummaryDataTables.aspx?G_ID=743&Parameter=Precipitation`.

Menchetti, F., Cipollini, F., and Mealli, F. (2021). Estimating the causal effect of an intervention in a time series setting: the c-arima approach.

Schaffer, A., Dobbins, T., and Pearson, S. (2021). Interrupted time series analysis using autoregressive integrated moving average (arima) models: a guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*, 21(58).

Switzer, J. (2020). Sept. 19 service change: Metro helping to prepare riders. `https://kingcountymetro.blog/2020/09/02/sept-19-service-change-metro-helping-to-prepare-riders/`.

The New York Times (2021). Coronavirus (Covid-19) Data in the United States. `https://github.com/nytimes/covid-19-data`.

U.S. Census Bureau (2020). 2020 American Community Survey 5-year Detailed Tables [Variable: B06011_001E; CSV file]. `https://www.census.gov/data/developers/data-sets/acs-5year.html`.