Analysis of Bureau of Transportation Statistics on Utilization and Spending Trends

Data 102

Daniel Chung, Alex Cui, Joseph Pang, Helen Rhee

1. Data Overview

Our raw dataset adapted from the US Bureau of Transportation Statistics (BTS) provides national transportation statistics between 1947 and 2021. This data is generated from regulatory and administrative data collections as well as surveys[1]. Participants are thus aware of this data collection in the latter case, although they may be unaware of data collection in other cases like taking public transit. Each row of data represents 136 transportation metrics at the beginning of each month, showing fluctuations in a year, meaning our findings are applicable by month only.

Initially, fewer features were recorded, but over time, more metrics like trade route operation and highway vehicle volume were recorded. This data is a holistic census of US transportation related spending, but there are missing features. For example, the data includes trade route metrics with Canada and Mexico, but excludes Central and Southern America, which comprised almost 250 million nominal dollars worth of goods in 2021[2]. Both Mexico and Canada are more developed than their Central American counterparts, introducing bias through the selective exclusion of less-developed trade partners[3]. Some metrics that would have been helpful include US president political affiliation and the national debt ceiling because they contextualize the US transportation budget constraint.

2. Research Questions

2.1 Prediction with GLMs (Question A)

Can we predict the 2019 State and Local Spending on Transportation in the US from previous general transportation industry statistics, using Bayesian Gaussian regression and random forests?

Answering the question would help government authorities allocate funds for more accurate budgeting for transportation. Through the analysis, they could learn about factors that affect transportation spending most, and prepare in response. The a) Bayesian Gaussian regression and b) random forests are good fits for answering this question because a) we have enough data to inform our prior and b) ensemble methods generalize better to unseen data.

2.2 Causality (Question B)

Does the change in state and local government spending on land passenger terminals (rail investment) have a causal effect on the number of passenger rail passenger miles in the U.S (rail utilization), and if so, what is its magnitude and sign?

This answer can help lawmakers plan how to cause changes in railway utilization. For instance, if increasing spending on land terminals causes a notable increase in passenger traffic, policymakers can advocate for budget changes to increase rail utilization. Causal inference is apt here because the goal is to determine the pure, causal effect of rail investment on rail utilization, beyond correlation and accounting for confounders.

3. EDA

3.1 Data Cleaning

Data cleaning focused on handling null values, which dominated earlier records in the dataset when those metrics were not tallied yet. For question A, with the goal of using historical

[1] Foreign Trade Data Dissemination., accessed. 2021

[2] Foreign Trade Data Dissemination., accessed. 2021

[3] U.S. Trade Representative Developing Country List

data to predict future transportation spending, we condensed the original dataset to cover the years 2005-2019 since instances prior to 2005 were missing over 80% of metrics and instances after 2019 were affected by the global COVID-19 pandemic.

Though condensed, Figure 1 shows that many rows were still missing 2 - 20% of total metrics. As our main objective is predicting transportation spending, we filtered through irrelevant features. Additionally, we plotted the distributions of features that contained missing values and substituted null values with either the mode or median depending on the skewness of the distributions (Figure 2).

3.1 Prediction with GLMs

Upon generating a correlation matrix between the response variable (transportation spending) and all other features, we witness a high association between transportation spending and population in two regards**: a) the number of users utilizing public transit, b) the population/volume of workers in the transportation sector**. Figures 4 and 5 show violin plots of features that had high and low correlations to the response variable, respectively. Both features were categorized into 4 levels of very low to very high based on the number of standard deviations away from the mean. Figure 5 suggests that features with low correlation to the response variable would have a greater variance in their prior distribution, informing the choices for priors in creating a GLM.

Based on the heatmap, it would be interesting to follow up on why transportation spending and mobile transportation features are more highly correlated than other modes of transportation. Understanding the domain knowledge behind why certain features have high or low correlation would improve our model's interpretability.

3.2 Causality

We grouped the 136 total columns by whether they described investment or utilization and then compared how many distinct spending types exist for different categories of infrastructure (see figure 6). This was relevant because we desired to study an area with simple (few forms of) investment, which was displayed in the plot through railways, making them an attractive focus area for studying infrastructure investment and utilization.

This motivated the scatterplot between the quantitative variables of state spending on land passenger terminal (LPT) construction and passenger rail passenger miles (PRPM) in figure 3. We saw there exists a notable correlation (0.48) between the two, suggesting a relationship between rail investment and utilization, further motivating a causal study on railways in particular. In fact, this correlation was one of the strongest out of all investment-utilization correlations for different infrastructure categories compared in figure 7, making it the most promising relationship to investigate further to determine causality.

Instead of dropping rows with null values like we did for GLM analysis, we filled null values with zero and then coded filter steps into our plots that refuse to plot a datapoint if one of its coordinates was exactly zero, as was done in the scatterplot in figure 3, in order to preserve the integrity of the correlation coefficient. Unlike before, this decision does not impact the veracity of our study because missing values are not correlated with anything besides time. All this did was limit the scope of the data we could confidently study.

