Economics, Politics and Regional Development
ISSN 2576-2060 (Print) ISSN 2576-2052 (Online)
Vol. 7, No. 1, 2026
www.scholink.org/ojs/index.php/eprd

Original Paper

How Does Public Transit Connectivity Affect Voter Turnout? The Case of US

Samuel Kye Wei Voon[1] & Jan P Voon[2]

[1] The College Degree, Bachelor of Arts in Economics and Political Science, The University of Chicago, Illinois, USA

[2] Department of Economics, Faculty of Social Science, Lingnan University, Tuen Mun, Hong Kong

Received: December 18, 2025 Accepted: December 24, 2025 Online Published: February 10, 2026

doi:10.22158/eprd.v7n1p1 URL: http://dx.doi.org/10.22158/eprd.v7n1p1

Abstract

This paper utilizes three methodological approaches to explore the relationship between public transit connectivity and voter turnout. The first approach compares the cases of Los Angeles County, where public transit was free on the day of the 2020 presidential election, and Alameda County, where there was no such program. The second approach utilizes a multi-linear regression to quantify the relationship between public transit connectivity and voter turnout. The third approach makes use of Random Forest, a machine learning model, to make more accurate predictions of voter turnout based on a range of independent variables including public transit connectivity. Overall, this paper finds that there is a positive relationship between public transit connectivity and voter turnout. However, on average across the US, the effects of improvements in public transit connectivity on voter turnout are minute; this paper finds that an average nationwide improvement in public transit connectivity by approximately 40% resulted in a roughly 0.028% average increase in voter turnout across the country. However, in specific cases, the effects can be drastic; for example, in Jackson County, Missouri, there was a 1.65% increase in voter turnout for every 1% improvement in public transit. These results suggest that while public transit can be useful in improving voter turnout, policies that improve public transit connectivity in order to increase voter turnout should be implemented on a case-by-case basis.

Keywords

Public transit, connectivity, voter turnout, public policy

1. Introduction

A good democracy relies on voters translating their lived experiences into political action. However, such democratic political action can only be materialized not just if each vote is valued and taken into account, but if people cast their votes in the first place. This is why voter turnout is an important metric in gauging the legitimacy of a country's electoral system. A democracy implies that everyone has a voice in the makeup of their government and, for everyone to vote, they should not be dissuaded from voting, whether due to costs of voting or a lack of belief in an electoral system. As stated by the Center for American Progress, "for the nation's democracy to function properly and for government to provide fair representation, all eligible Americans must have the opportunity to vote—and be encouraged to do so… collective self-rule is established and fostered through free, fair, accessible, and secure elections through which the voice of every eligible American is heard."[1] While the determinants of voter turnout have been researched extensively in the US, existing explanations of voter turnout often fail to explore a critical feature of a voter's ability to vote: the mode of travel by which they get to the polls. This is a topic that is especially pertinent for democracies around the world considering that in many countries, voters must physically travel to polling places to vote. For example, in Europe, 67% of the 27 countries in the European Union ban absentee voting unless voting from abroad and every single one of the other 16 non-EU countries in Europe bans absentee voting or at least requires a photo-ID to obtain a mail-in ballot.[2] In the US, most voters choose to vote in-person despite the option of absentee voting; 54% of voters in the 2020 presidential election voted in-person either early or on election day.[3] The percentage of in-person voters is even lower in US states where absentee voting laws are more restrictive. In Texas and Mississippi, two of the five states where voters were required to provide a non-pandemic excuse to vote by mail in the 2020 presidential election, mail voting accounted for only 11% and 10% of votes respectively in the 2002 election.[4] With such high rates of in-person voting, it is not extraordinary to imagine that a substantial percentage of voters might encounter problems that might inhibit one's desire or ability to vote when it comes to physically traveling to the polling place. In fact, according to the 2020 Survey on the Performance of American Elections, a nationwide average of 5.4% of people who

[1] Hananel Director, Sam, et al. "Increasing Voter Participation in America." Center for American Progress, 2 Nov. 2021, https://www.americanprogress.org/article/increasing-voter-participation-america/.

[2] Lott, John R. "Why Do Most Countries Ban Mail-in Ballots?: They Have Seen Massive Vote Fraud Problems." SSRN Electronic Journal, 2020, https://doi.org/10.2139/ssrn.3666259.

[3] Stewart, Charles. How We Voted in 2020: A First Look at the Survey of the Performance of American Elections. MIT Election Lab, 15 Dec. 2020, http://electionlab.mit.edu/sites/default/files/2020-12/How-we-voted-in-2020-v01.pdf.

[4] Baseballot. "What Absentee Voting Looked like in All 50 States." FiveThirtyEight, FiveThirtyEight, 9 Feb. 2021, https://fivethirtyeight.com/features/what-absentee-voting-looked-like-in-all-50-states/.

chose not to vote cited transport issues as their main reason for not voting in the presidential election of 2020; this percentage gets as high as 22.3% and 25% in the states of Nebraska and Indiana.[5] As it is inevitable that physical elections are followed with the issue of voters encountering transportation problems that inhibit their likelihood to vote, it would make sense that more accessible and convenient systems of transportation would lead to increases in voter turnout—systems such as mass transit.

Public transportation is a common method of transportation in the US. According to a survey conducted by the Pew Research Center in 2015, 11% of Americans say they take public transportation on a daily or weekly basis.[6] Furthermore, the US census bureau reported in 2019 that 8.7% of households in America do not have access to a vehicle.[7] This represents about 10.5 million households in America. These statistics indicate that there is a sizable portion of the American population that relies on public transit as a primary form of transportation. Consequently, this thesis anticipates that a large proportion of people who do not have access to a car also live in regions without convenient access to any form of public transit, and as a result are completely unable to vote in person. While such a case would be extreme, it is not entirely out of one's imagination to expect voters to be hindered from voting due to polling centers being too costly to travel to. This paper therefore hypothesizes that improvements to public transit access, such as lowering the cost of use, increasing the number and frequency of routes, or improving their speed and route design in relation to polling centers, can help to increase voter turnout.

2. Analytical Approach

This thesis will seek to explore the relationship between public transit access and voter turnout by answering two main questions. Firstly, what effect does improving public transportation have on voter turnout? That is to say, in general, for any given geographical area, does voter turnout increase when public transit connectivity is improved? Secondly, passengers that take public transit likely vary substantially by factors such as income, age, and area of residence. For example, in the US, 21% of urban residents use public transit on a regular basis compared with 6% of suburban residents and 3% of

[5] Stewart, Charles. "2020 Survey of the Performance of American Elections." Harvard Dataverse, Harvard Dataverse, 28 Mar. 2021, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FFSGX7Z.

[6] Anderson, Monica. "WHO Relies on Public Transit in the U.S." Pew Research Center, Pew Research Center, 30 May 2020, https://www.pewresearch.org/fact-tank/2016/04/07/who-relies-on-public-transit-in-the-u-s/.

[7] "ACS Demographic and Housing Estimates." 2015-2019 American Community Survey 5-Year Estimates, United States Census Bureau, 2019, https://data.census.gov/.

rural residents.[8] Hence, if public transport access does lead to an increase in voter turnout, which regions would see the biggest change? In other words, what are the characteristics of counties that have the most to gain from public transit connectivity with regards to voter turnout? Answering these questions will help to highlight how city governments may be able to change their policies on public transit in order to increase voter turnout; this could go on to help to encourage civic participation from the many marginalized social groups that rely on public transit. Such policies can include making public transit free on election days, which is a policy some cities in the US have already adopted, increasing transit coverage, or modifying routes to accommodate those that might not have a means of transportation to get to the polls.

The research question will be answered through three approaches. The first approach is an aggregate-level analysis that will explore the case study of Los Angeles County in the 2020 presidential election where a policy of free public transport was introduced in an attempt to encourage turnout at the polls, comparing it with Alameda County, a county with similar demographic qualities in California. The goal of this approach is to identify a potential relationship between improved mass transit connectivitiy and voter turnout by using a comparative method that will provide suggestive evidence for the argument that that lowering the costs of public transport increases voter turnout. The second approach is to test the findings from the first approach by using a county-level multiple linear regression analysis to quantitatively model the relationship between public transit connectivity and voter turnout in the United States at the county level while accounting for a range of socioeconomic determinants of voter turnout which vary from county to county such as household income, median age, and level of education attainment. This regression will then provide a basis for a predictive model that can predict for voter turnout after an improvement in public transit connectivity. The third approach will improve on the second approach by using Random Forest, a machine learning model that can produce more accurate and robust results, to predict for the effect that an improvement in public transit connectivity has on voter turnout. Through the analyses in the second and third approaches, it is not only possible to gain a more accurate picture as to how public transit connectivity is related to voter turnout, but it would also be possible to gain an insight into the types of places that benefit the most from improved public transit with regards to a positive effect on voter turnout, thereby helping to answer the second question that this thesis raises on the types of regions that are more likely to show higher turnout if transit connectivity is improved.

[8] Anderson, Monica. “WHO Relies on Public Transit in the U.S.” Pew Research Center, Pew Research Center, 30 May 2020, https://www.pewresearch.org/fact-tank/2016/04/07/who-relies-on-public-transit-in-the-u-s/.

Each of the approaches in this study are centered on regions in the US. The US was chosen as the country of focus for this study for three reasons. Firstly, despite concerns of "backsliding",[9] the US is known for its robust electoral system and its democratic values which it champions around the world. A Bright Line Watch survey of 1571 political scientists finds that the United States fully or mostly meets the standard that elections, free speech, and the judiciary, are free and fair;[10] in the US, voter turnout is a pertinent issue that captivates policymakers and constituents alike. Secondly, there is a wealth of data from various databases such as the Center for Neighborhood Technology and the US Census Bureau that make this research project possible. Lastly and most importantly, levels of public transit access vary drastically across regions in the US, which enables close regional comparisons. For example, in 2019, the Metropolitan Transit Authority, which operates most mass transit systems in the New York City Metropolitan Area, recorded a daily ridership of 5.5 million people on its buses and trains. This is in stark contrast to regions such as Jefferson County in Kansas, a rural area without any forms of public transport whatsoever. Thus, America works as a center of focus because it has some of the busiest mass transit systems in the world, places with no public transit at all, and everything in between, which allows for large-scale standardized comparisons between regions with high and low public transit connectivity.

