HARVARD Kennedy School
MOSSAVAR-RAHMANI CENTER
for Business and Government

Riding with Charlie: Public Transportation Policy and Its Impact on Business and Road Safety in Massachusetts

Honorable Mention, 2020 Dunlop Undergraduate Thesis Prize

Pedro Farias

May 2020

M-RCBG Associate Working Paper Series | No. 143

Mossavar-Rahmani Center for Business & Government
Weil Hall | Harvard Kennedy School | www.hks.harvard.edu/mrcbg

Riding with Charlie:
Public transportation policy and its impact on businesses and road safety in Massachusetts

Pedro Luis Cunha Farias

Presented to the Department of Applied Mathematics
in partial fulfillment of the requirements for a Bachelor of Arts degree with Honors.

Harvard College
Cambridge, Massachusetts

April 2, 2020

Abstract

Can public transportation help local businesses and save lives and costs? This paper considers the impact of the MBTA's late night service extension on local businesses and car crashes and provides two main contributions. The first is evidence from a differences-in-differences design that businesses may benefit from late night service, as the number of Yelp reviews for businesses near stops with late night service increased between 2.5% and 8% and the number of late night tweets in areas with the service increased 4%, with these estimates statistically significant at the 99% level. The second is that the service helps reduce the number of car crashes by 4% on average, though this is not statistically significant. Areas with a high proportion of young individuals and minorities experienced a stronger treatment effect of around 12%, with statistical significance varying based on each of the three empirical strategies employed, which include differences-in-differences, propensity score matching, and geographic proximity analysis. This indicates public transportation could be effective both to boost the performance of local businesses and save lives by reducing the incidence of car crashes.

Acknowledgments

I was lucky enough to have an extremely patient faculty advisor, thesis seminar leader, and peers who helped me navigate this process. I could not have completed this thesis without the support of Professor Edward Glaeser. His microeconomics class was the first Economics class I had ever taken, and it convinced me to study Applied Mathematics in Economics. His invaluable advice allowed me to deepen my exposure to empirical research and ensured these past months were the most rewarding of my academic career. I am deeply grateful for Zhengyun Sun's guidance, availability, and support throughout this process, and her dedication to our thesis seminar. My fellow classmates provided invaluable feedback and advice throughout the year. I would also like to thank Diane Sredl, the Harvard Institute for Quantitative Social Science, the Boston Area Research Initiative, and all the professors and peers who helped me grow during my time at Harvard College.

Contents

1 Introduction **5**

2 Institutional Context **9**

2.1 MBTA and Late Night Service . . . 9

2.2 Yelp and Twitter . . . 11

2.3 Massachusetts Car Traffic and Crashes . . . 12

2.4 The Road Safety Challenge . . . 12

3 Analytical Framework **14**

4 Data **17**

5 Empirical strategy **18**

5.1 Impact on Local Businesses . . . 19

5.2 Crash Analysis: Complete Dataset . . . 21

5.3 Crash Analysis: Young and Minority Regions . . . 23

6 Results **25**

6.1 Impact on Local Businesses . . . 25

6.2 Crash Analysis: Complete Dataset . . . 29

6.3 Crash Analysis: Young and Minority Regions . . . 32

7 Discussion **35**

7.1 Cost-Benefit Analysis . . . 35

7.1.1 Benefits to Local Businesses . . . 35

7.1.2 Car Crash Cost Savings . . . 37

7.2 Limitations and Extensions . . . 38

8 Conclusion **40**

9 Appendix **54**

9.1 MBTA Coverage and Area of Study 54

9.2 Crash and Demographic Data 55

9.3 Standard Error Clustering 56

1 Introduction

Public transportation has the potential to provide affordable transportation options to those who need it while helping local businesses and saving lives. Nearly 37,000 lives are lost and $1 trillion dollars spent yearly in the United States because of car crashes. Extended public transportation could help save these lives and reduce the cost of crashes, and this paper empirically estimates the effect of extending public transportation service. From March 28, 2014 to March 20, 2016 the Massachusetts Bay Transportation Authority (MBTA) instituted late night service in some of its stops in an attempt to boost the Greater Boston Area's economy and provide affordable and safe transportation options for commuters (Burton, 2013). This imposed a yearly operating cost of $14 million, contributing to the MBTA's deficit of $240 million (Kassiel, 2018). Considering the service's high cost, did the MBTA's decision help the local community by improving business outcomes and saving lives by reducing the incidence of car crashes?

My analysis begins considering the impact of additional public transportation supply on local businesses. Using an OLS differences-in-differences design, I estimate the average treatment effect on the treated (ATT) of late night service on the number of reviews businesses receive on Yelp — a crowd sourced review website — with reviews serving as a proxy for visits to these locations, and find an increase between 0.063 and 0.199 standard deviations, or between 2.5% and 8% in the number of Yelp reviews for businesses near late night stops. I also provide non-parametric evidence for an increase in the number of Yelp reviews by plotting the number of reviews across the treatment and control groups in Figure 2, including linear time trends in Table 2, measuring the treatment effect over time in Figure 3, and computing the difference between treatment and control groups in Figure 4. Figures 2 and 4 also strengthen the assumption of pre-trends, as they indicate the two groups evolved similarly prior to treatment and that the difference in Yelp reviews between them increased during the treatment.

A second analysis of business outcomes using Twitter data indicates there is also an increase in the number of late night tweets between 0.111 and 0.123 standard deviations, or about 4%, in the treated areas caused by late night service, as reported in Table 3. The number of late night tweets helps measure the corresponding activity in these areas and serves as a proxy for potential customers. The plot of tweets over time in Figure 5 indicates a slight increase in the number of tweets during the treatment, and the plot of the treatment effect over time in Figure 6 is consistent with a statistically significant effect, while Figure 7 indicates there is an increase in the difference of tweet counts between the two groups. These figures also support the parallel trends assumption — Figure 5 indicates the two groups evolved similarly, and this is supported by the small difference between the two groups shown in Figure 7 prior to treatment.

After the analysis on business outcomes, I consider the impact of late night service on reducing the number of car crashes, which saves both lives and costs. The analysis begins with an analytical model to consider the individual's decision to go out and the medium of transportation they choose to use. The paper then moves to an empirical analysis to measure the impact of late night service on car crashes. The first part of the empirical analysis estimates the average treatment effect on the treated for the entire Greater Boston Area. The pre-trend tests include a plot of the crashes and a plot of the treatment effect over time and indicate the parallel trends assumption required for a differences-in-differences estimate to be consistent is valid. The analysis estimates a -0.113 standard deviation or -4% reduction of weekend late night car crashes due to late night service, as reported in Table 4. Although small in magnitude, this effect represents a cost-reduction of approximately \$3 million. However, the standard errors on this point estimate are large and the treatment effect appears to be heterogeneous, varying widely based on the demographics of each region.

The inclusion of demographic characteristics indicates the treatment effect is heterogeneous based on demographics, and allows for an analysis on the impact of age, income, and racial and ethnic composition on the treatment effect — providing an ATT based on geography and demography. The result for each of these additional estimates is dependent on the demographic characteristic considered. The treatment effects are much stronger in areas with a high concentration of young individuals, low income households, and minorities, and these areas of higher ATTs are more closely studied in the second part of the analysis. I also use Twitter data to identify which areas of the region studied had the most activity in the weekend late nights, and find the treatment effect in these regions is three times the regions without late night Twitter activity.

To ensure the pre-treatment trends assumption holds, a requirement for the differences-in-differences analysis to be consistent, the paper presents the plot of crashes over time, shown in Figure 8, which indicates pre-treatment trends were parallel. A plot of the treatment effect over time, presented in Figure 9, is consistent with the regressions indicating there is a treatment effect.

The heterogeneity in the treatment effect leads to a second analyses, containing three different designs and employing a similar strategy as Chen et al. (2019), as presented in Table 5. These analyses are carried out on the regions with the highest concentrations of young populations and minority populations, as these groups have stronger non-parametric evidence of a relevant treatment effect relative to the low-income groups; the former have clearer observable decreases in crashes during treatment and increases in crashes after the treatment concludes, as seen in Figure 10. The first design employed is an OLS differences-in-differences. This strategy indicates a -0.320 or -12.5% reduction in the regions with the most young individuals and a -0.060 or -2.5% reduction in the regions with the most minorities, yet neither of these results are statistically significant. To include covariates in a non-parametric manner, I employ propensity-score weighting methods as a second analysis. The propensity score analysis indicates a -0.410 or -16% decrease in the regions with the youngest individuals and a -0.397 or -15.5% reduction in the regions with the most minorities; both of these values are statistically significant at least at the 90% level. Finally, in the third design I compare the treated areas with their closest non-treated areas, conducting a geographic proximity analysis. This analysis indicates a statistically insignificant reduction of -0.224 or 9% in the regions with the youngest individuals and a reduction of -0.457 or 18% in the regions with the most minorities, statistically significant at the 90% level.

The paper concludes with a cost-benefit analysis comparing the cost of late night service with the reduction in costs caused by the lower number of car crashes and outlining methods to calculate the dollar benefit to local businesses. The reduction of car crashes saves an estimated $3 million per year, less than a quarter of the $14 million in yearly operating costs for the service. However, a late night service design leveraging the heterogeneity of the treatment effect that targets areas with the highest measured treatment effects can reach the 17% reduction threshold needed for the service cost to be offset by the cost savings of fewer crashes.

This work does not imply the MBTA's late night service was a mistake nor that it did not provide benefits to the regions covered by the service. The service provided transportation options for those who may not have had low-cost transportation and does seem to provide benefits to local businesses. The low cost savings relative to the operating cost of the service do suggest the limits of late night service at such a scale, yet the heterogeneity provides an opportunity to consider service designs with potential for higher proportional cost offsets. Therefore, although the original late night service design did not offset its operating costs, there does seem to exist opportunity for a new design which leverages the heterogeneous treatment effect.

This paper proceeds as follows: Section 2 provides institutional context, with an overview of the MBTA, Twitter and Yelp, Massachusetts car transit, and academic literature on public transportation. Section 3 proposes an analytical framework with which to consider the consumer decision to go out and the preferred medium of transportation. Section 4 discusses the data sources and Section 5 introduces the empirical strategies employed in the analyses. Section 6 presents the empirical results. Section 7 contextualizes these results in a cost-benefit analysis of late night service. Section 8 concludes and Section 9 provides additional information that supplements the analysis.

2 Institutional Context

This section contains relevant institutional information that informs the empirical analyses of the paper. It begins by discussing the public transportation service offered in Massachusetts, then presents a brief overview of Yelp and Twitter. It provides summary data on crashes in Massachusetts and finalizes with a review of the academic literature on the road safety challenge.

2.1 MBTA and Late Night Service

The MBTA is the public agency responsible for operating the majority of public transit in the Greater Boston Area. It operates five major terrestrial mass transit vehicles, including light rail vehicles, heavy rail trains, regional rail trains, motor buses, and electric trolleybuses. It also operates a few boat and ferry lines, as well as "The Ride," a paratransit program for pre-scheduled routes. The MBTA system averages over 1 million passengers per weekday, the majority of whom use heavy rail and light-rail lines (Dickens, 2017). A breakdown of MBTA ridership across these offerings can be found in Section 9.1. The MBTA bus system is the sixth largest by ridership in the United States and contains over 150 routes, covering a larger area than the subway system.[1] The subway system has four lines: the Blue, Green, Orange, and Red Lines.[2] Figure 1 shows the MBTA subway lines.

The MBTA is primarily funded through 16% of the state sales tax not including the meals tax. Service fares also fund a significant portion of the expenses — subway fares are $2.40 for one-way trips while bus fares range from $1.70 for local buses to $5.25 for express buses from downtown Boston to the outer Boston suburbs. There are also timed passes;

[1] The majority of bus routes are contained within 13 cities, all of which are either in Essex, Middlesex, Norfolk, or Suffolk: Arlington, Belmont, Boston, Brookline, Cambridge, Chelsea, Dedham, Everett, Milton, Newton, Somerville, Revere, and Watertown.

[2] The majority of subway routes are contained within 11 cities, all of which are either in Essex, Middlesex, Norfolk, or Suffolk: Boston, Braintree, Brookline, Cambridge, Malden, Medford, Milton, Newton, Quincy, Revere, and Somerville. Most subway lines run radially from central Boston to outward regions of the Greater Boston Area, crossing downtown Boston. All have a direct transfer connection to each other, except for the Red and Blue lines.

a monthly subway pass costs $90.00. The MBTA is overseen by a fiscal and management control board charged with bringing financial stability to the agency and devising policies such as maintaining, revoking, or expanding late night service. This board was created in 2015.

The late night service trial considered in this paper was the second attempt by the MBTA to institute such a service and ran from March 28, 2014 to March 20, 2016, when it was discontinued due to high costs. It included all subway routes and the 15 bus routes considered "key bus routes,"[3] encompassing over 70% of MBTA ridership. The goal of late night service was to provide new mobility and to reduce overcrowding. After the introduction of the service, the MBTA observed a higher concentration of late night service users were minorities and low-income individuals relative to those during regular hours — 48% of late night riders were low-income individuals vs. 41% system-wide, while 51% of late night riders were minorities vs. 48% (MBTA, 2019). Over time, late night service ridership steadily decreased, from its peak of around 16,000 per week when the service was introduced to around 13,000 by late 2015, and the Massachusetts government decided these declining rates did not justify the estimated $14 million in operating costs required to maintain the service (Kassiel, 2018).[4]

When implementing the service, policy makers do not mention or state any connection to benefiting local businesses or reducing car crashes. Additionally, the timing of the trial coincides with the MBTA receiving state funding of $20 million (Burton, 2013), so it is unlikely to have been a program with previous extensive planning for the time period in which it occurred. It is also unlikely that the introduction is correlated with an increase in car crashes, as Table 1 indicates that in the years prior to late night service, there was no significant change in the number of crashes, increasing only 2% in 2012 and 2013, less than the 4% increase in 2011. There were also no significant changes in Massachusetts or federal legislation that would have a direct impact on either public transportation usage or private

[3]The key bus routes are the 15 routes with the highest ridership.
[4]The $14 million represented approximately 6% of the MBTA's $240 million deficit.

car usage. The end of late night service was also exogenous, as it was primarily caused by the MBTA's cost reduction agenda.

2.2 Yelp and Twitter

The analysis of the impact of late night service on businesses considers data from Yelp and Twitter, and this section briefly discusses both of these websites. Yelp is a crowd-sourced review website that allows users to rate and review various businesses and provides other business information such as operating hours and business category — including restaurants, nightlife, bars, and others. The service was founded in 2004 and by 2010, the earliest period considered in this paper, around 4.5 million crowd-sourced reviews had been published, the vast majority of which were in the United States. In 2019, Yelp reported over 60 million monthly average users on its desktop website and 76.7 million on its mobile application. It also has over 192 million reviews on its website. To write a review on Yelp, an individual must create a Yelp account with their email. Individuals can then submit a review on any business, but Yelp uses a filter to identify reviews they suspect are not real. Each review is tied to a specific business and contains a review score, some review text, the date and time it was submitted by the user, and other relevant data.

Twitter is a social network that allows users to post short comments, known as "tweets." The platform was created in 2006 and experienced rapid growth; by 2012 it had over 100 million users who tweeted 340 million times a day. Now, Twitter has more than 321 million monthly active users. Any user can create a tweet, and if a user's profile is not marked as private, their tweets are publicly accessible. Each tweet contains the tweet text, the date and time it was posted by the user, the user's geographic coordinates when they posted the tweet if the user allows Twitter to access that information, and other relevant data.