4 Prediction with GLMs and Nonparametric Methods

4.1 Methods

To predict the 2019 State and Local Spending on Transportation in the US from previous general transportation industry statistics, we created a heatmap to visualize all the 136 features

with our response variable. To account for the large range of spending, all values were log-transformed. For our model, we used 14 metrics where the absolute Pearson correlation coefficient was at least 0.5. Some included features are levels of transit ridership and transportation employment. This was to optimize the accuracy and interpretability of the model while minimizing overfitting.

For our GLM, we chose the Bayesian Gaussian model. First, we chose a Bayesian model because given the real-world application of allocating transportation spending, authorities allocate spending from a distribution of values rather than calculating a fixed number (Frequentist). Also, we have historical data to create an informative prior that accounts for exogenous variables that can impact transportation spending. such as national debt and political atmosphere. Next, we chose the Gaussian distribution to model the likelihood. Though studies suggest that spending tracks a Gamma distribution[4], taking the log of spending data revealed a Gaussian distribution instead.

To fit the priors for the 14 features we selected, we examined discourse on how each metric affected our target variable (transportation spending). For each feature, there were at least two contrasting perspectives quantifying its effect on spending. For example, transit ridership (correlation: 0.86) was strongly associated with large increases in transportation spending due to its ability to measure overall volume of US transportation[5]. Though this strong correlation suggests a lower variance for its respective prior distribution, other studies disagreed, suggesting that ridership was susceptible to uncontrolled variables such as supply chain disruptions[6]. Assuming that the contrasting perspectives hold equal importance for each feature, our prior distributions for the features were standardized to be objective normal distributions with mean 0 and variance 1. This assumes that all the features are equally expected to affect transportation spending, and could be further modified with greater domain knowledge.

For our nonparametric method, we selected the random forest ensemble method because they reduce variance and prevent overfitting through bootstrapping and aggregation. In addition, compared to a neural network, the features of a random forest are much more interpretable and use less computational power. In practical settings, if the city wanted to analyze what factors contribute the most to spending on transportation, a neural network would not be able to provide insight, while a random forest could.

The following hyperparameters were modified: number of trees (150), maximum number of features per tree (5), and maximum depth of tree (10). It is known as best practice to use ⅓ of the number of total features to generate each tree. The increased number of trees fit more to training data, while the limited depth of tree prevents overfitting, and the exact numbers were tuned by examining the mean RMSE of 100 bootstrapped random forest models with different hyperparameters.

Both models' performances were evaluated using root mean squared error (RMSE).

4.2 Results

The Bayesian Gaussian GLM was fit to data of 14 selected features from 2005 and 2018. Visually, the GLM captured the cyclic nature of US Transportation Spending, as seen in Figures 8 and 9. To contextualize the RMSE of the log-transformed values, we converted them back to their original units.

---

[4] Manning, Cost and Generalized Gamma Distributions, 2003

[5] Public Transportation Facts via American Public Transportation Sector, 2021

[6] Transportation and Supply Chain Disruptions, 2007

The training set RMSE (2005-2018 data) was approximately $194,200,157 while the test set RMSE of data in 2019 was $521,960,888. This means that on average, the generalized linear model mispredicted the actual spending amount between 2005 and 2018 by $194M and the spending in 2019 by $520M. Similarly, the random forest showed significantly lower training RMSE compared to test RMSE. The training set error was $64,152,531, while the test set error for random forest was $453,046,908. This is reflected in Figures 10 and 11, where it's seen that though the model is closely fit to training data, it does not capture the local optimum well.

Though these errors suggest a large degree of inaccuracy, it is important to note that total monthly US Transportation Spending is well over $1B. Statistically, this demonstrates that when training the model on data prior to 2019, the GLM mispredicted by around <20% of total spending, which increased to 45% when testing the model on 2019 data.

4.3 Discussion of GLM vs. Non Parametric Model

Figure 13 shows the significant jump from training to test RMSE for the random forest, showing the nature of overfitting. The GLM shows higher test RMSE than the random forest's test RMSE, which may be modified with more informative priors. The broadness of the prior distributions contributed to the high overall degree of error for the GLM.

Figure 12 shows that each model predicts the true spending with higher accuracy for different months. The RF predicts the 2019 log-spending better until mid-June, while the GLM predicts better between mid-June and September. On a month by month basis, the RF tracks the amount of 2019 spending with higher degree of accuracy (hence the lower test RMSE). However, we see that in terms of overall shape, the GLM captures the cyclic pattern of transportation spending better due to the relatively higher variance of the random forest (sensitivity to noise/outliers). Overall, we conclude that the RF was a better choice of model than the GLMwhen predicting future US Transportation Spending. However, with more informative priors, the GLM may capture more information than the random forest.

5 Causal Inference

5.1 Methods

Our treatment variable is whether state and local government construction spending on land passenger terminals (LPT) increased from the previous month, since this is an investment in their use by trains. Z=1 if this spending increased. The outcome variable is passenger rail passenger miles (PRPM). It represents the movement of 1 passenger for 1 mile, reflecting both rail passenger volume and cumulative rail travel distance, and thus describing railway utilization.