3. Context: Elections and Voter Turnout in America

As this paper is focused purely on voter turnout in the US, it is vital to understand how elections work and how people vote in the US. Elections in America are held for government officials in federal, state, and local offices. Federal elections are the narrowest in scope, featuring the presidential election which happens once every four years towards the end of each presidential term and attracts the largest number of voters and congressional elections which vote in members of the House and the Senate, the components of the legislative branch of the federal government of the United States. State elections take place for state leadership positions, such as the position of governor and state legislators. Local elections are held for the selection of local officials, who make up the vast majority of elected officials in the US. Examples of local officials include mayors, aldermen, and school board members. Most elections in the US use a first-past-the-post system where the highest polling candidate wins the election.[11] As a result of this, the U.S. political system is organized in a way that prefers a two-party

[9] Navarre, Brianna. Study Classifies U.S. Democracy as "Backsliding" for the First Time. 24 Nov. 2021, https://www.usnews.com/news/best-countries/articles/2021-11-24/study-classifies-u-s-democracy-as-backsliding-for-the-first-time.

[10] Bright Line Watch. Bright Line Watch, 23 Feb. 2017, http://brightlinewatch.org/results-from-the-bright-line-watch-u-s-democracy-survey/.

[11] "Voting and Elections." USAGov, https://www.usa.gov/voting.

system. As a testament to this, with the exception of one election, every presidential election since 1852, a candidate from the Republican or the Democratic party, the two predominant political parties in the US, has placed either first or second.[12] This results in party competitiveness and party ID being significant determinants of voter turnout.[13] While numerous elections are held in the US every year, voter turnout differs greatly among them. For example, midterm elections, which are held every 4 years, but staggered 2 years away from presidential elections, have seen regularly lower rates of voter turnout since the 1840s. In fact, while 57.1% of the voting age population voted in 2008 for President Barack Obama, only 36.9% of the same population voted in the midterm elections.[14] This phenomenon therefore partly explains, among other reasons, why this paper will only focus on presidential elections when making comparisons between geographic regions.

Furthermore, to vote in America, one must be registered to vote. The only exception to this is the state of North Dakota. Overall, voter registration rules differ on a state-by-state basis although generally all states require voters to be at least 18 years old, citizens of the US, and have not been convicted by felonies.[15] Voter registration is also almost always conducted in person. Thus, for many people to vote for the first time in an election, voters must not only travel to register but also travel to vote, incurring two times the travel costs. While the advent of mail-in voting has given voters an option to avoid having to travel to the polls to vote, voting in person has traditionally been the predominant form of voting, never falling below 75% until 2020.[16] Only in recent years has the number of in-person voters as a percentage of voters fallen drastically; 46% of voters voted by mail in the 2020 presidential election. Although even with such a drastic increase in the popularity of voting by mail, the bottom line is that the majority of voters in America still vote in person, therefore requiring them to travel to the polls, whether it is due to the need to register or the actual act of voting itself. However, it is important to note that voter turnout differs greatly by region. In fact, civic engagement in general is always in flux;

[12] Blake, Aaron. "Why Are There Only Two Parties in American Politics?" The Washington Post, WP Company, 25 Nov. 2021, https://www.washingtonpost.com/news/the-fix/wp/2016/04/27/why-are-there-only-two-parties-in-american-politics/.

[13] Hofstetter, C. Richard. "Inter-Party Competition and Electoral Turnout: The Case of Indiana." American Journal of Political Science, vol. 17, no. 2, 1973, p. 351., https://doi.org/10.2307/2110524.

[14] DeSilver, Drew. "Voter Turnout Always Drops off for Midterm Elections, but Why?" Pew Research Center, Pew Research Center, 30 May 2020, https://www.pewresearch.org/fact-tank/2014/07/24/voter-turnout-always-drops-off-for-midterm-elections-but-why/.

[15] "Voter Registration Rules." Vote.org, https://www.vote.org/voter-registration-rules/.

[16] Baseballot. "What Absentee Voting Looked like in All 50 States." FiveThirtyEight, FiveThirtyEight, 9 Feb. 2021, https://fivethirtyeight.com/features/what-absentee-voting-looked-like-in-all-50-states/.

according to Rosentone and Hansen in their seminal book on political participation, *Mobilization, Participation, and Democracy in America*, the nature of and frequency of politically mobilizing actions leads to high variances in voter participation rates across different times and places.[17] Additionally, according to Robert Putnam, civic engagement has been decreasing in America due to falling levels of social capital across the country.[18] This variety in civic participation and more specifically, voter turnout, lends itself to the main topic of investigation in this paper which seeks to explain if public transit, which is one factor that is greatly different between regionalities and is also related to how people travel to polls, has a relationship with voter turnout rates.

4. Context: Public Transit in America

The main independent variable that leads this investigation is levels of access to public transport. Public transit in the US can take numerous forms, including not just buses and trains but also forms such as bike sharing systems like Chicago's Divvy bikes, or cable cars like the Telluride gondolas in Telluride, Colorado. In the US, transit that is public means that it is available to the general public and usually requires a fare. Most forms of public transit run according to scheduled timetables although some forms like taxis and shared bicycles can be accessed whenever there is availability. The purpose of introducing or expanding public transit services is to decrease disparities in access to services, employment, and recreation opportunities, especially for people with low incomes, people with disabilities, or the elderly while at the same time reducing traffic congestion and the number of miles traveled by private vehicles.[19] It is important to note that the accessibility and quality of public transit is not the same across the US. Some cities, like New York and Washington D.C., have metropolitan public transportation systems that feature extensive networks of trains, taxis, buses, and ferries. On the other hand, some cities, like Waco, Texas, and Green Bay, Wisconsin, have public transit networks that predominantly rely on buses. Many regionalities don't have any forms of public transit at all. Additionally, public transit connectivity and usage can differ within cities. For example, the city of Chicago has a train service operated by the Chicago Transit Authority. In The Loop, Chicago's downtown area with a population of 42,498,[20] there are 8 stations that serve a total of 28,360

[17] Chapters 1, 2, 6 in Hansen, John Mark and Steven J. Rosenstone. *Mobilization, Participation, and Democracy in America*. Pearson (2002).

[18] Putnam, Robert D. "Tuning In, Tuning Out: The Strange Disappearance of Social Capital in America." *PS: Political Science & Politics* 28, no. 4 (December 1995): 664-683.

[19] "Public Transportation Systems." County Health Rankings & Roadmaps, https://www.countyhealthrankings.org/take-action-to-improve-health/what-works-for-health/strategies/public-transportation-systems.

[20] CMAP Community Data Snapshot | The Loop - Illinois. https://www.cmap.illinois.gov/documents/10180/126764/The+Loop.pdf.

passengers a day. Meanwhile, Wicker Park, a neighborhood with about 53,294 residents[21], is served by only 1 station that sees about 2,638 passengers a day.[22] Furthermore, transit use varies significantly by income, gender, race, and ethnicity; for example, a 1999 study by Garrett and Taylor concluded that, across the US, Hispanic and especially African-American workers have much higher rates of transit usage than non-Hispanic whites, and these differences are particularly pronounced for bus and subway use.[23] However, as demonstrated in the same study, the quality of public transit service can also differ among social groups. Garrett and Taylor found that while transit demand is concentrated in high-density, low-income areas, transport subsidies tend to favor low-density, higher income areas. As public transit services can sometimes be economically inefficient and/or socially inequitable, improved public transit access and connectivity can also come in the form of lower and/or more equitable costs of access, allowing for more people to use public transport and reducing traffic. This is on top of existing ways to improve connectivity including better route design, cheaper fares, higher frequencies, better passenger experiences, etc. How this independent variable will be measured quantitatively will be covered more extensively in the later discussion of AllTransit ratings, a rating system developed by the Center for Neighborhood Technology that evaluates public transit access based on a range of factors including but not exclusive to the number of passengers per capita and the distance between stations and places of work and recreation.

5. Literature Review

In order to parse out this topic's extant literature, it is first necessary to understand that voting has costs. This was critically highlighted by Anthony Downs in 1957, who established that for any individual, when the cost of voting is sufficiently high, it can be rational to abstain from voting.[24] The fact that it is extremely rare for 100% of registered voters to turn out at the polls at a given election indicates that despite the benefits of voting, the costs of voting are a relevant component in one's decision to vote. This analysis was furthered by Brady et al., who found that resources such as time, money, and civic skills, are essential to political activity and therefore voting participation.[25] This not only lends credence to the role that costs play in voting, but also to the possibility for socioeconomic factors like

[21] "Wicker Park Demographics." Point 2, https://www.point2homes.com/US/Neighborhood/IL/Chicago/Wicker-Park-Demographics.html.

[22] November 2021 - Transitchicago.com. https://www.transitchicago.com/assets/1/6/Monthly_Ridership_2021-11(Final).pdf.

[23] Garrett, Mark, and Brian Taylor. "Reconsidering Social Equity in Public Transit." Berkeley Planning Journal, vol. 13, no. 1, 2012, https://doi.org/10.5070/bp313113028.

[24] Downs, Anthony. An Economic Theory of Democracy. Addison-Wesley, 2001.

[25] Brady, Henry E., Kay Lehman Schlozman, and Sidney Verba. "Beyond SES: A Resource Model of Political Participation." American Political Science Review 89, no. 2 (June 1995): 271-294.

income and age, which are linked to how much money and time someone has respectively, to be related to voter turnout. The importance of the role that the costs of voting plays in voter turnout is exemplified by a 1982 analysis conducted by Sigelman and Berry that showed that the impacts of voting benefits pale in comparison to the factor that costs contribute to the voting calculus.[26] This suggests that the costs of voting are a greater determinant of voter turnout than the benefits of voting. Accepting this to be true, then it can be reasonable to question if some of the costs associated with voting are physical in nature and if such costs may have significant impacts on voter turnout. However, while the universe of existing literature on the costs of voter turnout is large, research on the physical hurdles of traveling to polling places and those effects on voter turnout is far less extensive.