2.3 Massachusetts Car Traffic and Crashes

An estimated 5 million people travel within Massachusetts each weekday; 69% do so on automobiles, while 8% do so with public transportation. Within the City of Boston alone there are an estimated 536,000 vehicles per weekday (Arthur, 2018). To limit drunk driving, Massachusetts adopted an anti-drunk driving law in 2005 known as "Melanie's law," which increased penalties for repeat drunk drivers. The state subsequently amended that law in 2012 to mend a loophole allowing certain repeat offenders to be considered first time offenders, but since then there has been no new legislation likely to impact the number of car crashes in the region.

Table 1 shows key driving statistics for Massachusetts. Vehicle miles traveled per capita is mostly constant throughout the period: staying close to 8,500 until the end of 2015, when the growth rate increases and the vehicle miles traveled per capita surpass 9,000. Fatal car crashes are relatively stable at 330 annually, but the total number of crashes has steadily increased, growing over 25% in the period from 2010 to 2017.

2.4 The Road Safety Challenge

The issue of car crashes pervades the United States — four fifths of city residents drive every day or nearly every day (Brenan, 2019), and their crashes cost the United States $836 billion yearly: $242 billion in economic costs and $594 billion in societal costs. Public revenues pay for approximately 8% of these costs — the equivalent of $156 in taxes for every household in the United States — and nearly 33,000 people were killed and 3.9 million were injured as a result of crashes in 2010 (Blincoe et al., 2015). Perhaps most alarming is that the majority of crashes are caused by three avoidable factors: drunk driving (22%), speeding (22%), and distracted driving (17%).

Most public policy literature on reducing crashes focuses on increasing the costs of careless driving, such as speeding and drunk driving. Yet the most common cause of crashes — drunk driving — remains a pervasive issue with over 1 million drivers arrested for driving under the influence in 2007 (Sourcebook, 2007). Past literature shows punitive measures such as criminal penalties and substance abuse have not substantially reduced the amount of drinking and driving (Caudill et al., 1990; Hansen, 2015; Ying et al., 2013), and discusses why punitive measures have had limited success (Ross, 1992). Economic literature also explores how other instruments such as anti-drunk driving PSAs (Niederdeppe et al., 2017) and gasoline prices (Chi et al., 2011), have led to modest reductions in fatal accidents. This body of research suggests punitive and other measures have limited impact if they do not also provide a substitute to private driving.

Even when such substitutes are provided, however, one must consider the demand for public transportation, which is modeled in the analytical portion of this paper based on each individual's decision to go out and their utility from each medium of transportation. Empirical work suggests trip time variability — which is included in this paper's model — has the greatest influence on the commuter's decision on the medium of transportation (Garcia et al., 2016), and many authors have modeled the demand for public transit in a variety of contexts to inform public policy (Ceder, 2007; Yao, 2007; Attaluri et al., 1997).

Demand for public transportation may also be evolving as private markets offer substitutes to private cars. These services, however, are unable to rival the low cost of public transportation. Although ride-hailing apps like Uber[5] and Lyft have decreased transaction costs by making it easier to hail and use taxi-like services, the financial cost of ride sharing applications is still significantly higher than the cost of public transportation.[6] Public transportation therefore serves as an important potential substitute for private cars to reduce drunk driving and car crashes, and economic literature still lacks sufficient evidence to consider the effectiveness of such policies.

Jackson and Owens (2010) contribute to this literature by considering how a gradual

[5] I contacted Uber to request data to supplement my paper's analysis, but they were unable to provide any, so it is not possible to test the effects of Uber directly.

[6] As an example, a one-way trip taking a local MBTA bus costs $1.70, while the base fare for Uber is $2.10, with an additional $0.88 per mile and $0.36 per minute. Furthermore, surge pricing can be common in the evenings, when demand for ride-hailing frequently surpasses supply; this is also when alcohol impairment among drivers involved in fatal crashes is four times higher than it is during the day (DOT, 2015).

extension of the Washington D.C. Metropolitan area's subway service impacted drunk driving and find a reduction of 14% in the number of DUIs close to bars. The Jackson and Owens (2010) paper differs from my thesis in four ways. First, the cancellation of late night service by the MTBA allows me to consider whether revoking the service led to a reversal in the car crash trend. Second, my paper considers all car crashes — which allows for a more holistic evaluation of late night service — and it discusses all public transit, as opposed to only subway transit. Third, this paper includes a cost-benefit analysis to evaluate whether the cost reduction in car crashes is comparable to the cost of late night service. Fourth, both the Jackson paper and this paper find much underlying heterogeneity, though the heterogeneity in their paper is based on the proximity of bars to subway stops while this paper's is based on regional demographics.

3 Analytical Framework

This section provides an analytical framework to contextualize the empirical analyses. Specifically, this model considers the individual's decision to go out at night, similar to the model proposed in Jackson and Owens (2010). It expands on that paper's model by further addressing the consumer's decision on the medium of transportation for going out.

Individual i derives utility from going out G_i, with price c_{gi}, and from consuming a numeraire good ϕ_i, with price normalized to 1. The utility of going out increases as others in the population I also choose to go out, such that $\sigma_i = \sum_{j \neq i}^{I} G_j$, and the utility for going out is $g(G_i, \sigma_i)$, where g is increasing in G_i and σ_i. The individual receives wage w_i and therefore maximizes:

$$\max U_i = f_i(g(G_i, \sigma_i), \phi_i) \text{ s.t. } c_{gi} \cdot G_i + \phi_i \leq w_i \tag{1}$$

The first order conditions are:

$$\frac{\partial f_i}{\partial G_i} - \lambda c_{gi} = 0 \tag{2}$$

$$\frac{\partial f_i}{\partial \phi_i} - \lambda = 0 \tag{3}$$

$$w_i - c_{gi} \cdot G_i - \phi_i = 0 \rightarrow G_i = \frac{w_i - \phi_i}{c_{gi}} \tag{4}$$

From Equations 2 and 3, we derive that $\frac{\partial f_i/\partial G_i}{\partial f_i/\partial \phi_i} = c_{gi}$ so that the individual's decision to go out is determined by their utility function.[7]

The individual can choose to go out either by car or by public transportation. Driving takes time $t_{ci}(\sigma_i)$ and costs $p_{ci}(\sigma_i)$[8]; both of these are increasing in σ_i as it determines congestion. Taking public transportation takes time t_{ti} and costs p_{ti}. Assume driving and taking public transport are normal goods and substitutes. We therefore express the cost of going out as:

$$c_{gi} = \min[t_{ci}(\sigma_i) \cdot w_i + p_{ci}(\sigma_i), t_{ti} \cdot w_i + p_{ti}] \tag{5}$$

Note that if there is no public transit, then the time for taking public transit $t_{ti} = \infty$ and the individual is forced to drive. Assume in the case the two options are equal, the individual prefers to drive, and use a strict inequality to evaluate when the individual chooses public transit:

$$t_{ci}(\sigma_i) \cdot w_i + p_{ci}(\sigma_i) > t_{ti} \cdot w_i + p_{ti} \rightarrow [t_{ti} - t_{ci}(\sigma_i)] < \frac{[p_{ci}(\sigma_i) - p_{ti}]}{w_i} \tag{6}$$

Equation 6 indicates the decision depends on the travel time differential, the price differential, and the wage. In most urban centers, $p_{ti} < p_{ci}(\sigma_i)$ due to high parking costs and congestion. When this is the case, then $\frac{[p_{ci}(\sigma_i) - p_{ti}]}{w_i} > 0$. The sign of $[t_{ti} - t_{ci}(\sigma_i)]$ is undetermined and will vary by individual. We consider the time differential when $p_{ti} < p_{ci}(\sigma_i)$ holds.

[7]Note that the decision to go out is represented as a continuous decision, whereas in reality for any given day the decision is binary. This assumption in the model can be assuaged by thinking of the decision as the frequency with which the individual goes out over the course of several days.

[8]This includes the cost of parking and gas; the model is interested in individuals who would substitute car usage for public transit, so car-related expenses other than those pertaining to usage for a trip are not relevant.

If $[t_{ti} - t_{ci}(\sigma_i)] \leq 0$, that is, it takes less time to use transit than to drive — perhaps because one is close to a transit station or parking is very time consuming — then the inequality holds trivially and the individual takes public transit.[9]

If $[t_{ti} - t_{ci}(\sigma_i)] > 0$, it is unclear whether the individual opts for public transit. As the time differential $[t_{yi} - t_{ci}(\sigma_i)]$ increases, the individual becomes less likely to do so. A decrease in wage w_i increases the likelihood of using public transportation. Finally, as the cost differential $[p_{ci}(\sigma_i) - p_{ti}]$ increases, the individual is more likely to use public transport.

We therefore write demand for public transportation as:

$$\text{Demand}_{\text{Public Transit}} = \begin{cases} 1, & [t_{ti} - t_{ci}(\sigma_i)] < \frac{[p_{ci}(\sigma_i) - p_{ti}]}{w_i} \\ 0, & [t_{ti} - t_{ci}(\sigma_i)] \geq \frac{[p_{ci}(\sigma_i) - p_{ti}]}{w_i} \end{cases} = \mathbb{1}\{[t_{ti} - t_{ci}(\sigma_i)] < \frac{[p_{ci}(\sigma_i) - p_{ti}]}{w_i}\} \tag{7}$$

This leads to the following predictions:

- Individuals with low wages w_i are less likely to go out, but those who do are more likely to use public transport. From a demographic perspective, this should be the case for low-income households.

- Individuals living in urban areas are more likely to go out as public transportation costs p_{ti} and transportation times t_{ti}, t_{ci} are likely to be lower relative to suburban areas. From a demographic perspective, this should be the case for minorities.

- Individuals with several friends who are going out have a higher σ_i and are more likely to go out and more likely to use public transportation as the costs of driving become higher. From a demographic perspective, this should be the case for younger individuals, who have more friends going out.

[9] This matches intuition — we are given that the cost of public transportation is smaller than the cost of driving, and when $t_{ti} < t_{ci}$, then public transportation is both faster and cheaper, and therefore the preferred option.

4 Data

Yelp data provided by the Boston Area Research Initiative is used to evaluate the impact of late night service on 2,766 local restaurants and bars located in Boston, Brookline, Cambridge, and Somerville, approximately 75% of which are in treatment areas. Yelp reviews serve as a proxy for the number of customers visiting businesses. The data contains 600,000 reviews from 2005 to 2019, but only the 553,000 reviews between 2010 and 2018 are used considering the limited number of reviews prior to 2010 and to align with the time frame of the other analyses in this thesis.

Twitter activity data within the city of Boston is provided by the Twitter Sandbox Archive and the geographic coordinates of tweets inform which regions had late night tweets.[10] To measure the change in tweets over time, the Twitter Sandbox Archive data is supplemented with subsets of the Archive.org Twitter stream data, which provides a larger sample of tweets. Twitter data ranges from 2010 to 2018.

The MBTA service schedule is provided by the MBTA's General Transit Feed Specification, an administrative dataset including all MBTA trips from March 3, 2009 to present day, including data on trips for each calendar day. The data provides the exact dates and stops with late night service and the geographic coordinates of each stop. To measure take-up of late night service, it would be ideal to have data on ridership for each trip, but the MBTA does not provide this and instead aggregates ridership daily across all bus and subway lines.

Crash data is provided by the Massachusetts Department of Transportation IMPACT Open Data Portal, an administrative dataset with all crashes occurring in Massachusetts.[11] The data contains geographic coordinates of 95% of the over 1 million crashes occurring between 2010 and 2019, and those without coordinates are excluded from the analysis. On

[10] The free Twitter Sandbox Archive allows only for the collection of a sample of tweets for each year of the analysis, as opposed to every tweet sent.

[11] Vehicle level data is used to obtain information on all vehicles involved in each crash. If a crash occurs involving two vehicles, it is listed twice in the data, allowing for a better estimate of the number of individuals who chose to drive in lieu of taking public transportation. For robustness, variations of the initial analyses considering the number of crashes instead of the number of vehicles are used, but doing so does not meaningfully alter the magnitude nor the statistical significance of the results.

weekends, the average daily number of late night crashes in the region is 0.174. The portal provides data from 2002 to 2019, though crashes after 2017 are provisional and subject to change and crashes before 2010 are not used in this analysis.

Demographic characteristics are pulled from the American Community Survey 2013 — 2018 5-year estimate.[12] The TIGER 2018 geographic shapefiles provided by the United States Census are used to allocate demographics to the smaller, rectangular regions used in this analysis and further described in the following section.

5 Empirical strategy

This section begins with a strategy to measure the impact late night service had on local businesses. Subsequently, it proposes a framework to consider whether late night service led to a reduction in the number of car crashes. The car crash analysis is divided into two parts. The first aims to calculate an average treatment effect on the treated of the entire Greater Boston Area. The second leverages the heterogeneity of the treatment effect and explores the areas with the strongest effect. For all three analyses, all variables have been normalized to be in units of standard deviations.

All analyses range from 2010 to 2018, the dates for which the majority of the data sources have available data. The regions considered for each design are divided into smaller rectangles of width 0.01 degrees of latitude (approximately 1.10 kilometers) and 0.01 degrees of longitude (approximately 0.82 kilometers), with an area of 0.90 square kilometers.[13] A treated rectangle is defined as one with a stop benefiting from late night service in the three-by-three area around it.[14]

One concern for these analyses are potential omitted variables, such as the increased

[12]This includes several data variables on population, age, household income, and racial and ethnic composition of residents in the relevant area of study.

[13]More details on the benefits and trade-offs of this division strategy and an alternative are in Section 9.1.

[14]For robustness, all specifications in Sections 5.1 and 5.2 are also run with a "stricter" treatment definition, which defines a rectangle to be in the treatment group only if the rectangle contains a stop with late night service. This does not materially alter the magnitude or significance of the majority of the analyses.

penetration of Uber, Lyft, and other ride hailing companies. Uber entered Massachusetts in 2011 and grew in both number of drivers and riders during the period of analysis. However, this occurred throughout the area of study homogeneously, since it is a single metropolitan area. The impact of these ride hailing applications on car crashes is captured in the time-fixed and location-fixed effects and should not interfere with the differences-in-differences estimators. Another consideration is whether differential weather effects could have impacted treatment and control areas differently. The small area of the region of study again makes this unlikely — weather effects in the region should be mostly homogeneous. Weather effects are also accounted for in season-fixed effects.

Per Chen et al. (2019), let i denote the geographic regions and $t = 1, ..., T$ denote the time periods included in the analysis from January 1, 2010 to December 31, 2018, with T denoting the last period in the series.[15] The treatment period T_p ranges from March 28, 2014 to March 20, 2016. For each rectangle i, we observe outcome Y_{it}^{obs},[16] treatment status $D_{it} = \mathbb{1}(t \in T_p)D_i$, and a vector of covariates X_i. The potential outcomes are $Y_{it}(1)$ and $Y_{it}(0)$ with $Y_{it}^{obs} = D_{it}Y_{it}(1) + (1 - D_{it})Y_{it}(0)$. I consider the average treatment effect on the treated (ATT), given by:

$$\tau = \mathbb{E}[Y_{it_p}(0) - Y_{it_p}(0)|D_i = 1] \tag{8}$$

Assume we have a balanced panel where $Z_i = (Y_i, X_i, D_i)$ which are independently and identically distributed.

5.1 Impact on Local Businesses

The first part of the analysis considers the impacts of late night service on local businesses. Two outcomes are considered: the number of Yelp reviews and the number of late night tweets. The number of Yelp reviews is used as a proxy for visits to local businesses, as an increase in the number of reviews of a business indicates there is more interest in that

[15]For Section 5.1, t represents months; for Sections 5.2 and 5.3, t represents days.

[16]For Section 5.1, the outcome variables are the number of Yelp reviews and the number of tweets. For Sections 5.2 and 5.3 the outcome variable is the number of weekend late night crashes.

business and serves as a proxy for a higher quantity of individuals going to the business. The number of late night tweets is used as a proxy for the quantity of individuals in an area, and an increase in the number of individuals in a region could also lead to more customers for the businesses in that region.