This study recognized two confounders in this relationship, namely month and Amtrak on-time performance (OTP). Month is a confounder because seasonal travel increases rail demand, prompting the government to spend more on infrastructure in advance and driving increases in passenger mileage during holidays. Amtrak OTP is a confounder because deteriorating rail efficiency prompts improvement through increased government construction and disincentivizes rail travel for travelers, decreasing passenger mileage. Unfortunately, however, there is not enough data (~50 records) on Amtrak OTP to integrate it into our study.

The unconfoundedness assumption that all confounding variables are observed and can be controlled for does not hold because we lack enough data on the confounding variable of Amtrak OTP to properly observe and control for it.

To handle the confounder of month, we used matching since month is categorical and has only 12 possible values, creating ample opportunities to match treatment-outcome pairs on it. In addition, this is an observational study, so we had no randomization to eliminate the confounding

influence of month. Matching is useful for situations like this because it eliminates the confounding influence by conditioning on it instead.

Rail transportation employment is a collider in this dataset because spending more on LPT construction necessitates more rail transportation staff to maintain them, and the volume of rail passengers as measured through PRPM also necessitates a corresponding amount of rail employees to meet demand. Thus, we avoided conditioning on this collider in the study.

5.2 Results

From matching, we found the ATE estimate is 23.490 million PRPM (see figure 14), which assumes month is the only confounding variable at play since it is the only one our results were conditioned on. The interpretation of this estimate is that increasing state and local government construction spending on LPTs causes an increase of 23.490 million PRPM in a given month, on average. This effect, although subtle, is positive, meaning these results support a positive causal effect of the treatment on the outcome.

This estimate is not without flaws. We performed bootstrapping by resampling records (sample size=34) from the original data to recompute ATE 10,000 times, as seen in figure 15. The bootstrap ATE variance is 2.319 1e16, meaning standard deviation is 4.816 1e7 and standard error is 4.816 1e5. This quantification means the bootstrapped ATE mean is likely very close to the true one, as visualized in figure 16, suggesting ATE is in fact positive on average. An ATE of 0, however, is within one standard deviation of the bootstrapped ATE mean (see figure 15), meaning the causal effect we discovered could still be negative with a non-negligible probability depending on the sample matches, making this positive effect a barely-perceptible one. Our estimate is also imperfect because our assumption that month is the only confounding variable does not hold. We could not control for the other confounding variable of Amtrak OTP because it was non-null in only 50 records, so its confounding influence on both rail investment and utilization is not reflected in our ATE estimate.

5.3 Discussion

Other limitations include our choice to transform the treatment into a binary variable, since it does not capture the degree of investment, only the direction. For example, treatment is still true if spending increases by 100% or 1%, meaning our causal effect is not as descriptive as it could be. Ideally, we would treat rail infrastructure investment as a continuous variable using 2SLS regression, but this requires an instrumental variable, which our dataset lacked.

Having more Amtrak OTP data would have been helpful because we would have been able to condition on it as a confounding variable, which would have accounted for the influence of all confounders in the infrastructure dataset and estimated ATE closer to the true causal effect.

In summary, we believe there exists some relationship between whether the government increases spending on LPTs and how much passengers utilize rail transportation. EDA showed a positive correlation (0.48) between construction spending on LPTs and PRPM, for instance, and the ATE of increasing LPT spending from the previous month is an increase of 23.490 million PRPM. We are, however, uncertain about our specific ATE estimate. The ATE can also be negative within one SD as found via bootstrapping, for instance, which is only countered by the fact that our standard error for the bootstrapped ATE is so low (482K PRPM). We also did not account for the influence of Amtrak OTP, but that influence would have to be around as great as the observed ATE to negate that causal effect from the rail investment alone. Nevertheless, this means our ATE estimate is likely different from the true ATE of increasing railway spending on rail utilization.

6 Conclusions

Using the original dataset from the Bureau of Transportation Monthly Transportation Statistics data, we were able to find all the relevant information to answer both our research questions.

6.1 Prediction with Models

Although the random forest had significantly lower training RMSE ($139,974,493) than the GLM, it performed only slightly better on the test set ($20,153,618). This suggests that the GLM, especially given more informative priors, generalizes better to unseen data and has more potential to improve given more domain knowledge.

Our model and methodology can be applied to different spending categories, such as spending on water supply or power, by modifying which features to use in the model. Both models predict spending with around 10-15% margin of error, and government authorities can use either model to set the range of how much spending should be allocated to a given category. However, our results may not be specific enough to allocate tight budgets or determine the tradeoff between multiple spending categories.

A limitation in the data is that we were not aware of conscious decisions made by policymakers, such as historical or policy changes that impacted spending in certain periods. For example, we do not know why transportation spending is cyclic. In addition, it was difficult to examine the relationships between different spending categories, though in reality, budgeting is done holistically while looking at all spending categories. Conducting additional research into these factors could improve the accuracy of the model by improving the prior distributions. However, another avenue that may be even more useful is the interpretation of the model, and determining which features are directly related to certain spending decisions, so that authorities can expect spending changes based on the events that occur throughout the year.