Naturally, any exploration of the physical costs of voting should begin with an analysis of the geospatial determinants of voter turnout like the locations of the polling places themselves relative to where voters live and the distances that voters have to travel to vote. The relationship between distance and voting turnout was empirically tested in a 2016 paper by Cantoni reported that a 0.245 mile increase in a region's average distance to polling places decreases voter turnout by between 2% and 5% in the 2012 presidential, 2013 municipal, 2014 midterm, and 2016 presidential primary elections.[27] In the same study, it was also found that the effects of distance on voter turnout are three times larger in high-minority areas than in low-minority areas, although this distinction only occurs in non-presidential elections. Additionally, a paper published in 2005 by Dyck and Gimpel found that voter turnout decreases with distance to about 10 miles whereupon it bottoms out and then begins to increase as people begin to resort to voting through mail.[28] A 2013 exploration by Gerber et al. found that a shift to all-mail voting in the state of Washington not only led to an increase in turnout by 2 to 4 percentage points but also reduced the turnout disparity between lower-participating registrants and frequent voters.[29] Brady and McNulty (2004) looked into both the effect of distance and location disruption on voter turnout. By observing the effect of a reshuffling of polling places in Los Angeles County for the 2003 gubernatorial election on individual voters, Brady and McNulty found that the negative effect that a change in polling place had on turnout was five times larger than the decrease in accessibility due to

[26] Sigelman, Lee, and William D. Berry. "Cost and the Calculus of Voting." Political Behavior, vol. 4, no. 4, 1982, pp. 419-428., https://doi.org/10.1007/bf00986972.

[27] Cantoni, Enrico. "A Precinct Too Far: Turnout and Voting Costs." American Economic Journal: Applied Economics, vol. 12, no. 1, 2020, pp. 61-85., https://doi.org/10.1257/app.20180306.

[28] Dyck, Joshua J., and James G. Gimpel. "Distance, Turnout, and the Convenience of Voting*." Social Science Quarterly, vol. 86, no. 3, 2005, pp. 531-548., https://doi.org/10.1111/j.0038-4941.2005.00316.x.

[29] Gerber, Alan S., et al. "Identifying the Effect of All-Mail Elections on Turnout: Staggered Reform in the Evergreen State." Political Science Research and Methods, vol. 1, no. 1, 2013, pp. 91-116., https://doi.org/10.1017/psrm.2013.5.

an increase in distance to the polling place when distances increased by about a sixth of a mile, but the effects were roughly equal when the distance increased by about a mile.[30] Thus, it is clear that Where polling places are located, in relation to where voters live, has a substantial effect on voter turnout.

While it is understood that geospatial elements that improve the accessibility of voting like keeping voters as close as possible to polling locations, encouraging voting by mail, and keeping polling places the same, are positively related to voter turnout, does the calculus change depending on the means of transportation that people use to get to the polls? And more importantly, does the lack of certain kinds of transportation methods impede voter turnout? In America, the car represents the primary means of transportation for most people. In fact, over 76% of Americans drive alone to work every day[31] and 91% of households have access to a vehicle.[32] A study conducted by de Benedictis-Kessner and Palmer, published as a working paper in 2020, explored the effect of car ownership and electoral participation. By looking at individual level voting and car ownership records in Michigan, it was found that in the 2018 general election, 36% of those without a car voted while 66% of those with a car voted.[33] Not only did this study conclude that, for those who had access to cars, voter turnout increased by 23%-27% and 16%-17% in the general and primary elections of 2018 respectively, it also showed that the time taken to get to polls is longer without a car. More specifically, the median additional time burden on voters without access to a car is about 18.5 minutes.

While de Benedictis-Kessner and Palmer focused on the effect of car ownership on electoral participation, this thesis instead seeks to focus on a topic concerning a different common mode of transportation: the effect of public transit access on electoral participation. This field is understudied, and existing studies are not only scarce but also seem to allude to the lack of a relationship. A brief report published by Bloomberg CityLab looked specifically into whether urban policymakers making public transport complimentary on election days increases voter turnout. Even though the report argued

[30] Henry, Brady, and McNulty John. The Costs of Voting: Evidence from a Natural Experiment. https://www.researchgate.net/publication/247821641_The_Costs_of_Voting_Evidence_from_a_Natural_Experiment.

[31] Tomer, Adie. "America's Commuting Choices: 5 Major Takeaways from 2016 Census Data." Brookings, Brookings, 9 Mar. 2022, https://www.brookings.edu/blog/the-avenue/2017/10/03/americans-commuting-choices-5-major-takeaways-from-2016-census-data/.

[32] Borrelli, Lena. "Car Ownership Statistics." Bankrate, https://www.bankrate.com/insurance/car/car-ownership-statistics/#:~:text=There%20are%20276%20million%20vehicles,secure%20experience%20than%20public%20transportation.

[33] De Benedictis-Kessner, Justin, and Maxwell Palmer. "Driving Turnout: The Effect of Car Ownership on Electoral Participation." SSRN Electronic Journal, 2020, https://doi.org/10.2139/ssrn.3714420.

that the topic of public transit in relation to civic engagement is understudied, its conclusion was that so far, reducing the cost of public transit on election days does not incentivize more voters to go to the polls and that it is still uncertain if policies such as "free transit on election days" are effective.[34] The report cited the fact that even though the Houston Metro system has provided free public transportation on election days since 1992, ridership on the system during voting periods generally increases by around 1%. The report also mentioned another similar program run by Dallas' public transportation network, where after providing free transit on election days for over 30 years, increases in ridership were minimal. However, these are only two examples and may not be indicative of a general trend. For example, the city of Tampa introduced a similar free transit program for election days and saw more success with voters constituting 4% to 5% of election day ridership from 2010 to 2014. Additionally, public transit differs in numerous ways, from ridership demographics to route design. Different public transportation systems around the country may therefore have differing effects on voter turnout depending on who their riders are and how well-connected their routes are.

Another study on whether public transit affects voter turnout found no clear relationship. In an undergraduate study conducted by Rosie Romano which was focused on the city of Washington D.C., voter turnout was compared against the distance of mass transit stops to polling places. In this case, the distance between mass transit stations and polling places acts as a proxy for mass transit accessibility, allowing for an investigation into the relationship between public transport access and voter turnout by using the distance between polling places and mass transit stations as an independent variable. After running a linear regression between the two variables of the percentage of voter turnout in the 2008 general election and the 2010 Democratic primary election and the distance in miles from polling places to the nearest Metrorail station, Romano was not able to reject the null hypothesis that proximity to a Metrorail station has no effect on the rate at which District of Columbia citizens vote.[35] However, distance to polls might not have been a good proxy for public transit access during elections; many other factors go into transit connectivity such as frequency, quality of service, and cost. Furthermore, Romano focused exclusively on the city of Washington D.C. and its public train service. Findings in just this city may not be indicative of a trend across the country. Additionally, a simple linear regression between two variables might not be sufficient in taking into account other variables that affect voter turnout. In the conclusions of both of the aforementioned studies on public transit and voter

[34] Sturgis, Sam. "Could Free Public Transit Get Americans to Voting Booths?" Bloomberg.com, Bloomberg, 3 Nov. 2014, https://www.bloomberg.com/news/articles/2014-11-03/could-free-public-transit-get-americans-to-voting-booths.

[35] Romano, Rosie. "Increasing Voter Turnout: Can Mass Transit Help?" Inquiries Journal, Clocks and Clouds, 1 Jan. 2012, http://www.inquiriesjournal.com/articles/1618/increasing-voter-turnout-can-mass-transit-help.

turnout, the authors argue for a need for greater study into the field. This thesis thus seeks to address this gap in the literature by providing a taking on a different methodological approach to considering the effect that public transit connectivity might have on voter turnout.

6. Aggregate Level Analysis: Context

This paper's first approach in investigating the relationship between transit connectivity and voter turnout is an aggregate-level analysis. Many cities all over the US have implemented policies to reduce the costs of public transit on election days with the goal of encouraging people to vote and therefore increasing voter turnout. Some examples of cities that have tried this in recent elections include Lawrence, Kansas; Goshen, Indiana; and Minneapolis, Minnesota. Some cities like Houston and Dallas, both in the state of Texas, have made public transit free on election days for decades. This phenomenon provides an excellent comparative-style field experiment between counties that have free public transit on election days and counties that do not. By comparing voter turnout between two counties that have had similar trends in the socioeconomic and demographic characteristics that determine voter turnout, the effect of improved public transit access, proxied by a program for free election day public transit, can be isolated for if one of the two counites being compared implemented such a policy.

Although the 2014 Bloomberg article that was discussed in the literature review portion of this study described the relationship between such policies and improvements in voter turnout as uncertain at best, they only looked into changes on ridership numbers but not the actual changes in voter turnout. For example, it is possible that commuter ridership decreased while voter ridership increased on election days. Furthermore, it is important to note that the policy of free transit on election days comes with different restrictions on use in different cities. For example, in Dallas, Tampa, and Houston, voters have to show voter ID or other forms of voter registration proof in order to ride on public transit for free on election days. On the other hand, in the Twin Cities area, public transit is completely free to all users on election days regardless of voter registration status; no checks are required.[36] Factors of these kinds may impact the effect that such policies have on voter turnout.

Furthermore, such kinds of free transit policies can sometimes exist not with the main aim of increasing voter turnout but with the aim of making public transit generally more accessible. If the time frame of these policies, which are the same in design as free election day transportation albeit with different aims, coincide with election days, then theoretically their effect on voter turnout should be the same. An example of this is the Los Angeles County Metropolitan Transit Authority removing public transit

[36] Sturgis, Sam. "Could Free Public Transit Get Americans to Voting Booths?" Bloomberg.com, Bloomberg, 3 Nov. 2014, https://www.bloomberg.com/news/articles/2014-11-03/could-free-public-transit-get-americans-to-voting-booths.

bus fares from March 2020 to January 2022 due to COVID.[37] It is important to note that since 2019, the Los Angeles County Metropolitan Transportation Authority also implements a free transit program on election days which applies to fares on trains, buses, and bike shares. In the next sections that cover one of this paper's methods, the aggregate-level analysis, Los Angeles County will be selected as a specific case study.

7. Aggregate-Level Analysis: Method

This paper's aggregate-level analysis will follow a method employed in a 2008 study by Stein and Vonnahme that aimed to deduce whether the introduction of Election Day Vote Centers, an innovation designed to increase voter turnout by decreasing the costs of voting at relatively inconvenient traditional precinct voting locations, increased voter turnout. From their aggregate-level analysis, they were able to find suggestive evidence that Election Day Voting Centers helped to increase voter turnout. The aggregate-level analysis will be conducted similarly in this study on public transit and voter turnout in two steps. The first step in an experiment of this kind is to select cases. The case selection methodology employed by this aggregate-level analysis will follow the method of the diverse cases as outlined by Gerring and Seawright in their guide to case selection in comparative politics.[38] Like Stein and Vonnahme's study, the aggregate-level analysis in this paper will take place on the geographic level of counties. In the US, counties are administrative divisions of states, constituent political entities of which there are 50 in the US. There are 3243 counties in the US. The county was chosen as the geographic level of analysis as it is the smallest unit of measurement with publicly available data on both the dependent variable of voter turnout and the range of independent variables that helped to ensure that it was reasonable to make comparisons between the two cases. On top of this, while factors such as election competitiveness are highly correlated with voter turnout nationally, on a county-to-county level of comparison, election competitiveness would be the same. Furthermore, many metropolitical transport organizations, municipal organizations responsible for the management and operation of public transportation networks in the US, operate in jurisdictions bounded by county lines. Consequently, the experiment in this paper will seek to compare presidential election voter turnout from 2004 to 2020 between two Californian counties: Los Angeles County and Alameda County. Alameda County will act as a control for the study of a free election day public transit program that was implemented by the Los Angeles County Metropolitan Transportation Authority during the 2020

[37] Lu, Yingtao. "Did Free Metro Buses Bring More Riders?" Crosstown, 1 Dec. 2021, https://xtown.la/2021/12/01/free-transit-los-angeles/.