The section employs a differences-in-differences design that optionally incorporates observable covariates with time-fixed effects γ_t and location-fixed effects ϕ_i to isolate the causal effect of the increase in public transportation supply on the number of Yelp reviews about local businesses and the number of late night tweets. For both the Yelp and Twitter datasets, the regression specification is:

$$Y_{it}^{obs} = \tau \mathbb{1}(t \in T_p, D_i = 1) + \alpha \mathbb{1}(D_i = 1) + \beta \mathbb{1}(t \in T_p) + \gamma + \delta_t + \phi_i + \epsilon_{it} + \text{Covariates} \quad (9)$$

Y_{it} is the number of reviews for business i in month t for the Yelp data and the number of late night tweets in region i in month t for the Twitter data. The average treatment effect on the treated is τ, while α provides the treatment group specific effect and β provides the time trend common to control and treatment groups. The covariate for the Yelp data is *avg_rating*, which represents the average rating of a business. For Twitter, the covariate is *pop*, the population in the region. Standard errors for Yelp are clustered by business-month pair and for Twitter are clustered by region-month pair.[17] The treatment period is from March 28, 2014 to March 20, 2016, and the treatment group is composed of the rectangles which have a station with late night service within their three rectangles by three rectangles area. The results for the Yelp regression are in Table 2 while those for Twitter are in Table 3.

Table 2 presents the results of several variations of the differences-in-differences regressions for the Yelp data. Column (1) presents the results without fixed effects, while Column (2) includes business and year fixed effects. Column (3) includes the business rating average control. Columns (4) and (5) include only Yelp reviews which occurred on weekends, the days of late night service. Column (4) includes fixed effects and Column (5) adds the rating

[17] The rationale for this is explained in Appendix 9.3.

average control.

Table 3 is structured similarly to Table 2, but it presents the results for the Twitter regressions. The columns are arranged in the same order — Column (1) has no fixed effects, Column (2) includes region and year fixed effects, Column (3) includes the population control, and Columns (4) and (5) include only weekend tweets.

For both the Yelp and Twitter datasets, I conduct three tests to supplement the parametric regressions. The first is a plot of the number of reviews over time for the treatment and control groups, shown in Figure 2 for the Yelp data and Figure 5 for the Twitter data. The second is allowing for linear time trends, which all regressions in Tables 2 and 3 include. The last supplemental test is a differences-in-differences regression which has interactions of the treatment dummy with monthly indicator variables per the following specification:

$$\begin{aligned} Y_{it}^{obs} =& \tau_1 \mathbb{1}(t = \text{Jan-2010}, D_i = 1) + \tau_2 \mathbb{1}(t = \text{Feb-2010}, D_i = 1) + ...+ \\ & \tau_{108} \mathbb{1}(t = \text{Dec-2018}, D_i = 1) + \beta_1 \mathbb{1}(t = \text{Jan-2010}) + ...+ \\ & \beta_{108} \mathbb{1}(t = \text{Dec-2018}) + \alpha \mathbb{1}(D_i = 1) + \gamma + \delta_t + \phi_i + \epsilon_{it} + \text{Covariates} \end{aligned} \tag{10}$$

Figure 3 plots the coefficients of these interaction terms for the Yelp data while Figure 6 does the same for the Twitter data. Figure 4 shows the difference between treatment and control groups for Yelp reviews, normalized to account for the different sizes of both groups, and Figure 7 does the same for the Twitter data.

5.2 Crash Analysis: Complete Dataset

Going beyond the impacts of late night service on local businesses, this section estimates how the service saved lives and costs by preventing car crashes. For this, the number of car crashes in four Massachusetts counties are considered: Essex, Middlesex, Norfolk, and Suffolk.[18] Approximately 7% of the over 5000 rectangles in the region are considered treated, and the remaining 93% form the control group.

[18]The majority of the MBTA's bus and subway lines are within these four counties as they compose nearly all the Greater Boston Area. They have similar demographics, including population density, median household income, poverty rate, and median age, and more information can be found in Section 9.2.

As in Section 5.1, this section employs a differences-in-differences design that optionally incorporates observable covariates with time-fixed effects[19] δ_t and location-fixed effects ϕ_i to isolate the causal effect of the increase in public transportation supply on the number of weekend late night car crashes. With the parallel trends assumption, we can consistently estimate τ with the OLS regression:

$$Y_{it}^{obs} = \tau \mathbb{1}(t \in T_p, D_i = 1) + \alpha \mathbb{1}(D_i = 1) + \beta \mathbb{1}(t \in T_p) + \gamma + \delta_t + \phi_i + \epsilon_{it} + \text{Covariates} \quad (11)$$

The ATT coefficient is τ, while α provides the treatment group specific effects and β provides the time trend common to control and treatment groups. The covariates include *young_pop*, *low_income_hh*, *minority_pop*, and *pop*, which represent the proportion of young individuals, the proportion of low income households, the proportion of minorities, and the population in each region. Standard errors are clustered by region-month pair.[20] The treatment period is from March 28, 2014 to March 20, 2016, and the treatment group is composed of the rectangles which have a station with late night service within their three rectangles-by-three rectangles area.

The results for this first part of the analysis are presented in Table 4. Column (1) indicates the differences-in-differences estimator without controls and Column (2) includes controls for demographic variables age, household income, and race and ethnicity.

An additional regression estimated is the differences-in-differences design interacting with age, household income, race and ethnicity, and Twitter data to allow for heterogeneous ATTs based on regional demographics and to highlight the heterogeneity of the treatment effect as a function of demographics.[21] For each of these define A_i as value of the demographic variable at region i measured in standard deviations, and run the OLS regression:

$$\begin{aligned} Y_{it}^{obs} =& \omega_v A_i \mathbb{1}(t \in T_p, D_i = 1) + \tau \mathbb{1}(t \in T_p, D_i = 1) + \alpha \mathbb{1}(D_i = 1) \\ &+ \beta \mathbb{1}(t \in T_p) + \theta_v A_i + \gamma + \delta_t + \phi_i + \epsilon_{it} + \text{Covariates} \end{aligned} \quad (12)$$

[19]Includes both year and season-fixed effects.
[20]The rationale for this is explained in Appendix 9.3.
[21]The age, income, and race and ethnicity data provide information on the individuals living in the region, whereas the Twitter data provides a different perspective by indicating which areas have activity during weekend late nights.

Here ω_v represents the additional ATT based on the demographics of the region and θ_v provides the trend for regions with similar demographics. Column (3) of Table 4 shows the coefficients with A_i representing the age variable, Column (4) with A_i representing the income variable, Column (5) with A_i representing the race and ethnicity variable, Column (6) with all three demographic variables, and Column (7) with A_i representing the Twitter activity variable.

I conduct three supplemental tests. The first is a plot of the number of weekend late night crashes for the treatment and control groups over time controlling for region, season, and year fixed effects and presented in Figure 8. The second is the inclusion of linear time trends, which are included in all regressions in Table 4. The third, shown in Figure 9, is the plot of the coefficients of interactions of the treatment variable with seasonal indicator variables, similar to Equation 10 in Section 5.1 but with indicators for seasons instead of months.

To provide further evidence of causality and test for omitted variables, I run placebo tests on crashes that should be unaffected by late night service. I alter the outcome variable in Equation 11 to be either weekend non-late night crashes, weekday non-late night crashes, or weekday late night crashes and report these results in Table 6. I also run placebo tests erroneously modifying the dates of treatment to be from January 1, 2012 to December 31, 2013 for the first regression and January 1, 2017 to December 31, 2013 for the second regression, presenting these results in Table 7.

5.3 Crash Analysis: Young and Minority Regions

The second part of the crash analysis provides a more rigorous understanding of two subsets — the regions which have the top quartile of young populations and the regions which have the top quartile of minority populations, as these two groups demonstrate the clearest observable treatment effect.[22]

[22] A third region, of the top quartile of low income households, is also considered, but the graph of crashes over time indicates weaker observable effects than in the other two regions.

Both subsets undergo three analyses, using a similar methodology to Chen et al. (2019). The first is a differences-in-differences design with tests for pre-treatment trends and a plot of crashes in the treatment and control groups. This is the same specification as in Equation 11 in Section 5.2, but limited to each subset instead of the full dataset. These differences-in-differences results are reported in Table 5, with the young region results without controls in Column (1) and the regression with controls in Column (2). For minority regions, the corresponding regressions are in Columns (5) and (6).

However, the identification strategies used here and in Section 5.2 are limited to the extent that regions receiving late night service differ in observable characteristics. These unbalanced characteristics weaken the parallel trends assumption.[23] Therefore, the second analysis within this section uses propensity score matching to incorporate observed covariates in a non-parametric manner. First I calculate propensity scores, defined as the probability $P(D_i = 1|X_i)$, and these are used to provide an estimator of the treatment effect which is consistent if the outcome regression model or the propensity score model is correctly specified. The results from the propensity score analysis are included in Table 5 in Column (3) for young regions and Column (7) for minority regions.

The third strategy, also based on Chen et al. (2019), compares the treated regions with their geographically most proximate non-treated regions, providing a geographic comparison of the treatment effect. Each treatment region i is paired with the geographically most proximate non-treatment region $\tilde{i}$ such that $\tilde{i}$ is in the control group and has the smallest distance to i's center. For each pair $\iota = (i, \tilde{i})$, we have $Y_{it}(0) = \delta_\iota + \alpha_{\iota t} + \epsilon_{\iota t}$ and $Y_{\tilde{i}t}(0) = \tilde{\delta_\iota} + \alpha_{it} + \tilde{\epsilon_{it}}$ with $Y_{it}(0) - Y_{\tilde{i}t}(0) = (\delta_\iota - \tilde{\delta_\iota}) + (\epsilon_{\iota t} - \tilde{\epsilon_{it}})$ so that I estimate:

$$Y_{it}^{obs} - Y_{\tilde{i}t}^{obs} = \tau \mathbb{1}(t \in T_p, D_i = 1) + \mu_\iota + \eta_{\iota,t} \tag{13}$$

This estimation is consistent if $\mathbb{E}[\eta_{\iota,t}|D_i] = 0$. Note this requires the pair has the same trend $\alpha_{\iota,t}$. The results for the geographic proximity analysis are presented in Table 5 in Column (4) for young regions and Column (8) for minority regions.

[23] There is, however, non-parametric evidence that parallel trends hold, as seen in the Results section.

6 Results

This section presents the results for all strategies described in Section 5. Tables 2 and 3 and Figures 2, 3, 4, 5, 6, and 7 present the results for the strategies discussed in Section 5.1. The results for the strategies in Section 5.2 are presented in Tables 4, 6, and 7 and Figures 8 and 9. Finally, the results for the strategies discussed in Section 5.3 are in Table 5 and Figure 10. Note that all variables are measured in standard deviations, except for dummy variables.

6.1 Impact on Local Businesses

This section presents the results from the strategies outlined in Section 5.1 to measure the impact of late night service on local businesses. There is parametric and non-parametric evidence consistent with an increase in both the number of Yelp reviews for treated businesses and late night tweets for treated areas.

The results for the impact of late night service on Yelp reviews are found in Table 2. Columns (1) and (2), which vary only by the inclusion of fixed effects, indicate a coefficient $\hat{\tau}$ of 0.199 standard deviations, statistically significant at the 99% level. This is the equivalent to an 8% increase caused by the treatment. The addition of year and business fixed effects vary the coefficients capturing the treatment group specific effects $\hat{\alpha}$ and the time trend common to the control and treatment groups $\hat{\beta}$, changing the sign of both and reducing the magnitude of $\hat{\beta}$.

The addition of the average rating control, represented in Column (3) of Table 2 indicates that controlling for the average review decreases the treatment effect to about a third of its previous estimate, reducing it to 0.063 standard deviations or 2.5%, but maintaining its statistical significance at the 99% level. The decrease in magnitude is caused by the very strongly negative coefficient -1.041 of the *avg_review* control, which dominates other coefficients and results in the coefficients of other variables also decreasing. The very small standard error indicates that the average rating is a strong predictor of the number of reviews. In this sample, the number of reviews is negatively correlated with the average rating. However, the non-parametric evidence provides visual support that is most consistent with Columns (1) and (2), as it indicates the treatment effect is visually perceivable and therefore likely higher than 2.5%.

Columns (4) and (5) show the results including only Yelp reviews on weekends. The point estimate increases to 0.207 standard deviations without controls and 0.153 with controls, and statistical significance decreases to the 90% level. The slight increase in the point estimate is expected, as late night service occurred during the weekends and therefore the effect is most felt during weekends, yet individuals don't always review their experiences the same day they have them, so an individual who has an experience during the weekend may wait until the weekday to rate it and vice-versa. Hence, one would expect only a slight increase when considering the weekend subset only, as is consistent with the results.

Figure 2 provides a plot of Yelp reviews over time, which accounts for both time and business fixed effects. Note the variation between the control and treatment groups prior to the service is consistent with the pre-treatment parallel trends assumption, as both groups vary relatively similarly over time. The difference in levels between the treatment and control groups is caused by the larger size of the treatment group relative to the control group, and so that difference is unlikely to impact the evolution of trends. There is evidence of an increase in the number of reviews for businesses in the treatment group shortly after late night service is established. This increase occurs primarily in the beginning of the service, and the number of reviews does not increase incrementally during service. After the conclusion of the service, there is a slight drop in the number of reviews in the treatment group. This evidence is consistent with there being a positive impact on the number of Yelp reviews for treated businesses.

To illustrate whether the treatment effect evolved over time, Figure 3 provides the plot of coefficients of the interaction of monthly indicator variables with the treatment variable, as per Equation 10 in Section 5.1. There is a noticeable increase in magnitude shortly after the beginning of the period, and another maximum approximately halfway through the treatment. During both of these maxima, the lower bound of the confidence interval does surpass the zero mark, providing evidence against the null hypothesis. The first maximum is likely explained by the treatment effect, while the second is aligned with the spike in the number of reviews-mid treatment, which seems to be periodic in nature. Throughout the remainder of the plot, the confidence interval is imprecise enough so that rejecting the null hypothesis is not possible. Figure 4 provides similar evidence as that of Figure 3, as it shows the normalized difference between the treatment and control group, indicating a small difference prior and after treatment, and a more significant difference during the treatment itself. This is consistent with the parallel trends assumption as it indicates the difference between the two groups prior to treatment was not significant.

The Yelp data therefore indicates an increase in the number of reviews in treated businesses. As an alternative proxy, I consider Twitter data to see whether treated regions had an increased number of late night tweets, which would also indicate higher late night activity. The number of late night tweets does increase, but with a smaller magnitude relative to the increase in Yelp reviews.

The results for the impact of late night service on the number of late night tweets are presented in Table 3. Columns (1) and (2), which vary only by the inclusion of fixed effects, indicate a treatment effect $\hat{\tau}$ of 0.123 standard deviations, a 5% increase statistically significant at the 99% level. The addition of region fixed effects changes the sign of the treatment group trend $\hat{\alpha}$, while the time trend $\hat{\beta}$ changes sign and decreases in magnitude with the addition of year fixed effects.

The inclusion of population as a control in Column (3) of Table 3 does not change the magnitude of the treatment effect, as the estimate drops only slightly to 0.111 standard deviations or 4%. The statistical significance of the point estimate decreases slightly, but remains significant at the 95% level. The lack of change in the estimates when controlling for population is likely because the population is relatively homogeneous in the city of Boston, which is the area for which the Twitter data is available.

Considering only tweets published on weekends leads to a higher point estimate of 0.173 standard deviations without controls significant at the 95% level in Column (4) and 0.154 standard deviations significant at the 90% level including the population control in Column (5). Note using the weekends-only subset increases the point estimate more strongly for tweets than it did for Yelp reviews — this is expected due to the nature of the two platforms. While a user might be tweeting while they are having an experience, they are unlikely to review the experience live on Yelp.