6.2 Causality

We conclude that an increase in U.S. government spending on LPTs causes, on average, an increase of 23.490 million PRPM in a given month. This ATE has a bootstrapped standard deviation of 4.819 1e7 with standard error 4.816 1e5, meaning that although the causal effect is likely positive, it is extremely subtle. This result is specific to rail infrastructure in the form of land passenger terminals and is thus not generalizable to other areas like air terminals, highways, etc. Because the causal effect we estimated is positive, we advocate for increases in state and local government budgets for the construction of LPTs to promote greater rail utilization by the U.S. public. Because the causal effect is slight, however, we recommend against prioritizing these budget policies if means other than terminal spending are found to boost rail utilization.

This result is limited by insufficient confounder data. As mentioned in our discussion, we lacked enough Amtrak OTP records to condition on that confounder in our analysis, violating the unconfoundedness assumption. Our ATE estimate is thus contaminated with unfactored, confounding influence from Amtrak OPT, which is why we hesitate to put full confidence behind it. We analyzed rail infrastructure, but future studies could compare the causal effects of investment on utilization for infrastructure beyond rail-based land terminals, like air and highway facilities. This matters because comparing such studies would comprehensively inform governments on which facets of infrastructure investment cause the greatest changes in infrastructure utilization. Such insights, furthermore, would allow U.S. government agencies to spend more efficiently to meet infrastructure utilization goals.

7. Visualizations

Figure 1: Percentage of Null Values per Row (2005 - 2021)

Visualized the percentage of null values in each row showing data from 2005 - 2021 using the following formula: $\frac{\#\ of\ null\ values}{\#\ of\ values\ in\ row}$

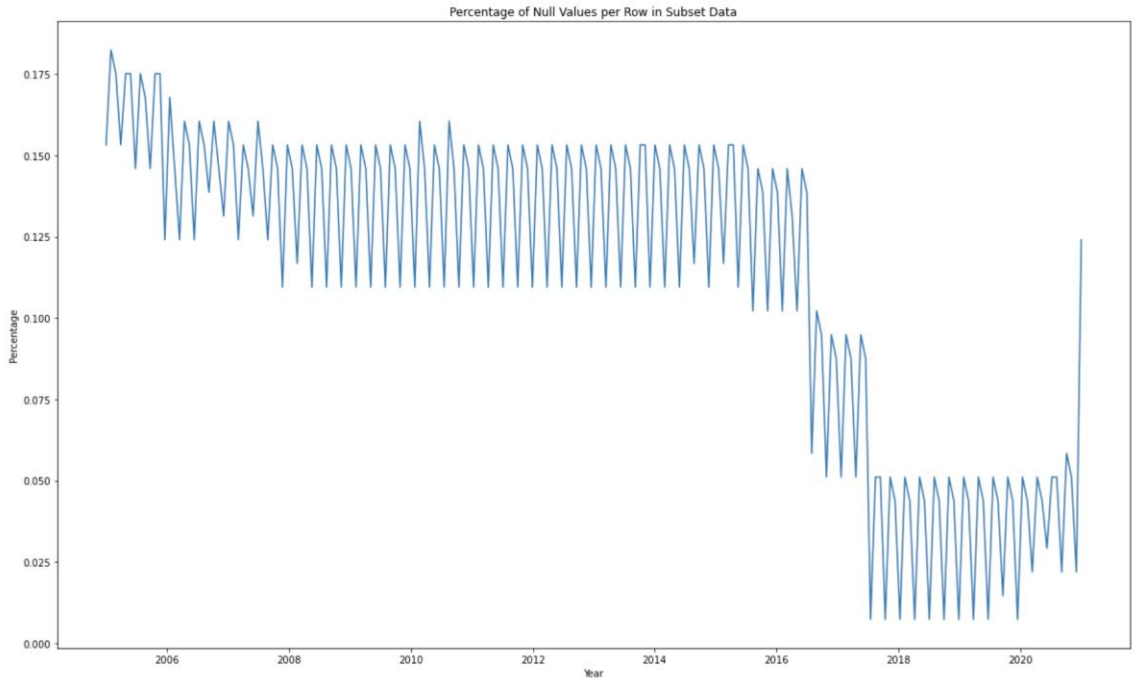


Figure 2: Distribution of Transborder US Freight

An example of plotting the distribution of the features with high percentages of missing values. If the distribution was normal, we used the mean to fill in NaN values, but if the distribution was skewed, we used the median.