[38] Gerring, John. "Case Selection for Case‐Study Analysis: Qualitative and Quantitative Techniques." Oxford Handbooks Online, 2009, https://doi.org/10.1093/oxfordhb/9780199286546.003.0028.

presidential election.[39] Why these counties were chosen in particular will be covered in the next section. Additionally, this method will only focus on presidential elections to reduce the exogenous effects attributed to different types of elections. The presidential elections of 2004 to 2020 were chosen in particular because the necessary MIT election return data used to calculate voter turnout in this experiment only goes as far back as 2004. The second step is to track and compare levels of voter turnout between the two counties. This can be done by using two time series on the same chart to visualize voter turnout in presidential elections from 2004 to 2020 in Alameda County and Los Angeles County, respectively. If the improvement in public transit access, represented by the reduced costs of election day public transport, does improve voter turnout, then we would see a spike in voter turnout in Los Angeles in 2020 out of the trends in voter turnout in the two counties.

8. Aggregate Level Analysis: Case Selection

For this study, the case of Los Angeles County will act as the source of useful variation in the dimensions of theoretical interest, which is whether public transportation is free and therefore more accessible. Specifically, with regards to this study, Los Angeles will act as the treatment group where public transit is made free on election days. Los Angeles County was chosen as a case for the study of free election day transit programs because of several reasons. Firstly, Los Angeles County has a large and varied population which is fairly representative of the America as a whole, therefore allowing for inferences to be made of the whole country and subsequently voter turnout in general. Composed of 88 incorporated cities, Los Angeles County is the most populous county in the United States of America with a population of just over 10 million people. The median age in Los Angeles County is 36.5 years, which is approximate to the median age in the US of 38.1. Racial makeup in Los Angeles County is also very representative of the US population. The US is 57.8% white, 18.7% Hispanic, 12.4% black, and 6% Asian. Los Angeles County is 52.1% white, 48.5% Hispanic (many people who identify as Hispanic in Los Angeles County also identify as white), 8.1% black, and 14.7% Asian. Next, public transportation is widely used in Los Angeles County. Run by the Los Angeles County Metropolitan Transit Authority, it is comprised of light and heavy rail lines, bus routes, and a vehicle share system that features both bikes and cars;[40] daily ridership as of 2019 was 1,174,751.[41] Lastly and most importantly, in the context of the diverse case method outlined by Gerring and Seawright, cases should feature two different types of the independent variable of interest. In this study, the independent

[39] Fonseca, Ryan. "La Metro and Other Public Transit Services Are Free on Election Day." LAist, 25 Feb. 2020, https://laist.com/news/ride-public-transit-free-on-election-day-march-3.

[40] "Metro: Bus, Rail, Subway, Bike & Micro in Los Angeles." LA Metro, 7 Feb. 2022, https://www.metro.net/.

[41] "Interactive Estimated Ridership Stats." Metro Ridership, Metro, https://isotp.metro.net/MetroRidership/Index.aspx.

variable of interest is public transit access on election days; therefore, Los Angeles County serves well as a treatment case as it offered free public transit on the presidential election in 2020 but in no other past presidential election.

Meanwhile, Alameda County was chosen as the control case because of its similarity to Los Angeles County in terms of its socioeconomic and political characteristics. For example, its median age is 37.6 years, close to the median of 36.5 in Los Angeles County. Its population is 45.5% white, 12.5% black, 22.4% Hispanic, and 33.7% Asian. This distribution is similar to that in Los Angeles County. Furthermore, as both counties are situated in the same kinds of relatively democratic spheres of influence in California, voters in both countries experience the same kinds of campaigns and candidate messaging that affect turnout and have very similar political attitudes. This is reflected in the party ID proportions in both counties. The people of Los Angeles County have voted for a democratic president at every presidential election since 1988.[42] In fact, as of March 2022, of the approximately 5.7 million registered voters in Los Angeles County, 52.7% claim to be Democratic while 17.1% claim to be Republican.[43] In the most recent presidential election, 71% of voters voted Democratic while 26.9% of voters voted Republican. Meanwhile, the people of Alameda County have also voted for a democratic president at every presidential election since 1988.[44] In the most recent presidential election, 79.83% of voters voted for a Democratic candidate whereas 17.72% of voters voted for a Republican candidate. In addition, Alameda County is also well served by public transportation. Its public transit options include the Bay Area Rapid Transit (BART), which is a heavy rail rapid transit system that serves the San Francisco Bay Area, and Alameda-Contra Costa Transit, a bus service that serves parts of Alameda County. Lastly, the percent of households without a vehicle in Alameda County is 8.8% which is almost the same as Los Angeles County's 9.4%.

While the two counties are similar in many ways, they do have some demographic differences which are likely cause voter turnout to be consistently higher in Alameda County. Namely, median income, which is known to be aligned with an increase in voter turnout,[45] is higher in Alameda County at $99,406 per household compared to $68,406 per household in Los Angeles County. Furthermore, the level of education attainment is also higher in Alameda County than in Los Angeles County. 44.16% of people in Alameda County have a bachelor's degree while only 29.9% of people in Los Angeles County do. However, while this creates some distinction between demographics of the two regions, the

[42] "Presidential Election Results Los Angeles County, 1852 - 2020." Presidential Election Results in Los Angeles County, California, http://www.laalmanac.com/election/el05.php.

[43] Report of Registration - Lavote.gov. https://www.lavote.gov/docs/rrcc/election-info/01042022.pdf.

[44] Politics & Voting in Alameda County, California, https://www.bestplaces.net/voting/county/california/alameda.

[45] Bureau, US Census. "2020 Presidential Election Voting and Registration Tables." Census.gov, 8 Oct. 2021, https://www.census.gov.

year-on-year change in these metrics in each county have not differed by much over the period from 2004 to 2020.

9. Aggregate Level Analysis: Comparison and Results

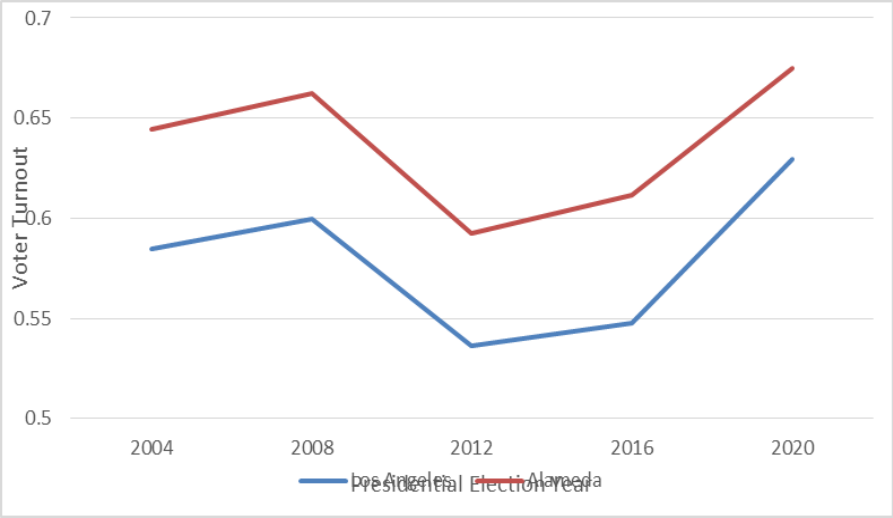

Figure 1. Presidential Election Voter Turnout in Los Angeles County compared with Presidential Election Voter Turnout in Alameda County[46]

Table 1. Voter Turnout at each Presidential Election from 2004 to 2020 in Los Angeles County and Alameda County with Election-to-Election Percentage Changes in Voter Turnout[47]

County	Voter turnout in 2004	Voter turnout in 2008	Percent Change from 2004 to 2008	Voter turnout in 2012	Percent Change from 2008 to 2012	Voter turnout in 2016	Percent Change from 2012 to 2016	Voter turnout in 2020	Percent Change from 2016 to 2020
Los Angeles	58.48%	59.99%	2.58%	53.63%	-10.60%	54.75%	2.09%	62.94%	14.96%
Alameda	64.47%	66.24%	2.75%	59.25%	-10.55%	61.16%	3.22%	67.46%	10.30%

It can be seen from figure 1 that although Los Angeles County's presidential voter turnout, demarcated by the blue line, has been consistently lower than Alameda County's presidential voter turnout, turnout during presidential elections in Alameda and Los Angeles counties have more or less followed the

[46] "Data." Data | MIT Election Lab, 17 Mar. 2022, https://electionlab.mit.edu/data.

[47] "Data." Data | MIT Election Lab, 17 Mar. 2022, https://electionlab.mit.edu/data.

same trend from 2004 to 2016. However, just from eyeballing the graph, it is evident that even though both counties saw an increase in voter turnout in 2020, Los Angeles County saw a significantly larger uptick in voter turnout from 2016 to 2020 than in Alameda County. In fact, as shown in Table 1, the percentage change in voter turnout between the two counties was always within 1.5% of each other. However, in 2020, voter turnout in Los Angeles County increased by 14.96% compared to 10.3% in Alameda County; this is a difference of 4.66%. As Los Angeles County implemented a program of free rides on public transportation in the 2020 election while Alameda County did not, figure 1 and Table 1 seem to lend support to the argument that the free public transit program was successful in helping to increase voter turnout.