The two plots presented in Figures 5 and 6 are consistent with these regressions. Figure 5 provides a plot of the number of tweets over time, which accounts for both time and location-fixed effects. The pre-treatment parallel trends assumption seems to hold, as the treatment and control group evolve similarly prior to treatment. As in the Yelp data, the difference in levels between the treatment and control groups is caused by the larger size of the treatment group relative to the control group, and so that difference is unlikely to impact the evolution of trends. The number of tweets during the treatment does appear to be higher than prior to the treatment, and slightly higher than the period after late night service ends.

Figure 6 presents the plot of coefficients of the interaction of monthly indicator variables with the treatment variable, as per Equation 10 in Section 5.1. The coefficients during the treatment have a larger magnitude relative to the coefficients before and after the treatment. The 95% confidence interval is less precise than it is in the regressions, making it impossible to reject the null hypothesis based on this plot. However, Figure 7 does indicate an increase in the difference between the treatment and control groups during treatment, both strengthening the parallel trends assumption and providing evidence consistent with a treatment effect indicating an increase in the number of late night tweets.

6.2 Crash Analysis: Complete Dataset

This section presents the results for the strategies discussed in Section 5.2. Table 4 presents the majority of these results, with the first two columns representing regressions as specified in Equation 11 and the other five columns representing the regressions specified in Equation 12.

Column (1) provides the regression coefficients for the regression without any demographic controls (Equation 11). The treatment effect $\hat{\tau}$ of -0.113 or -4.5% is small in magnitude and statistically imprecise. The addition of demographic controls in Column (2) does not lead to a more precise estimate of the treatment effect but does lead to a reduction in the coefficient of the treatment dummy $\hat{\alpha}$ from 0.243 standard deviations to 0.164.

The addition of interactions with demographic variables as per Equation 12 leads to more precise estimates. Variations of these regressions are in Columns (3) through (7). The coefficients of the controls including the treatment dummy $\hat{\alpha}$, the period dummy $\hat{\beta}$, the young variable $\hat{\theta_y}$, low income variable $\hat{\theta_i}$, minority variable $\hat{\theta_m}$, and population variable $\hat{pop}$ do not change significantly in any of the regressions.

In Column (3), the interaction is with the young variable. The differences-in-differences coefficient $\hat{\tau}$ becomes 0.090, altering sign and becoming statistically significant at the 90% level. The change in sign to a positive coefficient is explained by the negative coefficient $\hat{\omega_y}$ of the interaction term with the young variable of -0.335, statistically significant at the 95% level. The coefficient of -0.335 has the largest magnitude of all coefficients estimated in Column (3), and is more than twice any other coefficient. Columns (4) and (5), which have interactions with the income variable and the minority variable, present similar results as those in Column (3), as the treatment effect $\hat{\tau}$ is estimated at 0.110 and 0.112, respectively, and is statistically significant at the 99%, while $\hat{\omega_i}$ and $\hat{\omega_m}$ are -0.637 and -0.510 respectively.

The demographic variables have higher means in the treated areas than in the control areas and the control group is significantly larger than the treatment group. Therefore, the normalization of the variables leads them to have positive means in the treatment group. After normalization, the means in the treatment group are 0.032 for the young variable, 0.625 for the income variable, and 1.244 for the minority variable. This explains why the differences-in-differences coefficient $\hat{\tau}$ becomes positive, as the negative interaction coefficients $\hat{\omega_v}$ bring down the treatment effect within the treatment group.

Consistent with this logic, the inclusion of all three interaction terms in Column (6) leads to a higher estimate of the differences-in-differences coefficient $\hat{\tau}$ of 0.222, statistically significant at the 99% level, because of the additional strongly negative coefficients. The coefficients of the interaction terms are all negative and have larger magnitudes than the differences-in-differences coefficient. The coefficient $\hat{\omega_y}$ of the interaction with the young variable is -0.236 and is not statistically significant; the coefficient $\hat{\omega_i}$ of the interaction with the income variable is -0.345 and is statistically significant at the 95% level; and the coefficient $\hat{\omega_m}$ of the interaction with the minority variable is -0.401 and is statistically significant at the 99% level.

Column (7) considers Twitter activity instead of demographic variables, and uses the specification in Equation 12 with an indicator variable of whether an area had late night Twitter activity for the interaction. The differences-in-differences estimate $\hat{\tau}$ of -0.027 is statistically insignificant and the interaction term's coefficient $\hat{\omega_t}$ is estimated to be -0.048 and is also insignificant. The point estimates suggest areas with Twitter activity experienced a treatment effect of nearly three times that in areas without Twitter activity. This is likely because areas with Twitter activity are more likely to have young individuals active during late night, and young individuals are more likely to take public transportation according to Anderson (2016).

Turning to non-parametric evidence, Figure 8 shows a plot of the weekend late night crashes over time with vertical lines indicating the late night service start and end dates. The figure indicates similar trends in both groups prior to the treatment, consistent with the parallel trends assumption. Before the service, both treatment and control groups had a slightly positive trend and oscillated in similar directions. Furthermore, no treatment effect can be detected visually, consistent with the small estimates calculated in the first panel of Table 4, yet it is evident that the growth rate in the number of crashes slows during the treatment. The control group's crash growth rate also slows, suggesting there may be spillover effects from the treatment into the control group. Once the service is revoked, the number of crashes in both groups increases relative to their rates during treatment.

As an additional test of pre-trends, I interact seasonal indicator variables with the treatment variable as specified in Equation 10 to demonstrate an evolution of the treatment effect over time. Figure 9 shows the plot of these interactions. The 95% confidence interval does not exclude the null hypothesis, but there is a visible treatment effect during the treatment period which disappears after the conclusion of late night service. The gradual decrease of the treatment effect towards the conclusion of late night service may be caused by the decreasing ridership over the course of the service.

For robustness and to verify potential omitted variable bias, I run two different robustness checks. The first is changing the outcome variable of the regression specification in Equation 11 to be crashes which should not be impacted by late night service, as late night service occurred primarily during weekends. These other crashes include weekend non-late night crashes, weekday non-late night crashes, and weekday late night crashes. The results of these regressions are reported in Table 6. For all the regressions, regardless of additional controls, the differences-in-difference estimates are very small. Columns (1) and (2), which have as their outcome variable non-late night weekend crashes, provide a statistically insignificant estimate of -0.005. Columns (3) and (4) have an estimate of 0.003 for the outcome variable of non-late night weekday crashes, and this estimate is statistically significant at the 90% level. Finally, the regressions with an outcome variable of late night weekday crashes have a statistically insignificant coefficient of -0.002. The point estimates for the differences-in-differences coefficients across these three outcomes are very small in magnitude, which is consistent with there not being omitted variables during this analysis which impacted the number of crashes, and the lack of statistical significance suggest the data is not informative on whether there is a trend in these crashes, as expected.

The second robustness check I run are placebo regressions, in which I erroneously alter the treatment periods. Table 7 reports these results. In Columns (1) and (2), the erroneous treatment dates are set to be January 1, 2012 to December 31, 2014, and in Columns (3) and (4) they are January 1, 2017 to December 31, 2018.[24] The differences-in-differences estimate in Columns (1) and (2) is -0.020 and in Columns (3) and (4) it is 0.038; both of these point estimates are statistically insignificant. Again, the small magnitude indicates there is no treatment effect and the statistical insignificance indicate the data is not informative on such an effect; both of these results are expected since no real treatment occurred during the placebo dates.

6.3 Crash Analysis: Young and Minority Regions

The small magnitude of the treatment effect in the first part of the analysis and the impact adding demographic variables has on statistical significance lead me to investigate the heterogeneity of the treatment effect. Specifically, the areas selected for further analysis are those with a high proportion of young individuals, a high proportion of low-income households, and a high proportion of individuals identifying as minorities.

Figure 10 shows the evolution of crashes in these areas and includes season-fixed effects. The graph of the youngest regions shows the treatment and control groups have similar pre-treatment trends and volatility. During treatment, the treatment group experiences a visible decline in the number of crashes. Once the treatment is revoked, crashes begin to rise. The graph for the areas with the highest proportion of low-income households have slightly different pre-treatment trends, but the two groups still tend to co-move. During late night service, the treatment group experiences a visible drop in the number of crashes, but after the service concludes the number of crashes does not change drastically. The parallel

[24] These periods are chosen to prevent any overlap with the actual treatment period.

trends assumption is strongest in the regions with the most minorities, as the control and treatment groups vary closely in their number of crashes. There is also a clear spike in the number of crashes after late night service is revoked.

As the areas with the most young individuals and most minority individuals seem to have the strongest correlation between their treatment and control groups in the evolution of crashes prior to the treatment and the clearest visual treatment effect, I select these two subsets for further empirical analysis. Table 5 shows the results of the three strategies described in Section 5.3 — OLS differences-in-differences, propensity score matching, and geographic proximity analysis — for both the young and minority subsets.

Columns (1) through (4) of Table 5 present the results for the subset composed of the regions with the highest proportion of young individuals. Columns (1) and (2) show the results for the OLS differences-in-differences design, and indicate that controlling for demographics leads to a small change in magnitude, from -0.334 to -0.320 without improving the statistical significance of the results. The reduction in standard deviations is the equivalent to a 13% reduction in the number of car crashes.

For a more precise estimate, I consider the non-parametric approaches. The propensity score matching ATT, shown in Column (3), provides a statistically significant result at the 90% level and a point estimate slightly larger than in the OLS regressions of -0.410, the equivalent to a 16% reduction in crashes. The standard error is about half the magnitude of the point estimate, providing a precise estimate at the 95% level. The geographic proximity analysis, presented in Column (4), indicates a slightly smaller estimate of -0.224, or a 9% reduction, which is not statistically significant. The subset of young regions therefore has treatment effects estimated between 9% and 16% dependent on the design employed.

The last four columns of Table 5 display the results of the subset composed of the regions with the most individuals who identify as African American, Black, Hispanic, or Latino. Columns (5) and (6) provide the differences-in-differences estimates of -0.078 and -0.060, around 3%. The small estimates are also very imprecise, as the standard errors are more than five times the point estimates.

The non-parametric methods again provide more precise and higher point estimates. Propensity score matching, presented in Column (7), provides an estimate of -0.397 or 15.5% which is statistically significant at the 90% level. The geographic proximity analysis provides an estimate with larger magnitude of -0.457, or 17%, significant at the 90% level. The standard errors for both estimates are small relative to the estimates. The treatment effect in the minorities subset therefore ranges between 3% and 17%, dependent on the empirical strategy employed.

The difference in magnitudes from these estimates and the differences-in-differences estimates is likely caused by the location of the minority regions. Many regions with minorities are clustered together and in the treatment area. The differences-in-differences estimate does not account for this clustering, whereas the propensity score matching design allows for identifying control areas which most closely resemble the treatment areas,[25] thereby providing a non-parametric model which allows for comparison between regions with similar demographic characteristics. The geographic proximity analysis also restricts the number of comparable control regions, potentially causing the more precise estimate with a higher magnitude consistent with the propensity score matching design.

The heterogeneity in the treatment effect is likely because minorities and young populations are more likely to use public transportation relative to other demographics. As there is more demand for the service in these areas, take-up is larger and the reduction of crashes follows. Understanding public transportation demand based on demographics and allocating service in response to this demand can therefore play an important role in ensuring the effectiveness of late night service.

[25] The youngest regions are distributed more evenly across the studied area, so this disparity in estimates between the OLS differences-in-differences and the propensity score matching design due to geographical clustering does not occur.

7 Discussion

This section contextualizes the impact of late night service, outlining a cost-benefit analysis and discussing this paper's limitations and potential extensions to the literature.

7.1 Cost-Benefit Analysis

Late night service had an estimated yearly operating cost of $14 million and cost was the primary reason cited for its cancellation (Kassiel, 2018). To measure the benefits of the service, I first outline methods to value the benefits to local businesses, then calculate the cost savings from the reduction in car crash expenses.

7.1.1 Benefits to Local Businesses

The Yelp and Twitter data indicate local businesses may have benefited from late night service, as Section 6.1 shows the number of Yelp reviews for businesses close to late night stops increased and the number of late night tweets close to late night stops similarly increased. The financial impact on local firms could be a consideration for the MBTA Fiscal and Management Control Board when evaluating whether to reintroduce late night service. Unfortunately, this analysis does not contain sufficient data to calculate the dollar benefit of the service to businesses, but in this section I outline how one could do this.

Ideally, one would have a measure of monthly income. This could be self-reported by businesses, or potentially through payment processing companies such as Square or credit card companies such as Visa and Mastercard, should the businesses consent to this data being shared or should these service providers be able to provide de-identified data which indicates no identifiable business information except for an anonymous identifier and a variable indicating whether the business is close to a late night stop. Although ideal, it is unlikely this data would be attainable.

If panel financial data is not available, one potential proxy is to consider public tax filings or other financial sources available for businesses such that at least one recent month's or year's financial performance is available for each business analyzed; similar data could be obtained from databases SimplyAnalytics and Infogroup. One could then create a proxy for income over time by dividing the income by the number of Yelp reviews and create a "dollar per review" or divide by the number of tweets and create a "dollar per tweet" metric, which could then be applied to other years and thereby create a proxy for panel financial data. The main limitation of this method is the reliability of the estimated metrics. One can only use recent data, as the metric is consistent only if Yelp and Twitter are not experiencing organic growth in Boston; otherwise, the metric is unreliable as the number of Yelp reviews or tweets increases without correlation to business income. However, by the beginning of late night service (March 28, 2014), both Yelp and Twitter were established in the Boston area, so an analysis from the treatment date until several years after its conclusion could prove feasible.

Alternative methods can provide rough bounds for financial impact. One such method is to use Yelp's price range data, which provides for each business price levels represented by a number of dollar signs corresponding to a price range. These could be used to estimate ranges in revenue increases by multiplying the bounds of the price ranges by the increased number of Yelp or Twitter activity. A final possibility is to consider retail rents. If rents in areas with late night service increased during late night service, they could provide a minimum estimate of the increase in profits to the businesses in those areas, assuming firms pay these rent increases with new revenues as opposed to reducing their profit margin. This assumption could be validated through interviews with a sample of businesses. Rent data could be obtained through municipal administrative data, but the assumption that rate changes are correlated with increases in profit may not hold, in which case this method would not be consistent.

7.1.2 Car Crash Cost Savings

The OLS estimates in Columns (1) and (2) indicate an average treatment effect on the treated of -0.113 standard deviations, or approximately a 4.5% decrease. The average yearly number of weekend late night crashes in the years without late night service was 1,916. A reduction of 4.5% on the 1,916 crashes is equivalent to 86.22 fewer crashes per year. Although the number of crashes prevented seems low, their high cost leads to significant cost-savings.

Considering the financial and social costs of a crash, such as insurance costs and lost earnings potential, a fatal Massachusetts car crash costs approximately $1.168 million (CDC, 2015), while a car crash with a non-fatal injury costs $70,200 and a crash resulting only in property damage costs approximately $8,900 (Parks, 2012). The crash data indicates 0.3% of car crashes are fatal, 26.7% non-fatal injuries, and 73.0% property damage only. During late night hours, however, the percentage of fatalities increases to 1.2%, while non-fatal crashes and property damage only crashes are reduced to 26.4% and 72.4% respectively. Therefore the weighted average cost for a Massachusetts weekend late night crash is $38,992.40.

The reduction of 86.22 crashes corresponds to a total cost reduction of $3.362 million per year, or about 24% of yearly operating cost of the weekend late night service. The yearly income from weekend late night fares is $1.352 million. The cost-savings and ridership fare total is $4.714 million, or 33.7% of operating costs.