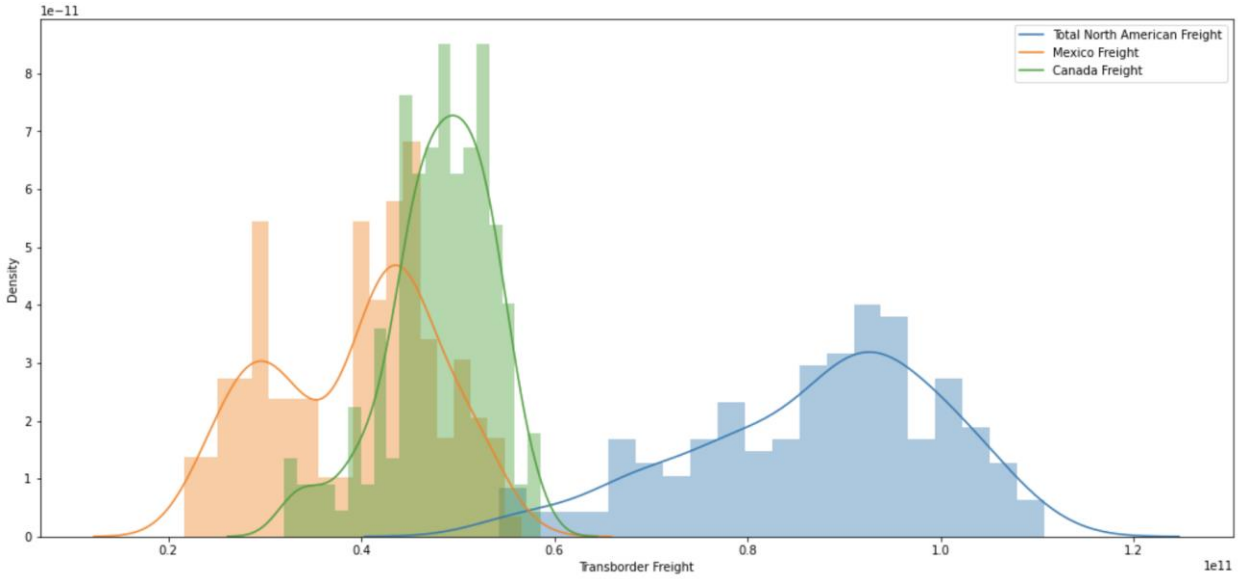


Figure 3: Rail Infrastructure Investment vs Utilization (Corr=0.48)

Used a scatterplot between LPT spending and passenger rail passenger miles during EDA along with a line of best fit to visualize the relationship and correlation between what would become the treatment and outcome variable. The positive relationship motivated us to use these variables.

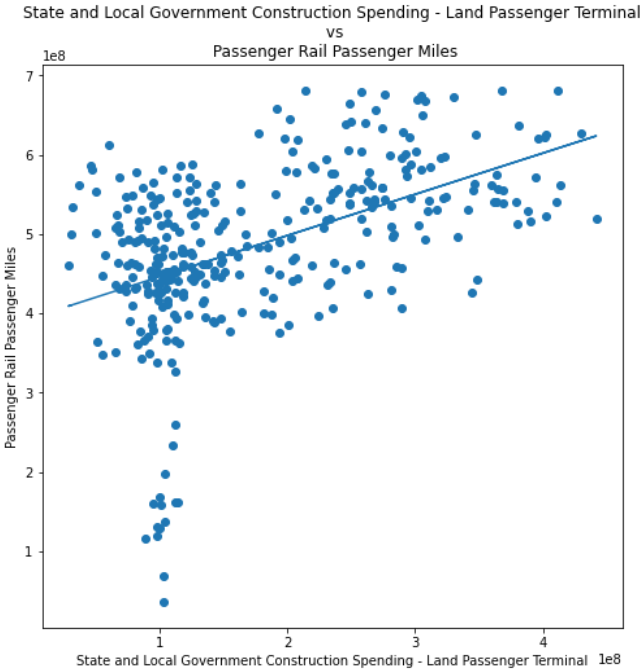


Figure 4: Transportation Spending vs. Levels of Transit Ridership - Other Modes (High Correlation)

Exploratory Data Analysis: Categorized “Transit Ridership - Other Modes” into different tiers according to how many standard deviations a value was away from the mean. Purpose was to see if there would be any trends within each tier of transit ridership.

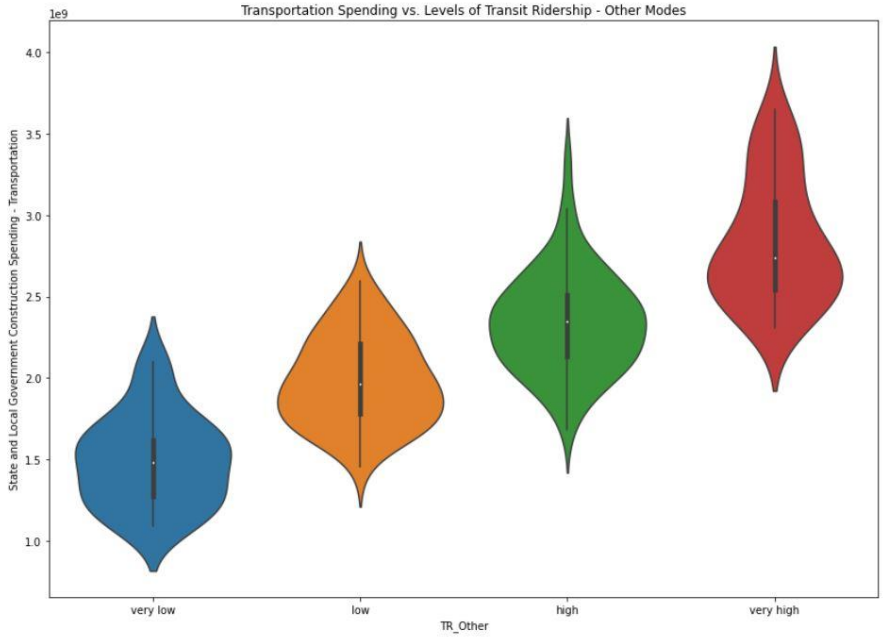


Figure 5: Transportation Spending vs. Levels of Transit Ridership - Fixed Route Bus (Low Correlation)

Exploratory Data Analysis: Categorized "Transit Ridership - Fixed Route Bus" into different tiers according to how many standard deviations a value was away from the mean. Purpose was to see if there would be any trends within each tier of transit ridership.