10. Aggregate Level Analysis: Discussion

This aggregate level analysis provides some evidence suggesting that improving public transit access has a positive effect on voter turnout. However, these results are suggestive at best and are ecological in nature. Even though the experiment was controlled insofar as diverse cases were selected, the cases, while similar, may have been susceptible to a range of variables that were out of the study's control. Also, in this method (and for the other two methods used in this study), voter turnout for each county was measured by dividing the total of number of votes in each county by each county's citizen voting age population (CVAP). This is different from the voting eligible population (VEP), which may be a more precise denominator in calculating voter turnout. This is because VEP estimates discount felons and other groups that are unable to vote, thus the CVAP underestimated turnout in counties with many felons and other ineligible groups.[48] VEP is not publicly available at the county level, so CVAP was used instead. Furthermore, while it might be true that the free transit program worked in 2020, the lack of repetition for these findings means that it is unknown if the effects of the program can be replicated for any given election; essentially, the analyses in this approach are small N-case studies and therefore do not hold much generalization power. In the end, these results suggest a relationship but are inadequate in proving one. In order to make a causal inference, research at an individual level must be done to determine if improved access to public transit truly affects an individual's decision-making when it comes to whether or not to go to the polls to vote.

While this approach suggests that making public transit free on election days could be an effective program in encouraging civic participation, it is important to note that free transit only on election days may not be as effective in encouraging increases in voter turnout as long-term free transit programs. There is a difference between free transit only on election days and long-term free transit programs. For election-day specific free public transportation, policymakers have to repeatedly get the word out before every election that public transit is free in order for potential riders to know about the program.

[48] McDonald, Michael P. and Samuel L. Popkin. "The Myth of the Vanishing Voter." American Political Science Review 95, no. 4 (December 2001): 963-974.*

On the other hand, a long-term free transit program would not require repeated messaging; riders may instead get used to such programs, relying more heavily on public transit, and therefore increasing ridership more than if public transit was only free on election days, which are usually a couple days each year.

11. Multiple Linear Regression: Overview

This study will now take on a different approach to exploring the relationship between public transit connectivity and voter turnout by developing a model that will aim to quantify the effect that improving public transit access has on voter turnout. While the aggregate-level analysis focused specifically on one "shock", which is the implementation of free transit on election days, this section of the study will have a much more expansive scope, aiming to determine how a general improvement in public transit will affect voter turnout. This is important as the price of fares is only one facet of public transit access. For example, even if rides are free on election days, if the routes are poorly designed or inaccessible to the kinds of people that are most likely to use public transit, then it could be expected that free transit would still have little effect on voter turnout. Improving public transit access can therefore come in many different forms, including but not exclusive to improving route design such that it is accessible to more people who are in need of public transport, improving the frequency and convenience of public transit services, and making mass transit services faster and more reliable. The numerous factors that affect public transit access and whether people are inclined to use public transit must be taken into account in determining if improving public transport access and the quality of the service itself have any positive effect on voter turnout. While the first method in this study exclusively explored the effect of reducing fare prices on voter turnout, this second method explores the effects of improving connectivity holistically; what if the service was better and would that increase voter turnout?

12. Multiple Linear Regression: Method

This approach will use a multiple linear regression to quantitatively determine the relationship between public transit connectivity and voter turnout. The aim is to quantify the relationship between public transit connectivity and voter turnout while also accounting for a range of other variables that may have an effect on voter turnout. After all, it is rarely the case that a dependent variable can be explained by only one variable. This also applies to voter turnout, which is affected by many different variables. Thus, for this investigation, rather than utilizing a simple ordinary least-squares regression, a multiple linear regression will be used instead. This does require some assumptions on the relationship between the dependent and the independent variable. Firstly, it is assumed that there is a linear relationship between the dependent and the independent variables. Secondly, it is assumed that there is no major correlation between the independent variables. While these assumptions are not proven, they are necessary in order to trust the output of the multiple linear regression.

Each observation in the regression will be a county in the United States, implying 3114 observations. The county was chosen as the level of measurement for the regression as it is the most granular level of analysis possible that is consistent across the dependent and independent variables. This therefore allows for a large sample size. Markedly, while there are 3243 counties in the US, there are only 3114 observations in the regression. This is because of two reasons; firstly, the state of Alaska was not included in the study as in Alaska, election returns are recorded at the borough level instead of the county level. This means that the values for the dependent variable in Alaska are not the same level of measurement as any of the independent variables. Thus, Alaskan counties were not included in the study. Secondly, several observations had issues with data reporting and featured erroneous or impossible data points. These observations were removed. As each observation is a county in the US, this regression will seek to explain the dependent variable of voter turnout at the county level in the presidential election of 2020, measured as the total number of votes in that election for a given county divided by the citizen voting age population (CVAP) for that county. The election of 2020 was chosen as the case for this regression because the AllTransit scores, the independent variable of interest, was dated for 2022, while the rest of the variables were recorded in 2019. Thus, the 2020 election felt like the election that provided the dependent variable with the closest match to the available independent variables. Meanwhile, the independent variables, of which there are 61 in total, are listed below in Table 2. Each variable, whether dependent or independent, has a value for any given observation in the model. These independent variables will also be used in this study's third methodological approach.

Table 2. List of Independent Variables in both the OLS regression, used in this study's second approach, and Random Forest, used in third approach[49]

Variable Number	Variable Name	Variable Description
1	total_pop	The total number of people.
2	pop_under5	The number of people under 5 years of age.
3	pop_10to14	The number of people between 10 and 14 years of age.
4	pop_15to19	The number of people between 15 and 19 years of age.
5	pop_20to24	The number of people between 20 and 24 years of age.
6	pop_25to29	The number of people between 25 and 29 years of age.
7	pop_30to34	The number of people between 30 and 34 years of age.
8	pop_35to39	The number of people between 35 and 39 years of age.
9	pop_40to44	The number of people between 40 and 44 years of age.
10	pop_45to49	The number of people between 45 and 49 years of age.

[49] "ACS Demographic and Housing Estimates." 2015-2019 American Community Survey 5-Year Estimates, United States Census Bureau, 2019, https://data.census.gov/.

11	pop_50to54	The number of people between 50 and 54 years of age.
12	pop_55to59	The number of people between 55 and 59 years of age.
13	pop_60to64	The number of people between 60 and 64 years of age.
14	pop_65to69	The number of people between 65 and 69 years of age.
15	pop_70to74	The number of people between 70 and 74 years of age.
16	pop_75to79	The number of people between 75 and 79 years of age.
17	pop_80to84	The number of people between 80 and 84 years of age.
18	pop_above85	The number of people over the age of 85.
19	median_age	The median age.
20	age_dependency_ratio	The age dependency ratio.
21	old_age_dependency_ratio	The old age dependency ratio.
22	child_dependency_ratio	The child dependency ratio.
23	perc_male	The percentage of males
24	perc_female	The percentage of females.
25	total_hh	The total number of households.
26	perc_hh_with_social_security_income	The percentage of households with social security income.
27	perc_hh_married_couple_families	The percentage of households that are families with married couples.
28	perc_hh_income_under10000	The percentage of households that have an income under $10,000.
29	perc_hh_income_10000to14999	The percentage of households that have an income between $10,000 and $14,999.
30	perc_hh_income_15000to24999	The percentage of households that have an income between $15,000 and $24,999.
31	perc_hh_income_25000to34999	The percentage of households that have an income between $25,000 and $34,999.
32	perc_hh_income_35000to49999	The percentage of households that have an income between $35,000 and $49,999.
33	perc_hh_income_50000to74999	The percentage of households that have an income between $50,000 and $74,999.
34	perc_hh_income_75000to99999	The percentage of households that have an income between $75,000 and $99,999.
35	perc_hh_income_100000to149999	The percentage of households that have an income between $100,000 and $149,999.
36	perc_hh_income_150000to199999	The percentage of households that have an income

		between $150,000 and $199,999.
37	perc_hh_income_above200000	The percentage of households that have an income above $200,000.
38	median_income	The median income.
39	mean_income	The mean income.
40	pop_above18	The number of people above 18 years of age.
41	perc_highschool	The percentage of people who only have high school degrees.
42	perc_somecollege	The percentage of people who only have a degree from some college.
43	perc_bachelorsorabove	The percentage of people who have a bachelor's degree or higher.
44	perc_hh_no_vehicle	The percentage of households with no vehicle.
45	perc_white	The percentage of people who identify as white.
46	perc_black	The percentage of people who identify as black.
47	perc_native	The percentage of people who identify as native American.
48	perc_asian	The percentage of people who identify as Asian.
49	perc_pi	The percentage of people who identify as a Pacific islander.
50	perc_otherrace	The percentage of people who identify as another race.
51	perc_hispanic	The percentage of people who identify as Hispanic.
52	unemployment_Rate	The rate of unemployment.
53	perc_walk	The percentage of people who walk to work.
54	perc_publictransport	The percentage of people who take public transport to work.
55	perc_governmentworkers	The percentage of people who are employed and working as government workers.
56	perc_hh_belowpovertyline	The percentage of households living below the poverty line.
57	perc_republican	The percentage of people who identify as Republican.
58	perc_democrat	The percentage of people who identify as Democratic.
59	perc_other	The percentage of people who identify with another political party.
60	perc_mail	The percentage of people who vote by mail.
61	AllTransit_score	The AllTransit Score.

In Table 2, Variables 2 to 22 are to do with age and the dependency ratios associated with age. Within these variables, variables 2 to 18 provide the age distribution; variables 28 to 39 are to do with income, with variables 28 to 37 providing the income distribution; variables 41 to 43 show levels of education attainment. The variables to do with age, income, and education attainment in each county were chosen because these factors have been shown to be a significant determinant in levels of voter turnout.[50] Age dependency ratios, presented in variables 20 to 22 were also included as economic adversity is known to impede political participation.[51] Gender ratios, shown in variables 23 and 24, were included because women have been shown to consistently turn out at polls at higher rates.[52] Access to a vehicle, depicted in variable 44 as the percentage of households that own a car, was included not only because it has a significant effect on voter turnout,[53] but because it may also help to show which counties are more receptive to public transit and will respond better to improvements in public transit in terms of voter turnout. In the same vein, variables 53 and 54 concern the type of transportation that people take to work which may not shed light on voter turnout but can potentially help to show which counties might be more receptive to public transit. Plus, while some of the variables used in this regression may not necessarily be related not voter turnout, it does not hurt the model to have more variables. Theoretically, the more variables there are, the more that biases associated with endogeneity can be reduced. Variables 45 to 51 are to do with racial makeups as there is a significant difference in levels of voter turnout depending on race.[54] Party ID, demonstrated in variables 57 to 59, was included in the

[50] Bureau, US Census. "2020 Presidential Election Voting and Registration Tables." Census.gov, 8 Oct. 2021, https://www.census.gov/newsroom/press-releases/2021/2020-presidential-election-voting-and-registration-tables-now-available.html#:~:text=High%20school%20graduate%20turnout%20was,that%20were%20not%20significantly%20different.