As the cost-savings provided by the treatment are smaller than the operating cost of the service, I estimate next an upper bound for the break-even treatment effect required to completely offset the operating costs. This is done by dividing the difference between the operating cost and the fare income by the weighted average cost of a crash. The estimate serves as an upper bound as it assumes no sponsorship income and that the fare income due to ridership will not change.[26] It also does not take into account the positive impacts on local businesses suggested by the Yelp and Twitter analyses in Section 6.1. For the cost-savings

[26] The reduction in car crashes is led by individuals substituting away from cars towards public transit, and therefore a higher reduction would imply that ridership also increases, thereby raising income from fares.

from prevented car crashes to equal the operating cost of the service minus the current fare income, late night service must prevent 324 crashes, a reduction of 17% in the number of weekend late night crashes.

Section 6.3 indicates routes in regions with the high proportions of young individuals and minorities could reach the break-even threshold. The regions that have the most minorities are clustered together, so it may be possible to define a late night service restricted to these areas, resulting in lower operating costs while maintaining a high treatment effect and corresponding cost reduction.

7.2 Limitations and Extensions

The first limitation is that the analysis does not consider spillover effects beyond the treatment area, yet these may be relevant in measuring the benefits of the service. Consider that a crash occurs in a treatment area, leading to road congestion continuing into a control area, which then causes more crashes in the control area. If the treatment area crash is eliminated with late night service and as a consequence the control area crash is also eliminated, the benefit from the reduced number of crashes in the control group is not captured in this analysis. Figure 8 indicates the growth of weekend late night crashes also decreased in the control group during late night service, suggesting the estimated treatment effect may be a lower bound; including the effect in the control group would likely increase the average treatment effect. A possible extension then is to analyze whether the reduction in car crashes in the control group is attributable to spillover effects from the treatment group; in this case the estimated treatment effect would increase, leading to a higher estimated benefit of late night service.

Another extension is to design optimal late night service routes based on the demographics of each region to leverage heterogeneity. This is accomplished by serving the regions with the highest average treatment effect on the treated: those with high proportions of young and minority populations. Such a design could include the predicted number of crashes prevented in each region, thereby providing a metric of evaluating the optimal routing for late night service, assuming the MBTA Fiscal and Management Control Board considers such an outcome relevant. Linear programming with the constraints of the existing network of routes could be used to optimize late night service scheduling.

As indicated by the Yelp and Twitter analyses in this paper and discussed in Section 7.1.1, more work can be done to explore the impact of late night service on local businesses. Additional social media data like Facebook Check-ins or Instagram Stories could be used to run regressions similar to those outlined in Section 5.1. Furthermore, obtaining data on financial performance and running similar designs would allow one to further estimate the impact of the service on local businesses.

Another interesting analysis and proxy for increased business activity is to use Google Trends data to analyze whether businesses in the treated area during see an increase in Google searches during late night service, indicating a higher interest in those businesses. A third potential outcome to be explored is whether late night service has impacts on crime. The Boston Area Research Initiative 911 Calls database could be used to measure changes in criminal activity during treatment. There are other potential outcomes impacted by late night service — including carbon emissions, productivity, [27] and congestion — all of which could be further explored.

Beyond other outcomes, it is also relevant to consider extending the analysis to new regions. Replicating it in other metropolitan settings could provide information on this analysis' external validity. It is unclear whether denser cities may benefit more than sparser cities. On one hand, dense centers may be more likely to have higher ridership during late night service, but they also may have few individuals opting to drive prior to late night service. On the other hand, while sparser areas may have lower ridership and limited routes, the marginal reduction in cars could be higher as a higher proportion of individuals are likely to be using private cars instead of public transportation prior to the service.

[27] Individuals who are not preoccupied with driving are likely to be more productive.

8 Conclusion

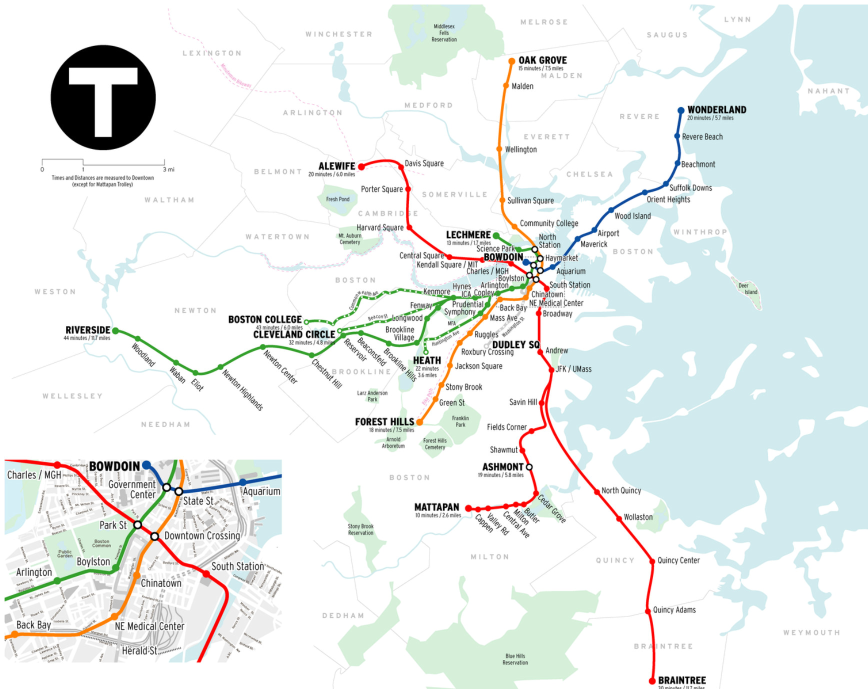
Late night service had detectable effects both on local businesses and in reducing the number of car crashes. Regarding business outcomes, the Yelp data indicates businesses near stops with late night service experienced a statistically significant increase between 0.063 and 0.199 standard deviations or 2.5% and 8% in their number of Yelp reviews. The Twitter data also indicates that regions with late night service had a higher number of tweets, increasing between 0.111 and 0.123 standard deviations or 4%. Both of these results indicate business traffic increased because of late night service, thereby benefiting local businesses.

The impact on car crashes indicates late night service caused a decrease of -0.113 standard deviations or -4% in the number of weekend late night crashes, which led to yearly cost savings of approximately \$3.4 million. The treatment effect is heterogeneous and becomes much stronger in areas with high proportions of young individuals, low income households, and minorities. For the second part of the analysis, the areas with the highest proportions of young individuals and minorities are further explored with three different designs. The treatment effect averaged across these designs was -0.318 or 12.5% in the areas with a high proportion of young individuals and -0.305 or 12% in the areas with a high proportion of minorities.

The heterogeneity suggests that a smaller scale of late night service could be designed targeting those areas with the highest treatment effect to produce cost savings from prevented crashes that are higher than the service operating costs. Even modest reductions like the one estimated in this paper are relevant cost levers when contextualized relative to the scale of the problem — car crashes across the United States continue to increase, and their cost is quickly approaching one trillion dollars annually. Public transportation is still an under-studied substitute that can be further leveraged in public policy toward improving business outcomes and reducing car crashes. Furthermore, considering the increase in Yelp reviews and the increase in Twitter activity in treated areas, further work to evaluate the economic benefits of late night service on nearby businesses could help inform policy makers considering implementing similar services.

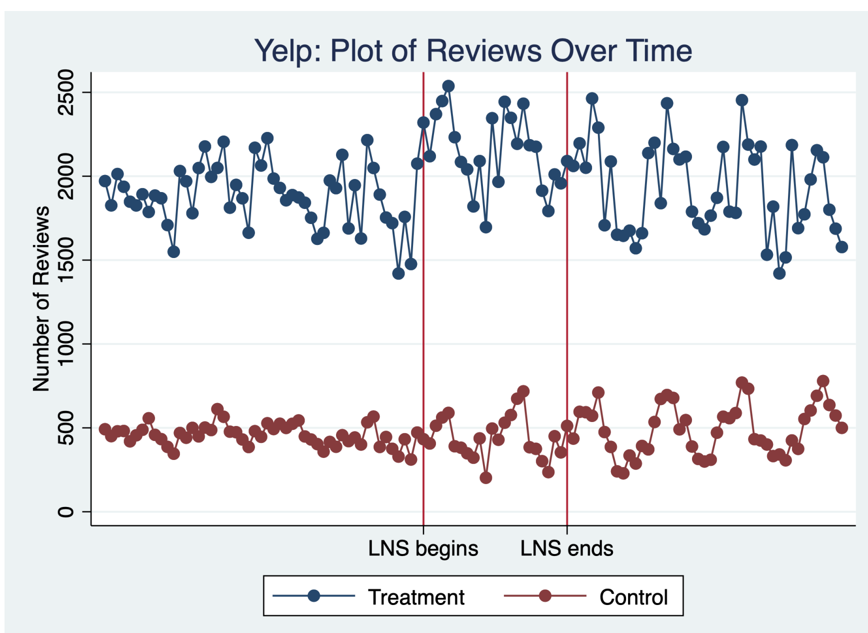
At the same time, one should be cautious about applying these results to other cities, as the different public transportation infrastructure in urban centers may limit the external validity of this analysis. Still, these results suggest the public transportation alternative should be further explored in an effort to support local businesses and resolve the road safety challenge.

Figure 1: Map of MBTA Subway Lines



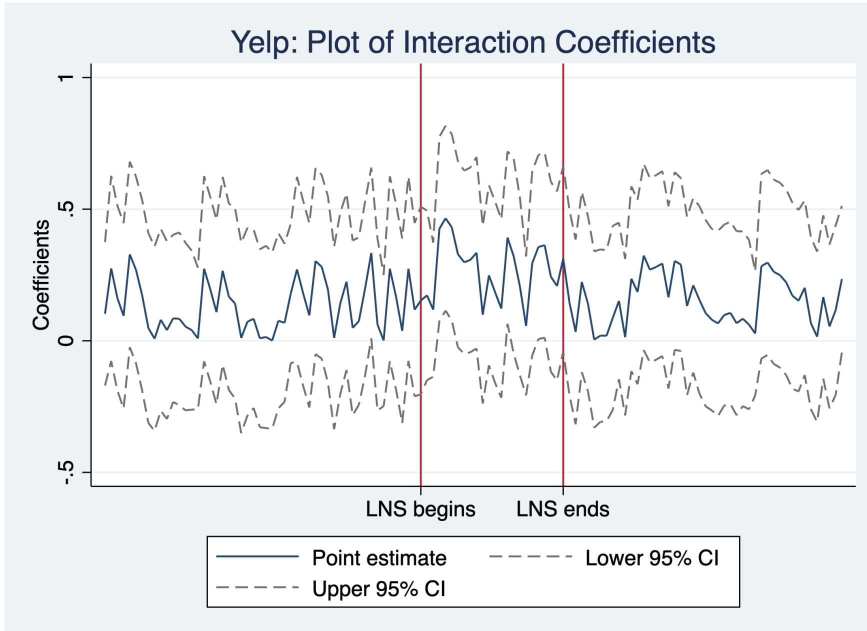
Source: WikiMedia Commons

Figure 2: Yelp: Plot of Reviews Over Time



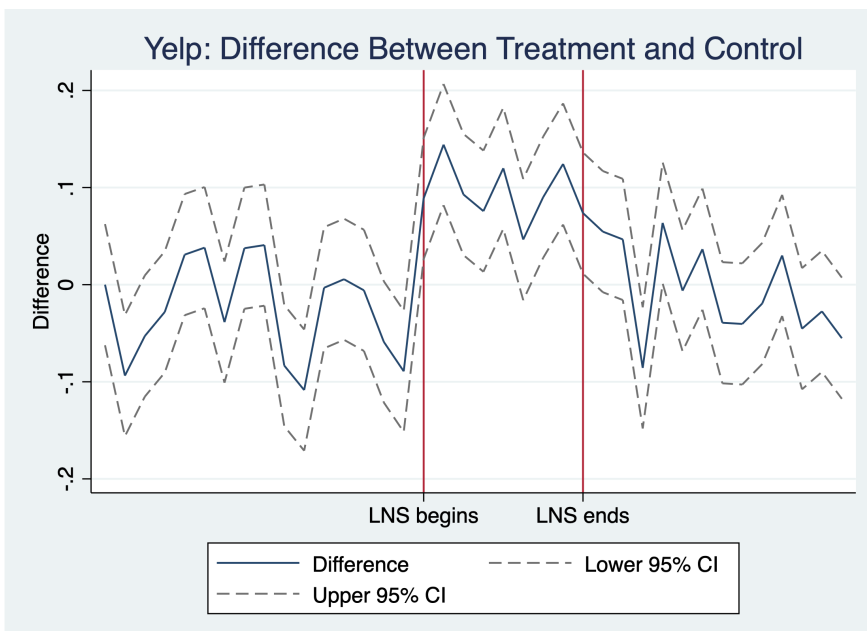
Each point represents the number of reviews during a month from January 2010 to December 2018. The plot accounts for both time and business fixed effects.

Figure 3: Yelp: Diff-in-Diff Treatment Effect Over Time



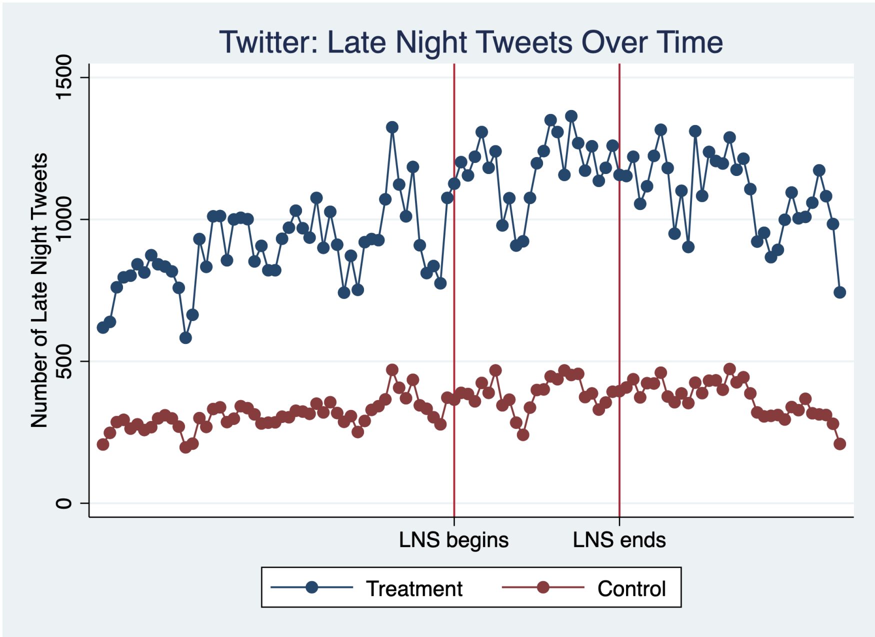
Each point represents a coefficient of the interaction of the treatment variable with an indicator month variable.