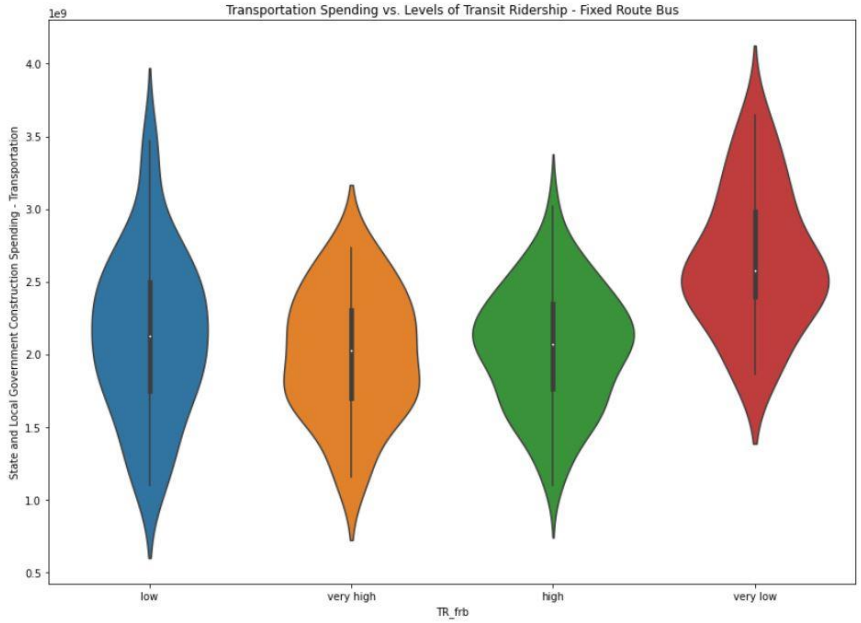


Figure 6: Relative Complexity of Investment by Infrastructure Category

Comparing the count of distinct spending types in each area of infrastructure reveals railways have relatively simple investment data, making rail investment an attractive treatment variable.

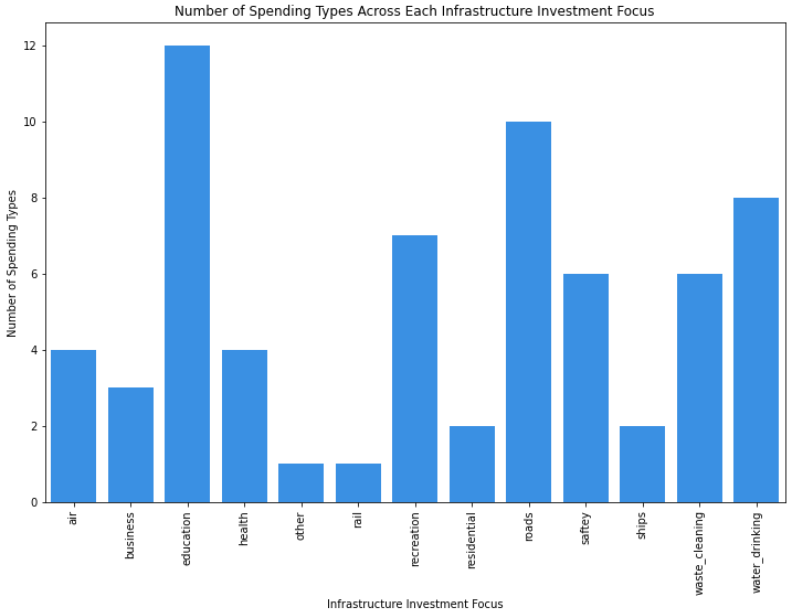


Figure 7: Comparison of Investment-Utilization Correlation by Infrastructure Category

Comparing the correlation between investment (our choice of relevant spending variables) and utilization (our choice of relevant infrastructure traffic variables) between infrastructure categories show rail exhibits a promising (and positive) investment-utilization relationship.