[51] Rosenstone, Steven J. "Economic Adversity and Voter Turnout." American Journal of Political Science, vol. 26, no. 1, 1982, p. 25., https://doi.org/10.2307/2110837.

[52] Igielnik, Ruth. "Men and Women in the U.S. Continue to Differ in Voter Turnout Rate, Party Identification." Pew Research Center, Pew Research Center, 2 Sept. 2020, https://www.pewresearch.org/fact-tank/2020/08/18/men-and-women-in-the-u-s-continue-to-differ-in-voter-turnout-rate-party-identification/.

[53] De Benedictis-Kessner, Justin, and Maxwell Palmer. "Driving Turnout: The Effect of Car Ownership on Electoral Participation." SSRN Electronic Journal, 2020, https://doi.org/10.2139/ssrn.3714420.

[54] Leatherby, Lauren, and About The Author Lauren Leatherby. "The Role of Race in Voter Turnout." The Journalist's Resource, 17 Dec. 2020, https://journalistsresource.org/politics-and-government/minority-voter-turnout-research/.

regression because, at least in the 2020 election, party ID does have an effect on voter turnout.[55] While levels of mail-in voting, noted in variable 60, was shown in some studies to have little effect on voter turnout,[56] this variable was shown by this study's regression model, covered in greater depth in the next section, to be the independent variable that was the strongest explainer of voter turnout. Therefore, the variable was still included. However, it is important to note that this variable was recorded at the state level; as county level data on mail voting was unavailable but state level data was, the percentage of voters who voted by mail was extracted at the state level and replicated for each observation depending on their state.

Variable 61, the last variable on the list, is the independent variable of interest. This is the variable that quantifies public transit connectivity and is composed of the latest version of the AllTransit rating for each county in the US. Published by the Center for Neighborhood Technology, the AllTransit ratings database is a ratings system that assigns a rating for every region in the US depending on their level of public transit connectivity. AllTransit starts as a comprehensive transit database of the locations of stops and routes and the types and frequency of all scheduled transit services in the US. This database feeds into two key measures; the first measure takes into account the distance of transit from homes, measured by the number of trips on public transport a week the average household can access by walking; the second measure evaluates distance of transit from jobs, measured by census data on the whether people use public transit to get to work. The final output, the AllTransit scores, scores the overall quality of transit as it pertains to the actual use of transit by combining the two aforementioned measures of transit accessibility and incorporating some locational and domestic control variables such as commuters per household, fraction of single family homes, and average block size.[57] Although ratings are constructed the level of the census block group, county-level ratings were used so that each AllTransit score could be matched with the observations of the dependent variable and the other independent variables, which were both measured by county. The county-level Altransit ratings were obtained by taking the average AllTransit rating of all the census block groups in a given county. The scores are measured in an index where the worst rating, applicable to counties without any forms of public transport, is 0, and the best rating, accorded to New York County, New York, is 78. The average

[55] Gramlich, John. "What the 2020 Electorate Looks like by Party, Race and Ethnicity, Age, Education and Religion." Pew Research Center, Pew Research Center, 29 May 2021, https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/.

[56] "Vote-by-Mail Had Surprisingly Little Effect on Turnout in 2020, New Study Shows." Stanford Institute for Economic Policy Research (SIEPR), https://siepr.stanford.edu/news/vote-mail-had-surprisingly-little-effect-turnout-2020-new-study-shows.

[57] Haas, Peter, and Preeti Shankar. AllTransit Methods - CNT. Center for Neighborhood Technology, 27 Mar. 2019, https://AllTransit.cnt.org/methods/AllTransit-Methods.pdf.

AllTransit rating in the US is about 0.8.

Multiple Linear Regression: Results

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.23404 -0.03116 -0.00177  0.02892  0.35142