Figure 4: Yelp: Difference Between Treatment and Control



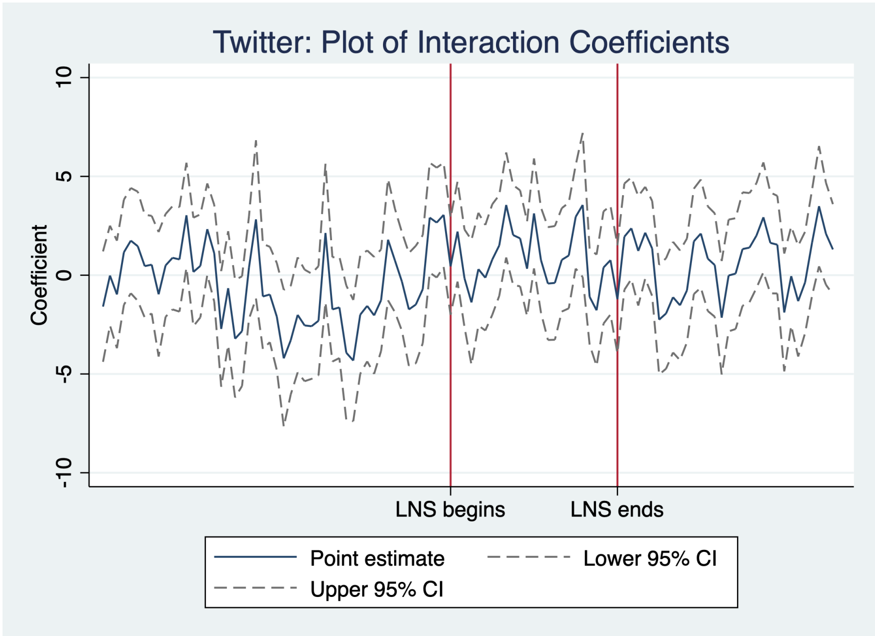
Each point represents the normalized difference of reviews between the two groups, defined as treatment minus control, for each season from March 2010 to November 2018. The plot accounts for both time and business fixed effects.

Figure 5: Twitter: Plot of Late Night tweets Over Time



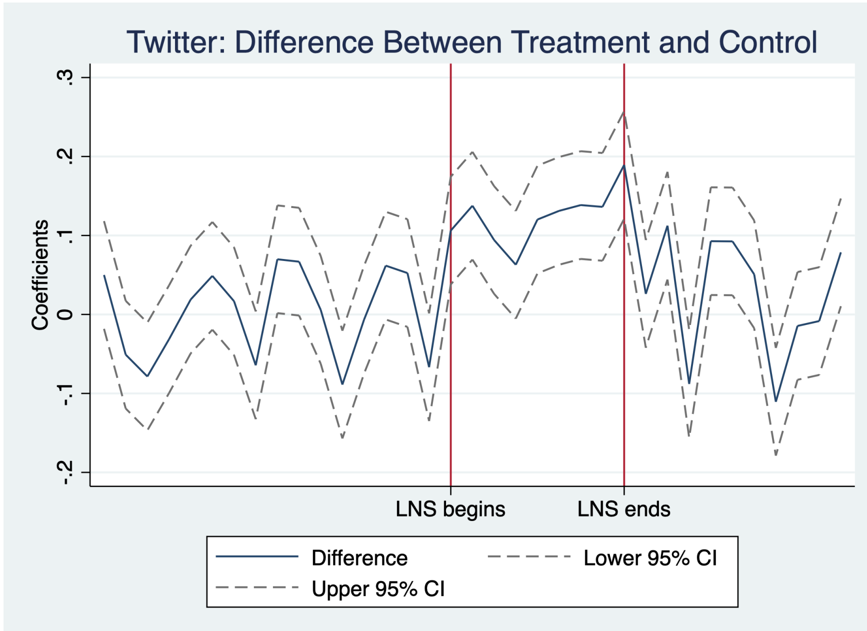
Each point represents the number of tweets during a month from January 2010 to December 2018. The plot accounts for both time and region-fixed effects.

Figure 6: Twitter: Diff-in-Diff Treatment Effect Over Time



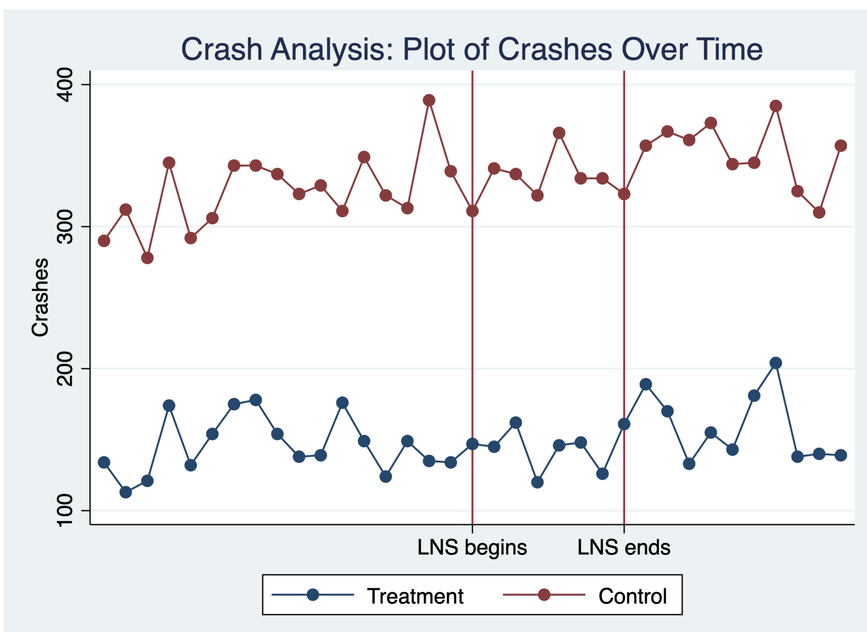
Each point represents a coefficient of the interaction of the treatment variable with an indicator month variable.

Figure 7: Twitter: Difference Between Treatment and Control



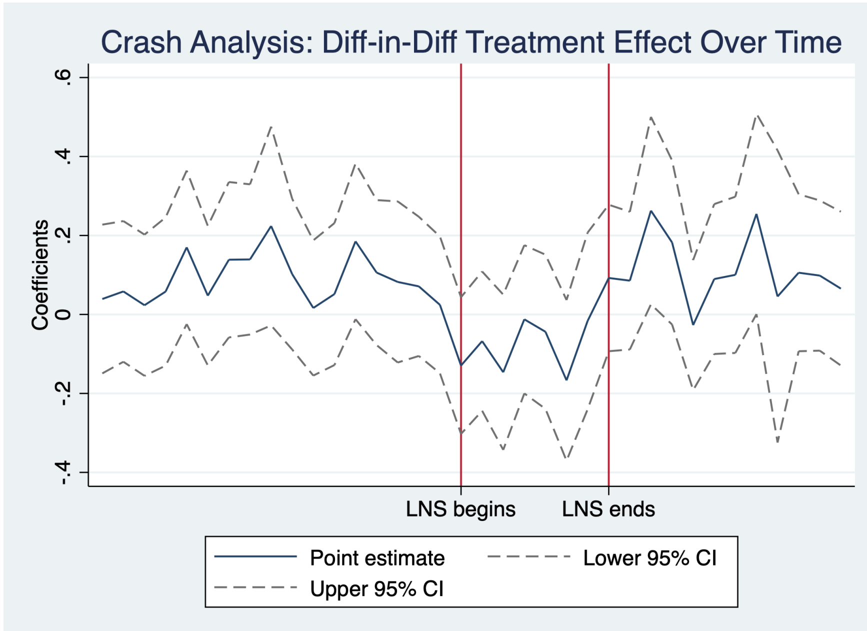
Each point represents the normalized difference of reviews between the two groups, defined as treatment minus control, for each season from March 2010 to November 2018. The plot accounts for both time and business fixed effects.

Figure 8: Crash Analysis: Plot of Crashes Over Time



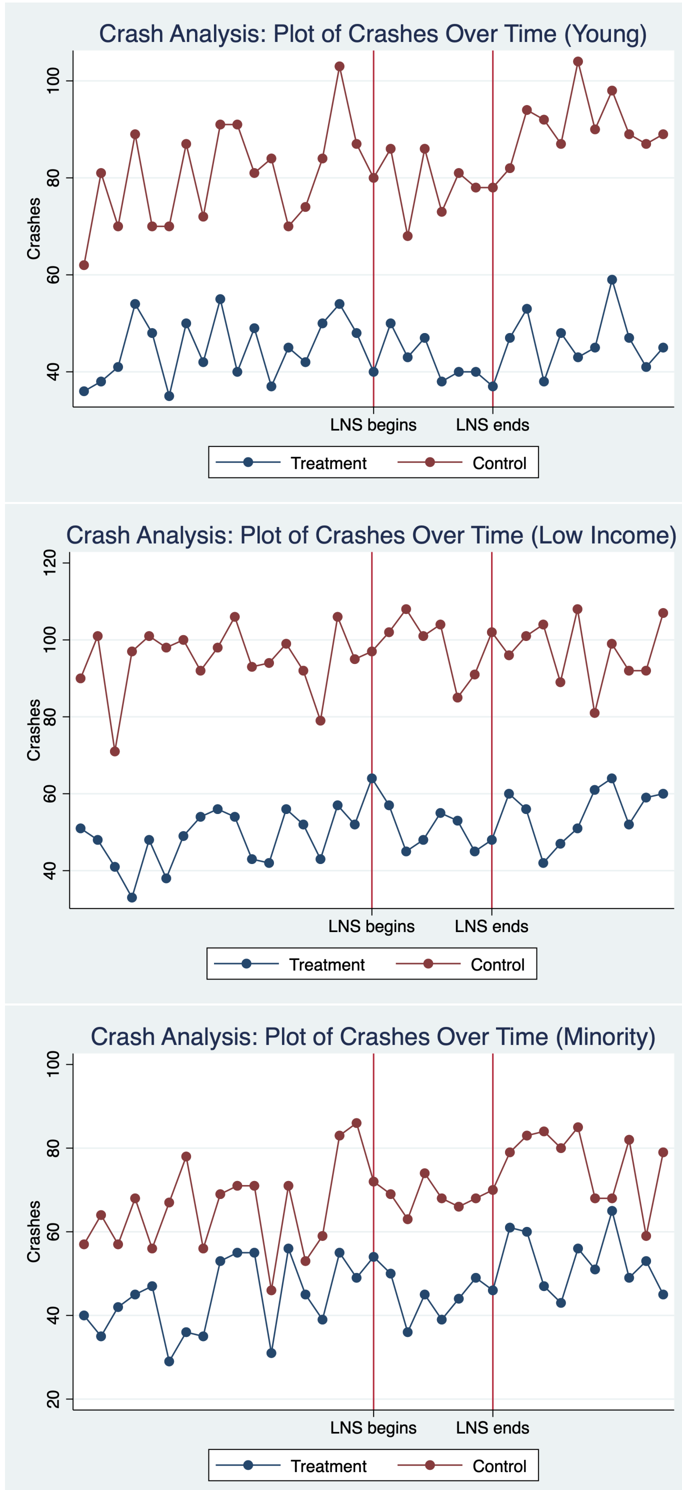
Each point represents the number of crashes during a season from Spring 2010 to Fall 2018. The plot accounts for both time and region-fixed effects. The number of crashes in the treatment group is translated down to allow for comparable scales.

Figure 9: Crash Analysis: Diff-in-Diff Treatment Effect Over Time



Each point represents a coefficient of the interaction of the treatment variable with an indicator season variable.

Figure 10: Crash Analysis: Plot of Crashes Over Time in Subsets



Each point represents the number of crashes during a season from Spring 2010 to Fall 2018. Young regions have a proportion of individuals under 25 of 0.338 or greater; low-income regions have a proportion of households with yearly income below $30,000 of 0.261 or greater, and minority regions have a proportion of Black, African-American, Hispanic, or Latino populations of 0.345 or greater. The plots account for both time and region-fixed effects.

Table 1: Key Driving Statistics in Massachusetts

Year	VMT		Fatal crashes		All crashes	
2010	8,279	(-0.41%)	333	(6%)	115,643	(-2%)
2011	8,285	(0.07%)	356	(7%)	120,632	(4%)
2012	8,395	(1.33%)	362	(2%)	122,646	(2%)
2013	8,387	(-0.10%)	334	(-8%)	125,285	(2%)
2014	8,509	(1.45%)	339	(1%)	130,233	(4%)
2015	8,720	(2.48%)	326	(-4%)	139,050	(7%)
2016	9,057	(3.86%)	360	(10%)	143,474	(3%)
2017	9,130	(0.81%)	334	(-7%)	145,068	(1%)

VMT is the number of vehicle miles traveled per capita. The values in parenthesis are year-over-year changes. Source: DOT (2019)

Table 2: Business Analysis: Yelp Reviews

	Dependent variable: Number of Yelp reviews per month				
	OLS No FE (1)	OLS Incl. FE (2)	OLS Controls (3)	OLS Weekends (4)	OLS Weekends w/Controls (5)
Diff-in-Diff	0.199***	0.199***	0.063***	0.207*	0.153*
Estimator ($\hat{\tau}$)	(0.076)	(0.076)	(0.024)	(0.111)	(0.086)
Treatment	0.422***	-0.399***	0.153***	-0.034***	0.155***
Dummy ($\hat{\alpha}$)	(0.067)	(0.010)	(0.003)	(0.007)	(0.003)
Period	0.497***	-0.134**	-0.042**	-0.097***	-0.031
Dummy ($\hat{\beta}$)	(0.062)	(0.062)	(0.019)	(0.037)	(0.022)
Business Avg.			-1.041***		-1.041***
Rating ($avg\hat{\_rating}$)			(0.000)		(0.000)
Year FE	N	Y	Y	Y	Y
Business FE	N	Y	Y	Y	Y
Linear time trends	Y	Y	Y	Y	Y
R^2	0.009	0.441	0.441	0.456	0.458
Observations	498,960	498,960	498,960	498,960	498,960

***p < 0.01, **p < 0.05, *p < 0.1

Robust standard errors reported in parentheses and clustered by business-month pair. Each observation is the number of Yelp reviews for a business for a month.

Table 3: Business Analysis: Late Night tweets Per Month

	Dependent variable: Number of late night tweets				
	OLS No FE (1)	OLS Incl. FE (2)	OLS Controls (3)	OLS Weekends (4)	OLS Weekends w/Controls (5)
Diff-in-Diff Estimator ($\hat{\tau}$)	0.123*** (0.039)	0.123*** (0.039)	0.111** (0.050)	0.173** (0.079)	0.154* (0.083)
Treatment Dummy ($\hat{\alpha}$)	0.620*** (0.125)	-0.105*** (0.009)	-0.105*** (0.009)	-0.251** (0.117)	-0.343*** (0.045)
Period Dummy ($\hat{\beta}$)	0.045*** (0.010)	0.008 (0.012)	0.008 (0.012)	0.100 (0.076)	0.085 (0.063)
Population ($\hat{pop}$)			0.031*** (0.011)		0.053** (0.027)
Year FE	N	Y	Y	Y	Y
Region FE	N	Y	Y	Y	Y
Linear time trends	Y	Y	Y	Y	Y
R^2	0.107	0.875	0.878	0.835	0.837
Observations	19,764	19,764	19,764	19,764	19,764

***p < 0.01, **p < 0.05, *p < 0.1

Robust standard errors reported in parentheses and clustered by region. Each observation is the number of tweets for a region for a month.