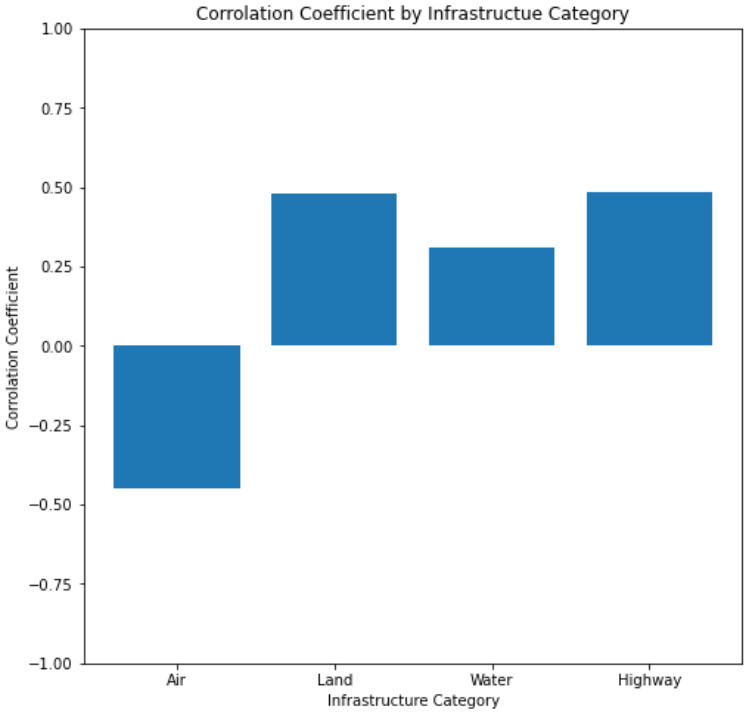


Figure 8: 2005 - 2018 Actual (Red) vs. Predicted (Green) Transportation Spending (GLM)
Plotted the actual transportation spending from 2005 - 2018 (training set) against the values predicted by the GLM to see if the model accurately captures the general trend of spending.

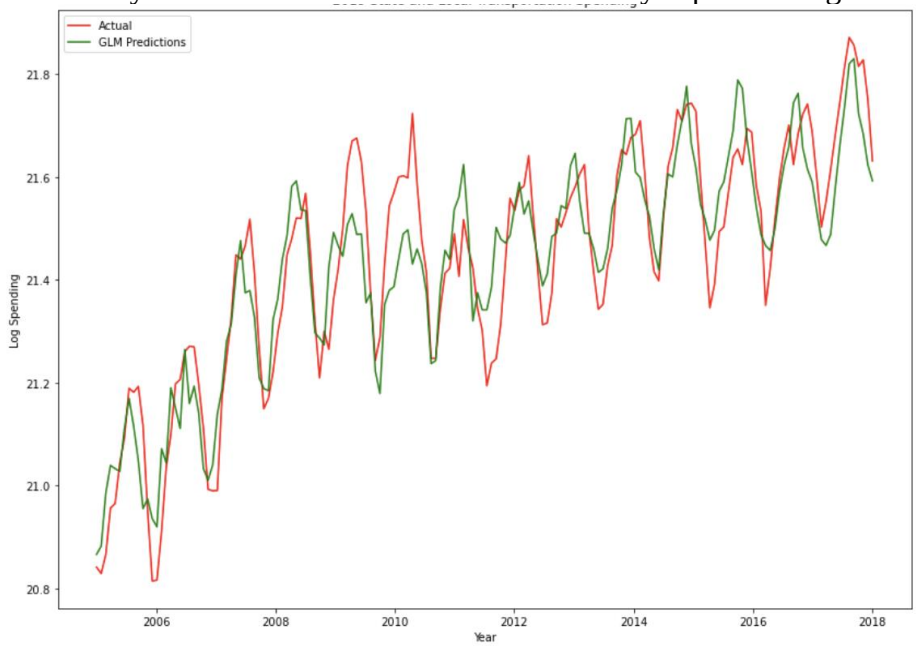


Figure 9: 2019 Actual (Red) vs. Predicted (Green) Transportation Spending (GLM)

Plotted the actual 2019 transportation spending (test set) against the values predicted by the GLM to see if the model accurately predicts 2019 spending.

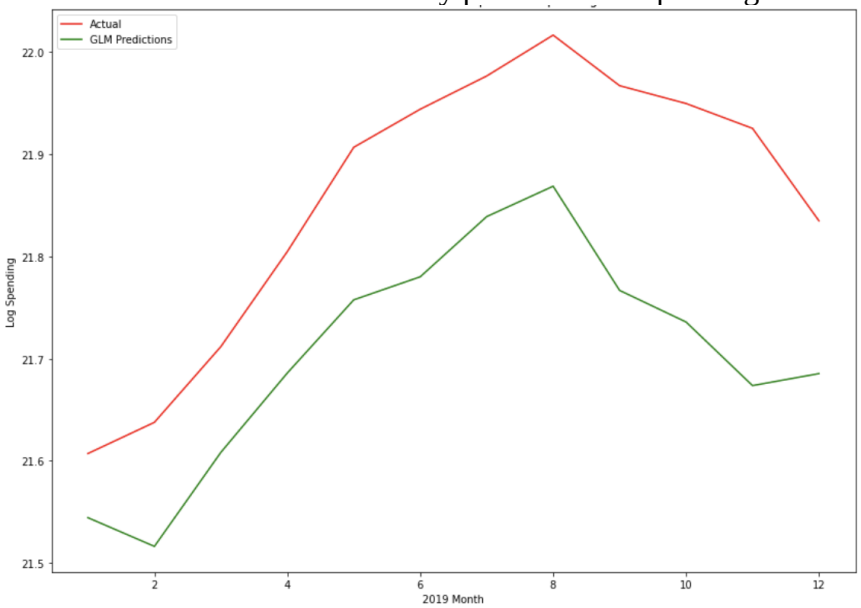


Figure 10: 2005 - 2018 Actual (Blue) vs. Predicted (Orange) Transportation Spending (RF)

Plotted the actual transportation spending from 2005 - 2018 (training set) against the values from the random forest to see if the model accurately captures the general trend of spending.