Coefficients: (2 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.023e+00  1.038e+00   0.985 0.324620
total_pop                       -3.441e-07  1.679e-06  -0.205 0.837606
pop_under5                      -1.010e-06  3.644e-06  -0.277 0.781778
pop_10to14                       1.081e-06  3.664e-06   0.295 0.768066
pop_15to19                      -1.851e-07  4.015e-06  -0.046 0.963232
pop_20to24                       4.953e-06  4.836e-06   1.024 0.305754
pop_25to29                       2.037e-06  4.623e-06   0.441 0.659483
pop_30to34                       6.046e-06  5.514e-06   1.097 0.272897
pop_35to39                      -2.390e-07  5.094e-06  -0.047 0.962578
pop_40to44                       1.588e-06  5.096e-06   0.312 0.755425
pop_45to49                       5.146e-06  5.206e-06   0.988 0.323016
pop_50to54                       5.476e-06  5.771e-06   0.949 0.342823
pop_55to59                       1.102e-06  4.899e-06   0.225 0.821970
pop_60to64                      -1.977e-06  4.684e-06  -0.422 0.673053
pop_65to69                       5.866e-06  5.242e-06   1.119 0.263214
pop_70to74                       5.133e-06  5.413e-06   0.948 0.343136
pop_75to79                      -8.046e-07  5.634e-06  -0.143 0.886439
pop_80to84                       4.453e-06  5.939e-06   0.750 0.453432
pop_above85                      7.168e-07  4.982e-06   0.144 0.885593
median_age                       9.697e-03  5.528e-04  17.541  < 2e-16 ***
males_per_100                   -2.108e-04  5.577e-04  -0.378 0.705534
age_dependency_ratio            -9.987e-03  1.821e-02  -0.548 0.583486
old_age_dependency_ratio         9.029e-03  1.823e-02   0.495 0.620431
child_dependency_ratio           1.465e-02  1.821e-02   0.804 0.421237
perc_male                       -1.526e-01  2.673e-01  -0.571 0.568101
perc_female                             NA         NA      NA       NA
total_hh                         3.511e-07  1.688e-07   2.080 0.037649 *
perc_hh_with_social_security_income  5.695e-03  3.432e-02   0.166 0.868218
perc_hh_married_couple_families  2.233e-01  2.807e-02   7.955 2.50e-15 ***
perc_hh_income_under10000       -1.809e+00  1.026e+00  -1.763 0.077990 .
perc_hh_income_10000to14999     -1.699e+00  1.028e+00  -1.653 0.098473 .
perc_hh_income_15000to24999     -1.896e+00  1.027e+00  -1.847 0.064861 .
perc_hh_income_25000to34999     -1.920e+00  1.025e+00  -1.873 0.061156 .
perc_hh_income_35000to49999     -1.924e+00  1.026e+00  -1.876 0.060737 .
perc_hh_income_50000to74999     -1.866e+00  1.026e+00  -1.819 0.069018 .
perc_hh_income_75000to99999     -1.707e+00  1.025e+00  -1.666 0.095909 .
perc_hh_income_100000to149999   -1.398e+00  1.028e+00  -1.360 0.173782
perc_hh_income_150000to199999   -1.238e+00  1.030e+00  -1.202 0.229509
perc_hh_income_above200000      -1.279e+00  1.037e+00  -1.233 0.217529
median_income                   -1.553e-06  4.442e-07  -3.497 0.000478 ***
mean_income                     -1.338e-06  4.326e-07  -3.093 0.002001 **
pop_above18                     -2.574e-06  4.969e-06  -0.518 0.604505
perc_highschool                 -1.278e-02  3.650e-02  -0.350 0.726216
perc_somecollege                 1.407e-01  3.259e-02   4.318 1.62e-05 ***
perc_bachelorsorabove            4.125e-01  3.841e-02  10.739  < 2e-16 ***
perc_hh_no_vehicle              -6.038e-01  5.182e-02 -11.651  < 2e-16 ***
perc_white                      -4.656e-01  6.486e-02  -7.178 8.87e-13 ***
perc_black                      -4.260e-01  6.408e-02  -6.649 3.48e-11 ***
perc_native                     -5.809e-01  5.952e-02  -9.760  < 2e-16 ***
perc_asian                      -6.542e-01  8.301e-02  -7.882 4.47e-15 ***
perc_pi                         -3.603e-01  1.376e-01  -2.618 0.008891 **
perc_otherrace                  -4.994e-01  7.012e-02  -7.122 1.32e-12 ***
perc_hispanic                   -6.286e-02  1.274e-02  -4.936 8.42e-07 ***
unemployment_Rate               -5.196e-04  5.361e-02  -0.010 0.992269
perc_walk                        9.829e-02  4.495e-02   2.187 0.028830 *
perc_publictransport            -6.762e-03  7.846e-02  -0.086 0.931326
perc_governmentworkers          -1.004e-01  2.267e-02  -4.428 9.84e-06 ***
perc_hh_belowpovertyline        -1.755e-01  4.462e-02  -3.934 8.56e-05 ***
alltransit                       2.339e-03  9.959e-04   2.349 0.018896 *
perc_dem                         1.498e+00  1.418e-01  10.561  < 2e-16 ***
perc_rep                         1.343e+00  1.385e-01   9.701  < 2e-16 ***
perc_oth                                NA         NA      NA       NA
perc_mailin                      1.327e-01  5.812e-03  22.828  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0511 on 3035 degrees of freedom
Multiple R-squared:  0.7515,    Adjusted R-squared:  0.7466
F-statistic:   153 on 60 and 3035 DF,  p-value: < 2.2e-16
```

Figure 2. OLS Regression Output Showing Coefficients in the Column Titled “Estimate”, Standard Error in the Column Titled “Std. Error”, and T-Value, in the Column Titled “T Value”

Once all the values for the dependent variable and the independent variables were compiled, a linear model was fitted for the points. The outputs are as shown in figure 2. According to the regression, the coefficient between AllTransit scores, which is the proxy for public transit connectivity, and voter turnout, is 0.0002339. This coefficient implies that for every increase in AllTransit score by 1, there is a 0.02339% increase in voter turnout. With a standard error for the coefficient of 0.009959%, a t-value of 2.349, and a p-value less than 0.05, these results are robust and statistically significant and allow the null hypothesis to be rejected even though the relationship is not particularly strong. This regression seems to conclude that an improvement in transit connectivity is correlated with very minute improvements in voter turnout; albeit it is an improvement, nonetheless. Other noteworthy significant relationships include the strong correlation between mail-in voting and voter turnout, the relationship between median age and voter turnout, the relationship between educational attainment and voter turnout, and the strong negative correlation between not having a car and turnout out to vote.

The OLS regression was then used as a foundation for a predictive model. This meant using the results of the regression, which modeled the relationship between voter turnout and a range of independent variables, to determine voter turnout when certain variables are changed. In the case of this study, the independent variable that was changed was the AllTransit scores. The results of the predictions are in figure 3 where voter turnout is on the y axis and the x axis shows the increase in AllTransit score.

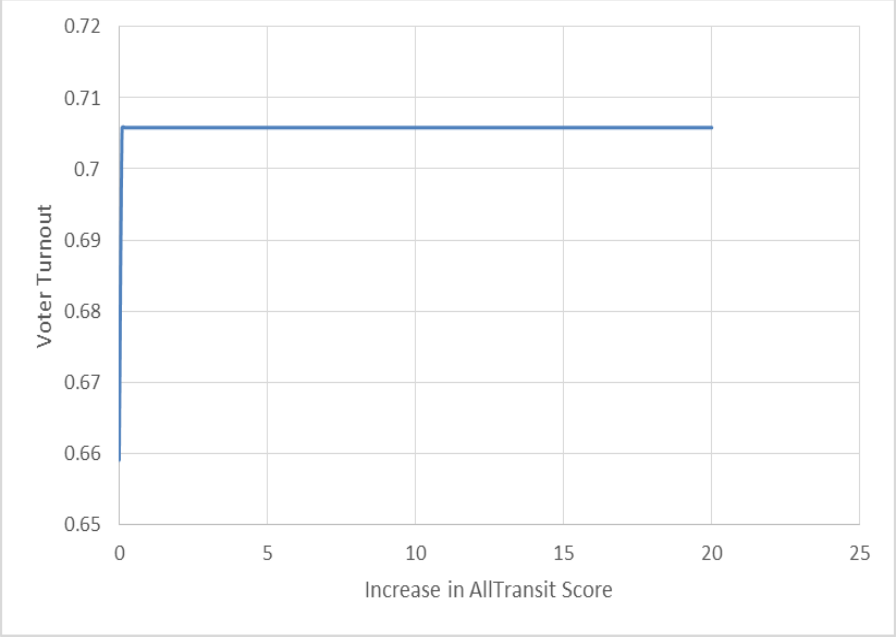

Figure 3. OLS Prediction Output Comparing Voter Turnout Against Increases in Alltransit Score

According to the prediction model that is powered by the OLS regression, any increase in public transit connectivity, indicated by any increase in AllTransit score, leads to an increase in voter turnout by 4.7%. This lends further support to the positive relationship between public transit connectivity and voter turnout and the notion that improvements in transit connectivity increases voter turnout.

13. Multiple Linear Regression: Discussion

This multiple linear regression provides a good quantitative perspective into the relationship between public transit connectivity and voter turnout. It is a powerful model in a sense that it also takes into account the interrelationships between the selected range of independent variables. Furthermore, by utilizing it for a prediction model, it was possible to isolate the effect that an increase in public transit connectivity might have on voter turnout in the midst of many other factors that contribute to voter turnout. One issue with the predictive model is that because the model is linear, it is only able to predict one possible outcome in terms of the size of the change in voter turnout relative to any given change in public transit connectivity. In addition, there are inherent weaknesses in OLS as a method of regression because of its key assumptions. Firstly, it assumes that there is a linear relationship between the dependent and independent variables. Secondly, it assumes no major correlation between the independent variables. However, neither of these are necessarily the case. Furthermore, the data inputs might not be completely exact; for example, the independent variables were obtained from the US Census Bureau were 2019 figures and are therefore not chronologically matched with the values for the dependent variable which comes from the 2020 presidential election. However, this was because the figures for 2019 are the most recent ones available and therefore the analysis was done with those values. Ultimately, these results are still suggestive at best. While they indicate a quantifiable and significant positive correlation between public transit and voter turnout, they do not prove a causal relationship.

14. Random Forest: Overview

The third approach used in this study to evaluate the relationship between public transit connectivity and voter turnout will use Random Forest, a machine learning model, to determine predictions of voter turnout depending on a change in transit connectivity that are more accurate than the ones made in the last model. How Random Forest works will be covered in the next section. This approach will aim to improve upon weaknesses in the previous model in two ways. Firstly, Random Forest does not require the two assumptions that OLS does. Therefore, it can operate freely and detect and suggest a nonlinear relationship albeit cannot prove one. Secondly, it is more accurate and robust as a means of prediction. While the first two approaches in this study helped to identify a positive relationship, this approach will aim to isolate and quantify the exact effect that an improvement in transit connectivity has on voter turnout as accurately as possible via machine learning.

15. Random Forest: Method

This predictive approach will use a regression, built by machine learning, to predict how voter turnout changes with public transit access. Given a number of data points for different variables for a certain county, a machine learning model will compute a regression model that can then be used to predict for voter turnout even if independent variables change. The first step in building the mode is training it. This means running the machine learning algorithm on a set of independent and dependent variables to find the best predictive fit. In other words, once trained, if independent variables were to change, the model would be ready to make a prediction for the dependent variable. Random Forest is the machine learning model that this approach will use. Random Forest is an ensemble learning method that can be applied on regressions by constructing a multitude of decision trees at training time. A single tree is composed of a chosen number of decisions made between randomly selected variables. This process is then repeated a chosen number of times to essentially perform a bootstrap, with each repeat featuring a different random selection of variables. The results of each tree are averaged to arrive at the final result.[58] Figure 4 is a visualization of Random Forest decision trees.

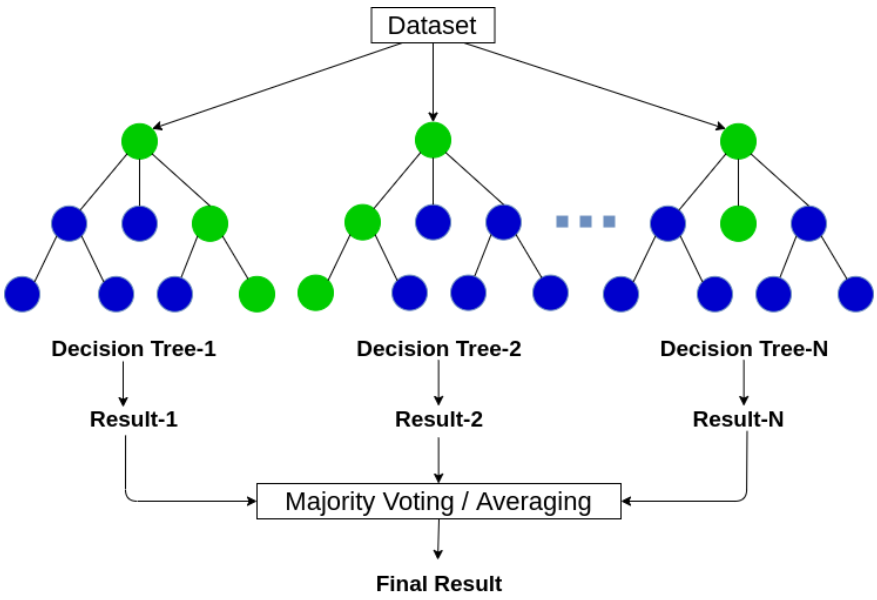