Table 4: Crash Analysis: Complete Dataset

	Dependent variable: Weekend late night crashes per day						
	OLS (1)	Added Controls (2)	Interaction w/Age (3)	Interaction w/Income (4)	Interaction w/Minority (5)	Interaction w/all three (6)	Interaction w/Twitter (7)
Treat. Effect (See notes)	-0.113 (0.257)	-0.113 (0.257)	0.079 (0.185)	-0.288* (0.159)	-0.522*** (0.136)	-0.500 (0.265)	-0.031 (0.107)
Diff-in-Diff Estimator ($\hat{\tau}$)	-0.113 (0.257)	-0.113 (0.257)	0.090* (0.054)	0.110*** (0.032)	0.112*** (0.030)	0.222*** (0.062)	-0.027 (0.026)
Interaction w/ Young ($\hat{\omega_y}$)			-0.335** (0.149)			-0.236 (0.145)	
Interaction w/ Low-inc. ($\hat{\omega_i}$)				-0.637*** (0.136)		-0.345** (0.153)	
Interaction w/ Minority ($\hat{\omega_m}$)					-0.510*** (0.102)	-0.401*** (0.113)	
Interaction w/ Twitter ($\hat{A_t}$)							-0.048 (0.051)
Treatment Dummy ($\hat{\alpha}$)	0.243*** (0.011)	0.164*** (0.012)	0.164*** (0.012)	0.159*** (0.012)	0.153*** (0.012)	0.153*** (0.012)	0.188*** (0.011)
Period Dummy ($\hat{\beta}$)	-0.016*** (0.006)	-0.016** (0.006)	-0.016** (0.006)	-0.016** (0.006)	-0.016** (0.006)	-0.016** (0.006)	-0.016** (0.006)
Young Prop. ($\hat{\theta_y}$)		0.004** (0.002)	0.005*** (0.002)	0.004** (0.002)	0.004** (0.002)	0.005*** (0.002)	
Low-income Prop. ($\hat{\theta_i}$)		0.048*** (0.003)	0.048*** (0.003)	0.050*** (0.003)	0.047*** (0.003)	0.049*** (0.003)	
Minority Prop. ($\hat{\theta_m}$)		0.081*** (0.005)	0.081*** (0.005)	0.082*** (0.005)	0.086*** (0.005)	0.085*** (0.005)	
Twitter Activity ($\hat{\theta_t}$)							0.237*** (0.010)
Population ($\hat{pop}$)		-0.014*** (0.002)	-0.014*** (0.001)	-0.014*** (0.001)	-0.014*** (0.001)	-0.014*** (0.001)	-0.006*** (0.001)
Year FE	Y	Y	Y	Y	Y	Y	Y
Season FE	Y	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y	Y
Linear time trends	Y	Y	Y	Y	Y	Y	Y
Observations	5,215,582	5,215,582	5,215,582	5,215,582	5,215,582	5,215,582	5,215,582
R^2	0.009	0.009	0.009	0.010	0.010	0.010	0.006

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Robust standard errors reported in parentheses and clustered by geocode x month pair. Each observation is the number of crashes for a region for a day. "Prop." is the abbreviation for "proportion", which is calculated as the percentage of individuals in that demographic (young, low-income, or minority). The "Treat. Effect" is the treatment effect calculated by adding the diff-in-diff estimator $\hat{\tau}$ to the product of the interaction coefficient of that regression ω_v times the treatment group mean of that demographic variable.

Table 5: Crash Analysis: Young and Minority Regions

	Dependent variable: Weekend late night crashes per day							
	Young				Minority			
	OLS	Added Controls	P. Score Matching	Geo. Pairs	OLS	Added Controls	P. Score Matching	Geo. Pairs
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Effect Estimator ($\hat{\tau}$)	-0.334 (0.440)	-0.320 (0.440)	-0.410** (0.199)	-0.224 (0.176)	-0.078 (0.439)	-0.060 (0.431)	-0.397* (0.227)	-0.457* (0.252)
Treatment Dummy ($\hat{\alpha}$)	0.230*** (0.019)	0.146*** (0.019)			0.198*** (0.021)	0.081*** (0.022)		
Period Dummy ($\hat{\beta}$)	-0.029** (0.014)	-0.029** (0.013)			-0.067** (0.032)	-0.067** (0.032)		
Young Prop. ($\hat{\theta_y}$)		0.007** (0.003)				0.042*** (0.005)		
Low-income Prop. ($\hat{\theta_i}$)		0.045*** (0.006)				0.098*** (0.008)		
Minority Prop. ($\hat{\theta_m}$)		0.120*** (0.008)				0.083*** (0.008)		
Population ($\hat{pop}$)		-0.009*** (0.002)				-0.047*** (0.006)		
Year FE	Y	Y			Y	Y		
Season FE	Y	Y			Y	Y		
County FE	Y	Y			Y	Y		
Linear time trends	Y	Y			Y	Y		
R^2	0.006	0.016			0.003	0.009		
Observations	850,314	850,314	850,314	850,314	838,335	838,335	838,335	838,335

***p < 0.01, **p < 0.05, *p < 0.1

Robust standard errors reported in parentheses and clustered by geocode x month pair. Each observation is the number of crashes for a region for a day. "Prop." is the proportion of individuals in the region in a demographic (i.e. young, low-income, or a minority). For propensity score matching, covariates include proportion of young population, proportion of households classified as low-income, proportion of population classified as a minority, and total population.

Table 6: Crash Analysis: Robustness Check w/Other Crashes

	Dependent variable:					
	Non-late night weekend crashes		Non-late night weekday crashes		Late night weekday crashes	
	OLS	Added Controls	OLS	Added Controls	OLS	Added Controls
	(1)	(2)	(3)	(4)	(5)	(6)
Diff-in-Diff Estimator ($\hat{\tau}$)	-0.005	-0.005	0.003*	0.003*	-0.002	-0.002
	(0.017)	(0.017)	(0.002)	(0.002)	(0.010)	(0.010)
Treatment Dummy ($\hat{\alpha}$)	0.403***	0.292***	0.587***	0.457***	0.091***	0.063***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.004)	(0.004)
Period Dummy ($\hat{\beta}$)	-0.016***	-0.016***	-0.004	-0.004	0.005*	0.005**
	(0.004)	(0.004)	(0.005)	(0.004)	(0.002)	(0.002)
Young Prop. ($\hat{\theta}_y$)		-0.001		0.001		0.002***
		(0.001)		(0.001)		(0.001)
Low-income Prop. ($\hat{\theta}_i$)		0.068***		0.088***		0.018***
		(0.002)		(0.002)		(0.001)
Minority Prop. ($\hat{\theta_m}$)		0.111***		0.120***		0.027***
		(0.002)		(0.002)		(0.001)
Population ($p\hat{o}p$)		-0.031***		-0.048***		-0.007***
		(0.001)		(0.001)		(0.000)
Year FE	Y	Y	Y	Y	Y	Y
Season FE	Y	Y	Y	Y	Y	Y
County FE	Y	Y	Y	Y	Y	Y
Linear time trends	Y	Y	Y	Y	Y	Y
R^2	0.018	0.045	0.020	0.046	0.001	0.003
Observations	5,215,582	5,215,582	13,152,449	13,152,449	13,152,449	13,152,449

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Robust standard errors reported in parentheses and clustered by geocode x month pair. "Prop." is the proportion of individuals in the region in a demographic (i.e. young, low-income, or a minority).

Table 7: Crash Analysis: Robustness Check w/Placebo Dates

	Dependent variable: Weekend late night crashes per day			
	OLS (2012 — 2013) (1)	Controls (2012 — 2013) (2)	OLS (2017 — 2018) (3)	Controls (2017 — 2018) (4)
Diff-in-Diff Estimator ($\hat{\tau}$)	-0.020 (0.025)	-0.020 (0.025)	0.038 (0.032)	0.038 (0.032)
Treatment Dummy ($\hat{\alpha}$)	0.245*** (0.011)	0.166*** (0.012)	0.232*** (0.011)	0.154*** (0.011)
Placebo Dummy ($\hat{\beta}$)	0.009** (0.005)	0.009** (0.004)	0.007 (0.005)	0.006 (0.005)
Young Prop. ($\hat{\theta_y}$)		0.004** (0.002)		0.004** (0.002)
Low-income Prop. ($\hat{\theta_i}$)		0.048*** (0.003)		0.048*** (0.003)
Minority Prop. ($\hat{\theta_m}$)		0.081*** (0.005)		0.081*** (0.005)
Population ($p\hat{o}p$)		-0.014*** (0.002)		-0.014*** (0.002)
Year FE	Y	Y	Y	Y
Season FE	Y	Y	Y	Y
County FE	Y	Y	Y	Y
Linear time trends	Y	Y	Y	Y
R^2	0.004	0.009	0.004	0.009
Observations	5,215,582	5,215,582	5,215,582	5,215,582
Treatment dates	Placebo (2012 — 2013)	Placebo (2012 — 2013)	Placebo (2017 — 2018)	Placebo (2017 — 2018)

***p < 0.01, **p < 0.05, *p < 0.1

Robust standard errors reported in parentheses and clustered by geocode x month pair. Each observation is the number of crashes for a region for a day. "Prop." is the proportion of individuals in the region in a demographic (i.e. young, low-income, or a minority).

9 Appendix

9.1 MBTA Coverage and Area of Study

The majority of ridership is attributable to the MBTA's four subway lines. This is the primary reason late night service covered over 70% of ridership. Figure 11 shows the breakdown of the MBTA's overall ridership. The MBTA's main service lines and the vast majority of ridership are contained within four counties in Massachusetts: Essex, Middlesex, Norfolk, and Suffolk; these counties were selected for the analysis. Table 8 provides relevant demographic characteristics for each county in Massachusetts and Figure 12 provides a map of Massachusetts which highlights the selected counties.

These counties are further divided for more granularity. Two approaches are considered. The first is to use United States Census block groups, which facilitates the incorporation of demographic characteristics in the analysis, since the American Community Survey data is at the block group level. This division strategy also provides relative consistency in population because block groups are designed to include a population between 2,500 and 8,000 individuals (State, 2019). However, there is stark variation in block size, as small blocks are more than one hundred times smaller than large blocks. Standardized areas would be preferable since distance to public transportation is an important parameter in the individual's choice of transportation.

The second approach, which this paper adopts, is to divide the area using geographic coordinates, creating rectangular regions of 0.01 degrees of latitude in width (approximately 1.10 kilometers) and 0.01 degrees of longitude in length (approximately 0.82 kilometers), yielding rectangles with area of approximately 0.90km^2. Population varies between these rectangles and demographic characteristics no longer have a direct mapping to the census data, as certain rectangles can intersect multiple block groups. However, the area is kept constant throughout each region. Furthermore, the selected counties have similar population densities, so there is limited variation in populations across blocks.

Distance to a stop is a major parameter in the decision to use public transportation, so the second method is most applicable and adopted for the analysis. The rectangular regions are linked to their respective demographic characteristics by using the data from the block group containing the region's center. This assumes the demographics are consistent within each block group, which is likely considering the small population size contained within each block group.

9.2 Crash and Demographic Data

This section provides more information on the crash and demographic data used. A definition of key variables and the balance of the treatment and control groups can be found in Table 9. Each crash is classified into a region-day pair, then the number of crashes within each pair is summed. Note each crash is assigned to the region in which it occurred.[28] Each region-day pair is classified with an indicator variable as a weekend observation or not, and as a late night observation or not.[29] Each observation is also classified with indicator variables as before, during or after the treatment, where the treatment is defined to be from March 28, 2014 to March 20, 2016.

Beyond the indicator variables, the three key demographic variables — proportion of individuals under age 25, proportion of households earning under $30,000 (in 2017 adjusted dollars)[30], and the proportion of individuals identifying as African American, Black, Hispanic, or Latino — are continuous. Figure 13 shows how these are distributed at a tract level.[31] The figure shows young regions are relatively dispersed, whereas low-income regions are focused mostly in the city center and to a lesser extent Essex county. Minority regions are the most clustered, primarily in the city center and with a cluster in Northern Massachusetts.

[28] It would be ideal to have the start and end location of each trip to assign crashes to their region of origin, as having late night service in the rectangle of origin would likely have contributed more towards preventing the crash. However, the approach adopted here can still be used as the rectangle in which the crash occurred must have necessarily been somewhere along the trip, and so this proxy is informative nonetheless.

[29] Late night crashes are defined as between 12:30am and 3:00am. Note this is 30 minutes past the end of the MBTA's late night service as the crash reporting time could be delayed.

[30] This represents the lowest quartile of household income in Massachusetts (Bureau, 2019).

[31] Note this is at a higher level than the analysis, as visualizing these at the rectangle level is not feasible.

To supplement demographic characteristics and provide information on which regions have late night activity and therefore would be more likely to have demand for late night service, a sample of tweets with geographic coordinates in the Greater Boston Area between 12:30a.m. and 3:00a.m. on weekends are collected from the Twitter Sandbox API for each year of study, from 2010 to 2018. An indicator variable informs whether that region had Twitter activity or not.

Table 10 provides correlations between the three demographic variables and the Twitter activity variable. It shows minority populations are correlated with low income households, while age is not correlated with minority populations nor low income households. Having Twitter activity is also not strongly correlated with any of the demographic variables.

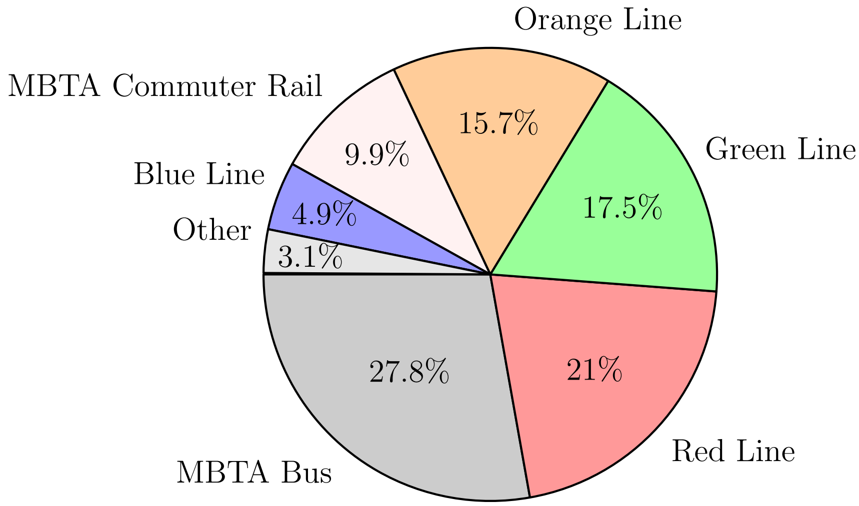
9.3 Standard Error Clustering

I adopt the Abadie et al. (2017) sampling design rationale as this analysis contains a sub-sample of the population of interest and the analysis attempts to understand how the sub-sample informs the impact of late night service more broadly. For this analysis, the sub-sample are the four selected counties and the population is the Boston Greater Area. I cluster standard errors by location and time as there are clusters of both time and location in the population of interest that are not represented in the study's sample. Monthly clustering is selected for the temporal dimension as yearly clustering would not capture the seasonality in crashes and daily clustering would likely not lead to correct standard errors since days are unlikely to be independent.[32] Hence, the monthly level provides the optimal temporal clustering. For the spatial dimension, the county level is unlikely to be insightful as there are only four counties in the sample and they are internally heterogeneous, suggesting correlations between residuals would not be addressed by clustering by county. The rectangular regions allow for this heterogeneity. It is possible there is some correlation between rectangular regions, and that a slightly higher level of clustering is needed. For

[32]Consider a snow storm which limits visibility and impedes travel for several days, for example.