Figure 4. Visualization of a Random Forest

The independent variables that this model will use is identical to the independent variables in the second approach of this study. Therefore, in the context of this thesis, an individual Random Forest tree with three decision levels may predict voter turnout for a certain county based on three independent

[58] Breiman, Leo. "Random Forests." Machine Learning, vol. 45, no. 1, 2001, pp. 5-32., https://doi.org/10.1023/a:1010933404324.

variables randomly selected from the various county characteristics included in the dataset such as age, population, and income. Subsequently, the accuracy of results will come from bootstrapping. An N number of trees will be generated at random to increase the accuracy of results. In reality, in the Random Forest generated for this study, each tree has 8 decisions, which is approximately the square root of the number of independent variables, and 1000 trees were formed. The square root was chosen for the number of decisions as it is the square root of the number of independent variables is the default number of decisions in a Random Forest. Once the model has been trained by developing a model to fit for voter turnout given the selected range of independent variables, the model will predict for voter turnout given average increases in AllTransit scores from 0 to 20. This range was chosen as it demonstrates the largest amount that voter turnout increases by given an average increase in AllTransit scores across the country.

16. Random Forest: Results

The results from the prediction are shown in table 3, where the original voter turnout is shown in the row where there was 0 increase in Alltranit Score. The model found that as average AllTransit scores in the country increased, voter turnout would increase by up to 0.0277 percentage points. This happened when average AllTransit scores are increased by 0.3, around a 40% increase from the current average. After this, the model sees generally diminishing returns as a result of increased average AllTransit scores. Voter turnout even goes so far as to decrease once AllTransit scores increase by more than 8. This trend is graphically represented in Figure 5 where voter turnout is on the y axis and increase in AllTransit score is on the x axis.

Table 3. Increases in Voter Turnout Depending on Increases in AllTransit Score

Increase in AllTransit Score	Voter Turnout
0	65.904%
0.1	65.925%
0.2	65.930%
0.3	65.932%
0.4	65.931%
0.5	65.931%
1	65.928%
1.5	65.924%
2	65.927%
3	65.923%
4	65.920%
5	65.917%

6	65.915%
7	65.914%
8	65.907%
9	65.907%
20	65.899%

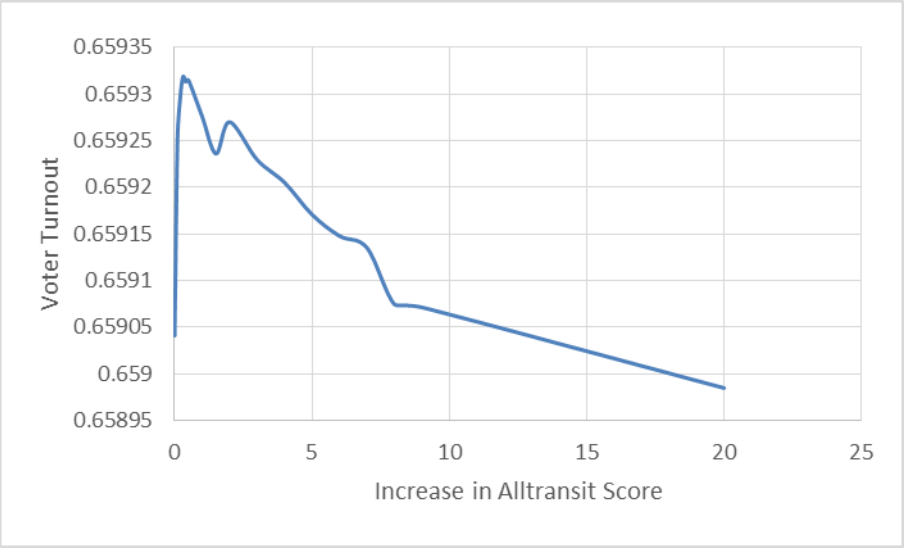

Figure 5. Voter Turnout Compared with Increases in AllTransit Score

The results from this predictive model reinforce the existence of a positive relationship between public transit connectivity, proxied by AllTransit scores, and voter turnout. However, it is unclear as to why increases in AllTransit scores are met with diminishing returns in voter turnout beyond a certain point and even results in decreases in voter turnout. One reason for this may be due to the fact that cities are known to have lower levels of voter participation. In fact, residents of most major American cities typically vote at rates 5 to 15 percent lower than their suburban neighbors.[59] Because cities are where most public transit systems in the US are located, it is likely that the predictive model is assuming that an increase in AllTransit score would cause a county to become more city-like and therefore see lower levels of voter turnout.

Furthermore, because this predictive model gives individual predictions for voter turnout for every county, the outputs also show the counties that benefit the most from an increase in public transit connectivity with regards to voter turnout. This helps to answer the second question posed in this study

[59] Diana Lind, contributor. "Cities Lead the Nation in Many Ways, but Not in Voter Turnout." The Hill, The Hill, 7 Dec. 2016, https://thehill.com/blogs/pundits-blog/campaign/309190-cities-lead-the-nation-in-many-ways-but-not-in-voter-turnout/.

which is: what are the characteristics of counties that have the most to gain from public transit connectivity with regards to voter turnout? By looking at the 10 counties that saw the biggest increases in voter turnout after an average national increase in AllTransit scores by 0.3, chosen as the value that led to the biggest increases in voter turnout across the country, 3 notable socioeconomic characteristics were isolated. This is shown in table 4.

Table 4. Table Showing the 10 Counties That Were Predicted by Random Forest to Have the Greatest Increases in Voter Turnout When Alltransit Scores Were Increased by 0.3

County	Percentage of people above the age of 18 with bachelor's degrees	Percentage of households with no vehicle	Percentage of people below Poverty Line	Original Voter Turnout	Percentage Increase in Voter Turnout After 0.3 Increase in AllTransit Score
Jackson County, Missouri	29.747%	8.899%	10.700%	38.531%	11.574%
Dooly County, Georgia	10.891%	11.135%	21.700%	37.557%	8.272%
Honolulu County, Hawaii	31.913%	9.659%	5.400%	54.139%	8.235%
La Paz County, Arizona	11.145%	6.859%	17.400%	46.467%	8.118%
Rolette County, North Dakota	17.343%	7.791%	20.800%	39.708%	7.802%
Dallam County, Texas	14.794%	2.115%	6.000%	40.941%	7.737%
LaGrange County, Indiana	9.181%	29.444%	4.100%	40.991%	7.722%
Kenedy County, Texas	1.000%	7.614%	2.600%	53.151%	7.471%
Lake County, Tennessee	8.344%	11.547%	25.600%	32.389%	7.218%
Jackson County, Arkansas	9.480%	5.279%	18.200%	37.475%	7.143%
Table Average	14.384%	10.034%	13.250%	42.135%	8.129%
National Average	20.183%	6.138%	10.821%	65.904%	0.028%

The three notable socioeconomic characteristics of the counties that were predicted to have the greatest increases in voter turnout stemming from improvements in transit connectivity are levels of education, proxied by the percentage of people above the age of 18 with bachelor's degrees, the percentage of households with no vehicle, and the percentage of people below the poverty line. Judging from table 4, it appears that counties that have the most to gain in terms of voter turnout from improvements in public transit have generally lower levels of educational attainment, lower percentages of households with no vehicle, and higher percentages of people below the poverty line. Jackson County, which is the county that was shown by the predictive model to have the greatest increase in voter turnout stemming from improvements in public transit connectivity, saw an increase in voter turnout of 11.574% after its AllTransit Score increased by 0.3, a 7% increase for this county. This equates to a 1.65% increase in voter turnout for a 1% improvement in public transit connectivity.

17. Random Forest: Discussion

While Random Forest on its own cannot quantify a relationship or the strength of a relationship as it has no coefficients, it is still a better means of prediction than OLS. Using Random Forest as the basis for a predictive model drastically improves the accuracy of predictions. This improvement in accuracy in evident in figures 6 and 7, which compares the voter turnouts that the OLS and Random Forest models respectively predicted for using existing unchanged independent variables with actual voter turnouts. If the points are closer to the line $x = y$, then the model is more accurate as the mean standard error of the predictions is lower. That is to say that the predictions deviate less from the actual values. Judging from figures 6 and 7, it is clear that the Random Forest predictive capabilities are far more accurate. In fact, mean standard error for the Random Forest predictions was only 1.620 percentage points in voter turnout whereas the mean standard error for the OLS predictions was 3.849 percentage points in voter turnout. However, Random Forest is by no means a perfect predictor. As shown in figure 7, it seemed to consistently under predict and over predict voter turnouts in counties where they were high and low, respectively. Incorporating more elections would increase the number of observations hence potentially making the predictions more accurate as it would have more observations to learn from. However, a predictive approach alone, even via machine learning, is still not enough to prove causation; it is merely what an algorithm thinks voter turnout would be based on large amounts of data.

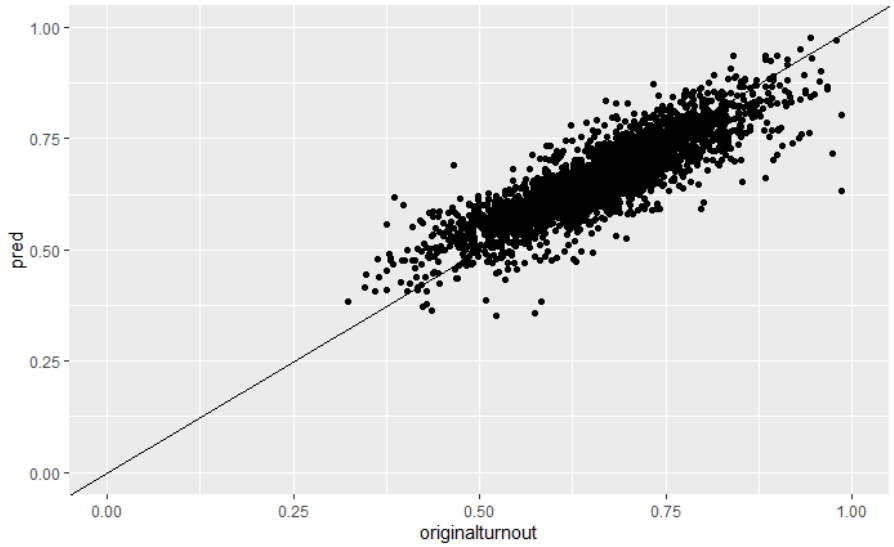

Figure 6. Graph Comparing Original Voter Turnout on the X Axis Against the OLS Predicted Turnout on the Y Axis

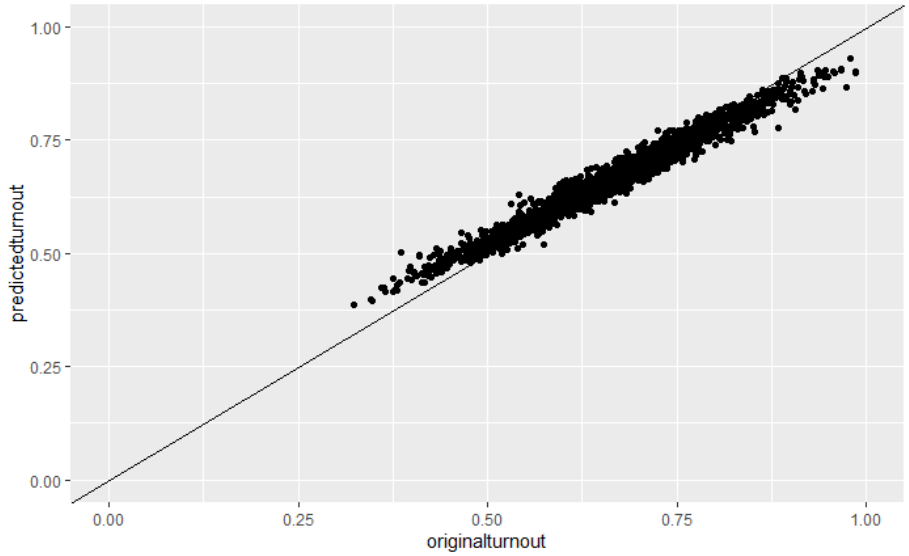

Figure 7. Graph Comparing Original Voter Turnout on the X Axis Against the Random Forest Predicted Turnout on the Y Axis

18. Conclusions

By using three different methodological approaches, this study was able to identify a positive relationship between public transit connectivity and voter turnout. Each of the three methodological approaches had results that are coherent with one another. The first approach compared the cases of Los Angeles County, where public transit fares have been free on election days since 2019, and Alameda County, where fares are not free on election days. This approach found that, between the 2016 and 2020 presidential elections, Los Angeles County saw a noticeably larger increase in voter

turnout relative to the increase in voter turnout in Alameda County. This was at least partly attributed to free election day transit. The second approach, which utilized a multi-linear regression, identified a statistically significant and positive albeit weak relationship between public transit connectivity and voter turnout. The third approach made use of Random Forest, a machine learning tool, to predict for changes in voter turnout depending on various improvements in public transit connectivity. These predictions were more accurate than the ones made using OLS and reinforced the positive relationship between public transit connectivity and voter turnout by showing that a national increase in public transit connectivity by about 40% can increase voter turnout nationally by about 0.028%.

Ultimately, although the proliferation of public transit is one step towards the democratization of our world, just because improving public transportation connectivity increases voter turnout does not mean that it is necessarily something policymakers should pursue. How such policies will be implemented by policy officials still depends on whether the costs justify the benefits. For example, the construction of public transit systems is usually expensive; instead of spending a million dollars on a new subway line, that money could instead be directed to setting up more polling stations which might be a more effective program in increase voter turnout. Therefore, policies that improve transit connectivity and are aimed at increasing voter turnout should be implemented on a case-by-case basis. For example, Jackson County, Missouri, which was predicted by the Random Forest model to be the county that would see the greatest increase in voter turnout when transit connectivity was improved, would have more to gain in terms of voter turnout from better public transit than somewhere like New York County, New York, where the Random Forest Model predicted no change in voter turnout if public transit improved.