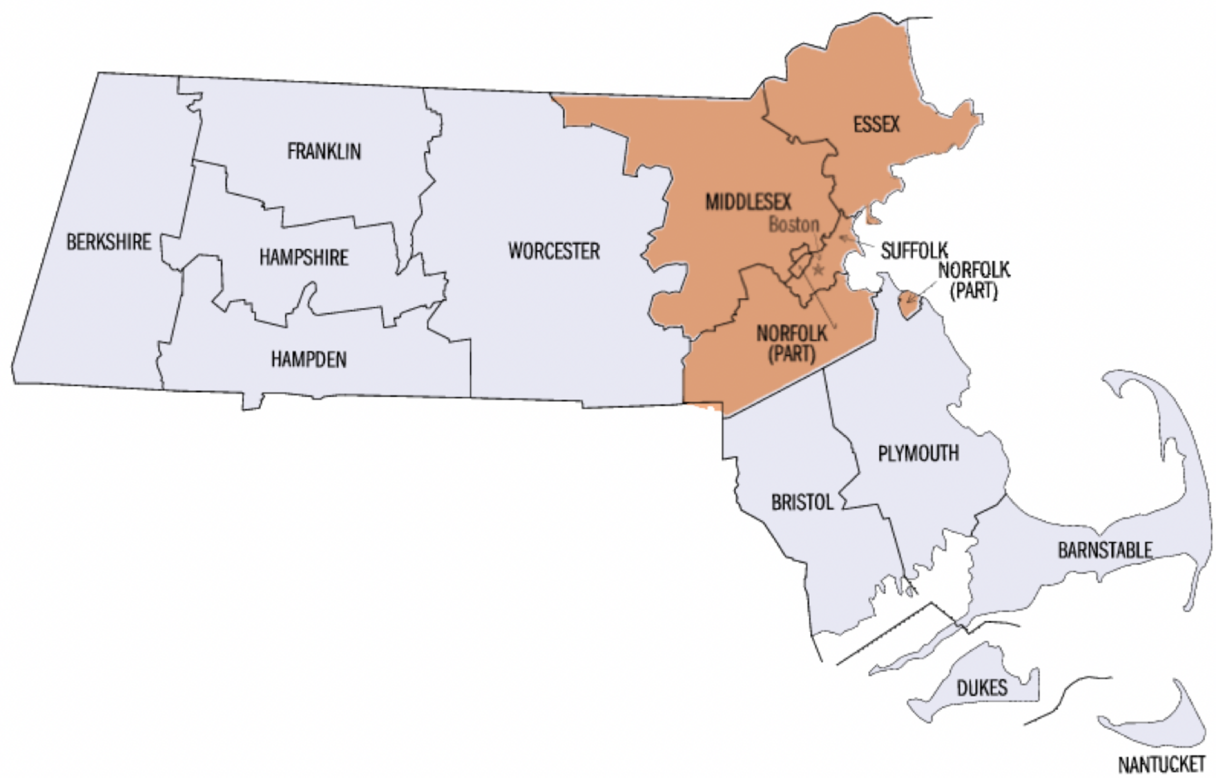
robustness, I test clustering by the 3x3 rectangular region and compare the standard errors to clustering at the rectangle level. There is no substantial difference in the standard errors for the coefficients of interest, so the spatial clustering is done at the rectangular level.

Figure 11: Ridership per MBTA Service



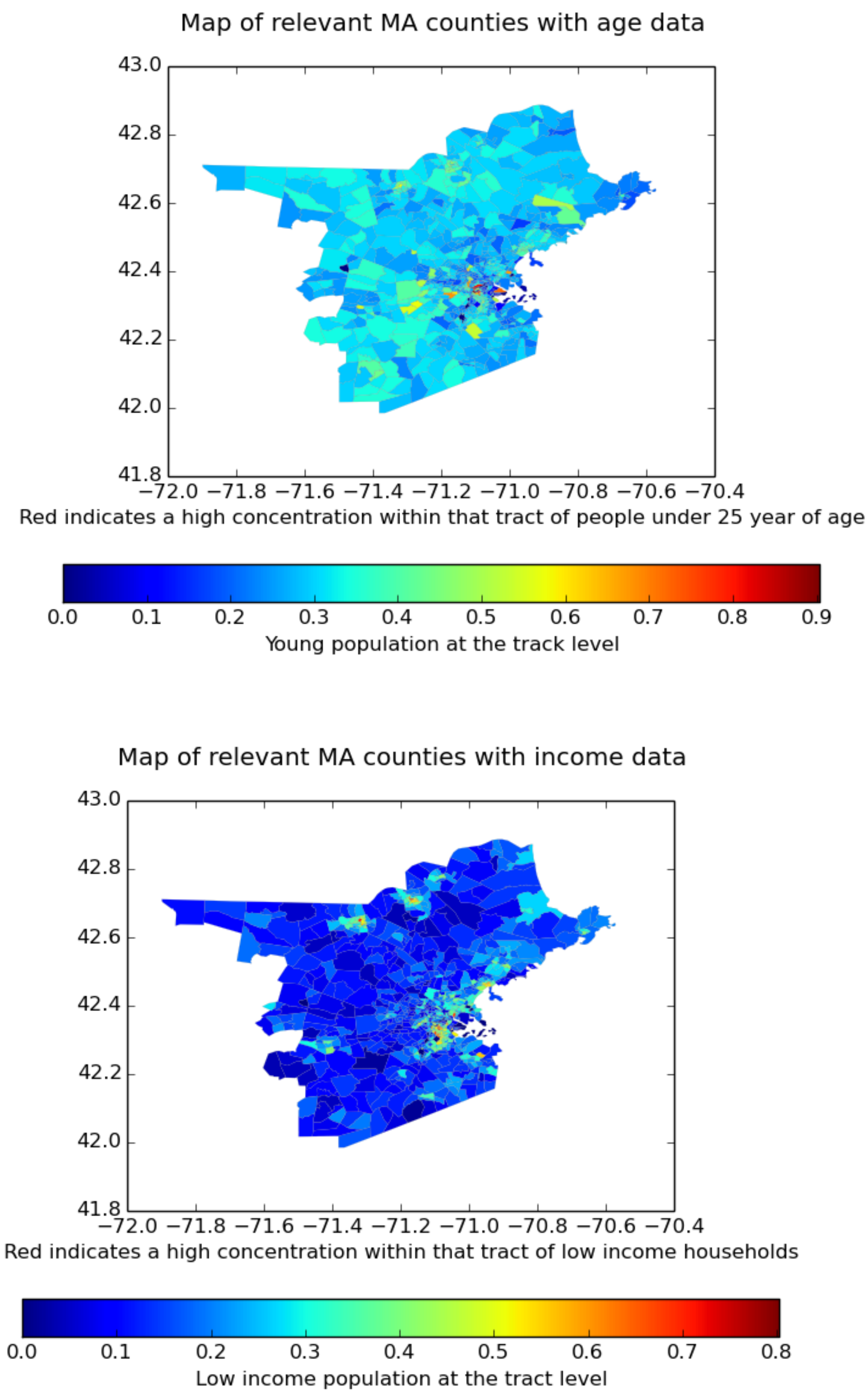
The “Other” category includes the Silver Line, The Ride, and the MBTA Boats.

Figure 12: Map of Massachusetts Counties



The highlighted counties are those selected for the analysis.

Figure 13: Demographic Characteristics



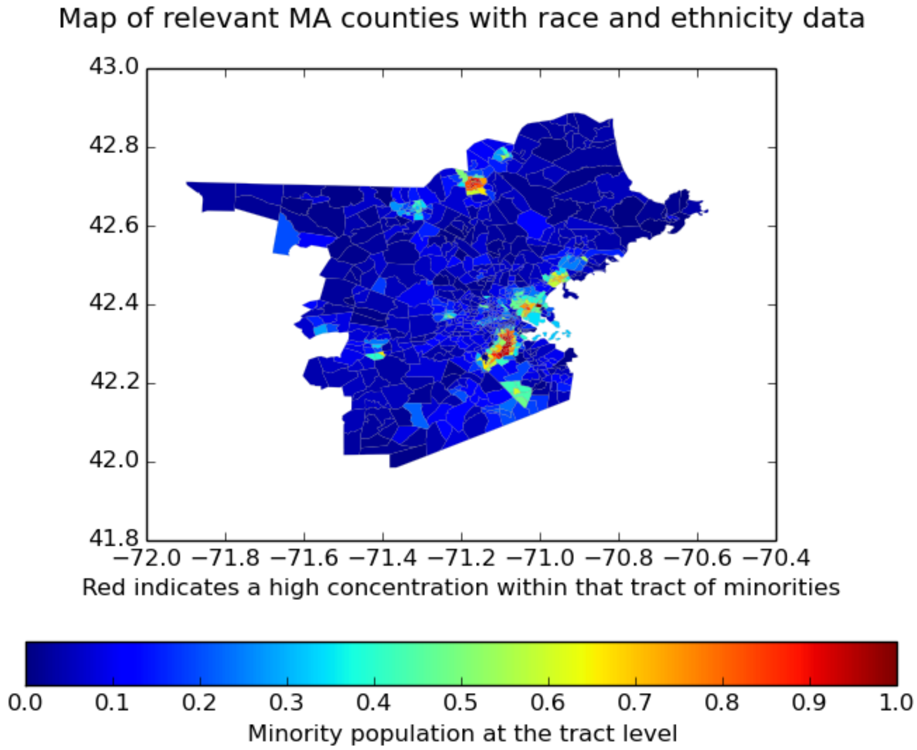

Map of relevant MA counties with race and ethnicity data
43.0
42.8
42.6
42.4
42.2
42.0
41.8
−72.0 −71.8 −71.6 −71.4 −71.2 −71.0 −70.8 −70.6 −70.4
Red indicates a high concentration within that tract of minorities
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Minority population at the tract level

Table 8: Demographics of Massachusetts Counties

County	Density (sq. km)	Median HH income	Poverty rate	Median age
Essex*	583	$75,480	10.9%	40.8
Middlesex*	710	$98,555	8.22%	38.4
Norfolk*	654	$100,829	6.54%	41.2
Suffolk*	5,360	$66,459	19.6%	32.9
Barnstable	210	$68,048	7.49%	52.4
Berkshire	53	$55,190	11.3%	46.5
Brisol	392	$67,842	12.2%	41.2
Dukes	60	$67,535	8.4%	45.9
Franklin	37	$57,307	10.9%	45.9
Hampden	290	$51,726	17.2%	38.8
Hampshire	100	$64,974	13.8%	36.1
Nantucket	82	$91,942	11.2%	39.8
Plymouth	290	$86,910	7.99%	42.9
Worcester	204	$70,402	11.1%	40.5

The first four counties, marked with an asterisk, are those selected for the analysis. Bristol county has similar densities as the selected counties but is excluded due to limited MBTA coverage beyond the Commuter Rail. Population density indicates whether public transport is a viable alternative, as the MBTA services denser areas and routes along denser areas are more likely to have ridership. Median household income and poverty rate are relevant for affordability — the analysis is focused on those who could drive but opt instead to substitute into public transportation. Finally, median age is relevant to ensure the average person is physically and legally able to choose between driving or taking public transit.

Table 9: Balance of Treatment and Control Groups and Variable Definitions

	Definition	Treatment Mean	Control Mean	Diff.
Crashes	Outcome variable; number of late night weekend crashes in a region on a day	2.048 (22.727)	0.303 (8.528)	1.745
Proportion young	Percentage of population in a region under 25 years old.	0.301 (0.147)	0.298 (0.082)	0.003
Proportion low income	Percentage of population in a region under yearly income of under $30,000	0.189 (0.157)	0.122 (0.092)	0.067
Proportion minority	Percentage of population in a region identifying as Black, African-American, Hispanic, or Latino	0.237 (0.263)	0.066 (0.102)	0.171
Twitter activity	Indicator variable with value 1 if a region has Twitter activity in the late night weekend, 0 otherwise	0.337 (0.473)	0.056 (0.230)	0.281

Standard deviations are in parenthesis. Difference is treatment minus control. The means and standard deviations of the "Crashes" variable are multiplied by 100 to facilitate reading the values. All demographic variables are subsequently normalized to be measured in standard deviations.

Table 10: Correlations of Demographic Characteristics

	Young Prop.	Minority Prop.	Low-inc. Prop.	Has tweet
Young Prop.	1	0.076	-0.074	-0.012
Minority Prop.	0.076	1	0.429	0.158
Low-inc. Prop.	-0.074	0.429	1	0.165
tweets	-0.012	0.158	0.165	1

"Prop." is the proportion of individuals in the region in a demographic (i.e. young, low-income, or a minority). "Tweets" is an indicator variable of value 1 if the block had a late night tweet and 0 otherwise.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *National Bureau of Economic Research*, 2017.

Anderson, Monica, "Who Relies on Public Transit in the U.S.," *Pew Research Center*, 2016.

Arthur, Christian, "Late night MBTA service has been a success," *University Wire*, 2018.

Attaluri, Prabhakar, Prianka N. Seneviratne, and Massoud Javid, "Modeling Demand for Public Transit Services in Rural Areas," *Journal of Transportation Engineering*, 1997, *123* (3).

Blincoe, Lawrence, Ted R. Miller, Eduard Zaloshnja, and Bruce A. Lawrence, "The economic and societal impact of motor vehicle crashes, 2010. (Revised)," *Report No. DOT HS 812 013. National Highway Traffic Safety Administration*, 2015.

Brenan, Megan, "83% Of U.S. Adults Drive Frequently; Fewer Enjoy It a Lot.," *Gallup*, 2019.

Bureau, US Census, "US Census 2010," *Census.gov*, 2019.

Burton, Paul, "MBTA to Add Late Service Under Pilot Program," *Bond Buyer*, 2013.

Caudill, Barry, Glenda K. Kantor, and Steven Ungerleider, "Driving While Intoxicated: Increased Deterrence or Alternative Transportation for the Drunk Driver.," *Journal of Substance Abuse*, 1990, *2* (1).

CDC, "Motor Vehicle Crash Deaths: Costly but Preventable," *Centers for Diseas Control and Prevention*, 2015.

Ceder, Avishai, *Public transit planning and operation: Theory, modelling and practice*, Butterworth-Heinemann, 2007.

Chen, Jiafeng, Edward L. Glaeser, and David Wessel, "The (non-) Effect of Opportunity Zones on Housing Prices," *National Bureau of Economic Research*, 2019.

Chi, Guangqing, Xuan Zhou, Timothy E. McClure, Paul A. Gilbert, Arthur G. Cosby, Li Zhang, Angela A. Robertson, and David Levinson, "Gasoline prices and their relationship to drunk-driving crashes," *Accident Analysis and Prevention*, 2011, *43* (1).

Dickens, Matthew, "Transit Ridership Report Fourth Quarter 2016," *American Public Transportation Association*, 2017.

DOT, "2015 State Statistical Abstracts," *Department of Transportation Office of Highway Policy Information*, 2015.

_ , "Research Note: Traffic Safety Facts.," *United States Department of Transportation, National Highway Traffic Safety Administration.*, 2019, *DOT HS 812 603.*

Garcia, David L., Mercedes López Sáez, and Jose Francisco Morales, "Influence of information about trip time variability, personal benefits, and environmental harm from cars versus public transportation on the choice of transportation mode," *Anales De Psicología*, 2016, *32* (1).

Hansen, Benjamin, "Punishment and Deterrence: Evidence from Drunk Driving," *The American Economic Review*, 2015, *105* (4).

Jackson, C. Kirabo and Emily G. Owens, "One for the Road: Public Transportation, Alcohol Consumption, and Intoxicated Driving.," *Journal of Public Economics*, 2010.

Kassiel, Ira, "Late-night service cut by MBTA.," *University Wire*, 2018.

MBTA, "Ridership and Service Statistics," *Massachusetts Department of Transportation*, 2019.

Niederdeppe, Jeff, Rosemary Avery, and Emily N. Miller, "Alcohol-control public service announcements (PSAs) and drunk-driving fatal accidents in the United States, 1996–2010," *Preventive Medicine*, 2017.

Parks, Jason, "How Much Do Car Accidents Cost the Average Citizen?," *AutoKnow – SafeAuto Blog*, 2012.

Ross, H. Laurence, *Confronting drunk driving: Social policy for saving lives* , Yale University Press, 1992.

Sourcebook, "Sourcebook of Criminal Justice Statistics," *School of Criminal Justice, Hindelang Criminal Justice Research Center*, 2007.

State, Michigan, "LibGuides: Finding Census Tract Data: About Census Tracts.," *Michigan State University Libraries*, 2019.

Yao, Xiaobai, "Where are public transit needed – Examining potential demand for public transit for commuting trips," *Computers, Environment and Urban Systems*, 2007, *31* (5).

Ying, Yung-Hsiang, Chin-Chih Wu, and Koyin Chang, "The effectiveness of drinking and driving policies for different alcohol-related fatalities: a quantile regression analysis," *International Journal of Environmental Research and Public Health*, 2013, *10.*