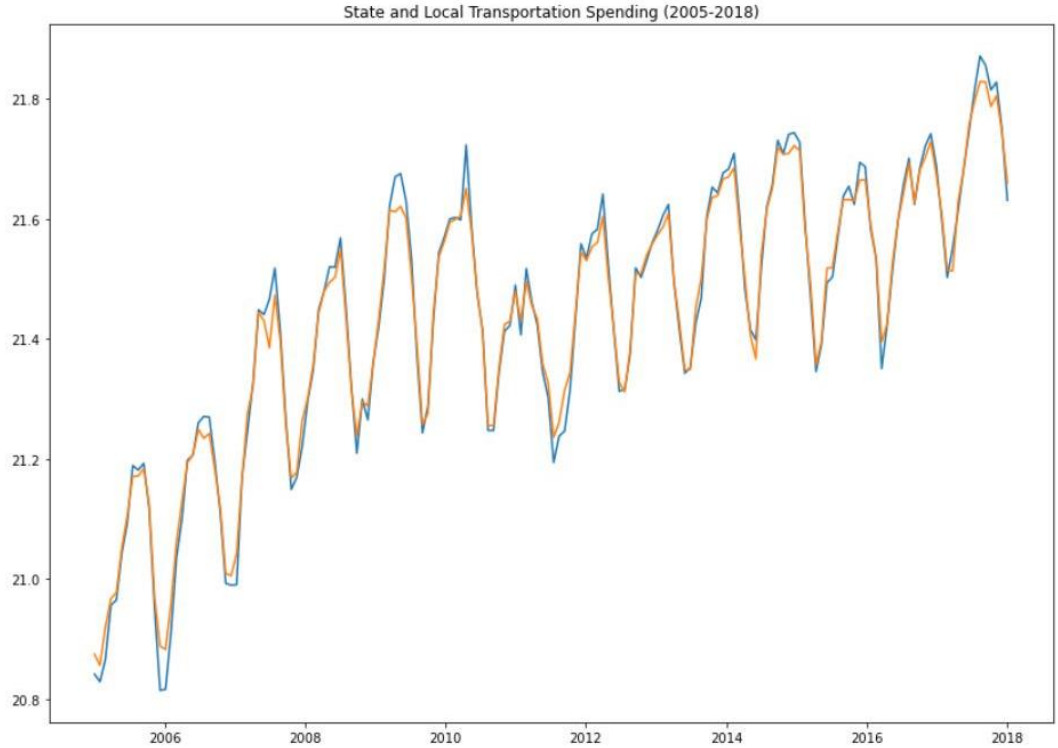


Figure 11: 2019 Actual (Blue) vs. Predicted (Orange) Transportation Spending (RF)

Plotted the actual 2019 transportation spending (test set) against the values predicted by the random forest to see if the model accurately predicts 2019 spending.

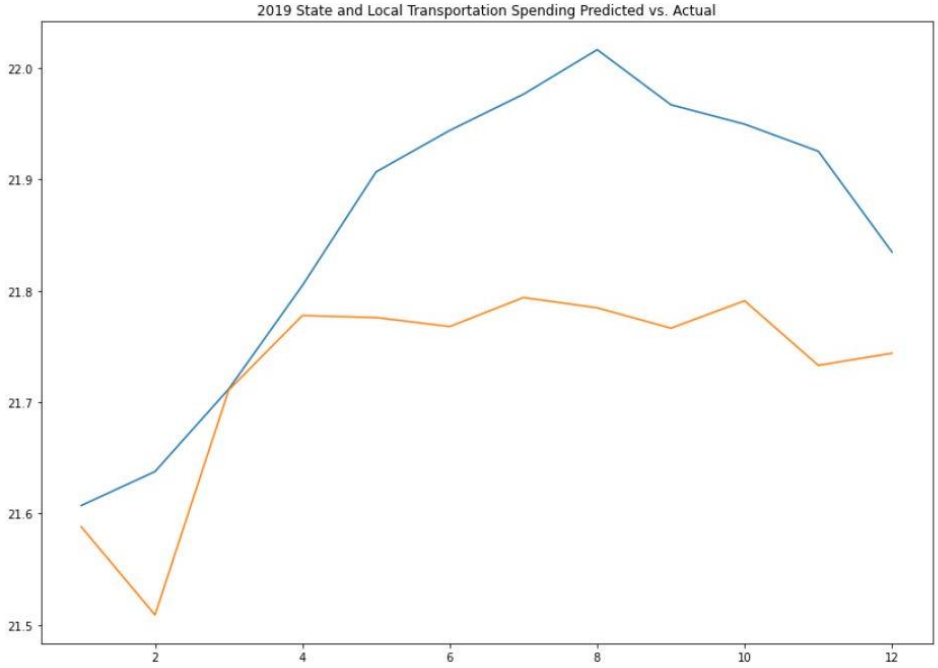


Figure 12: 2019 Transportation Spending - Actual (Blue) vs. GLM (Green) vs. RF (Orange)
Plotted the actual 2019 transportation spending (test set) against the values predicted by the GLM and random forest to compare the models' predictions.

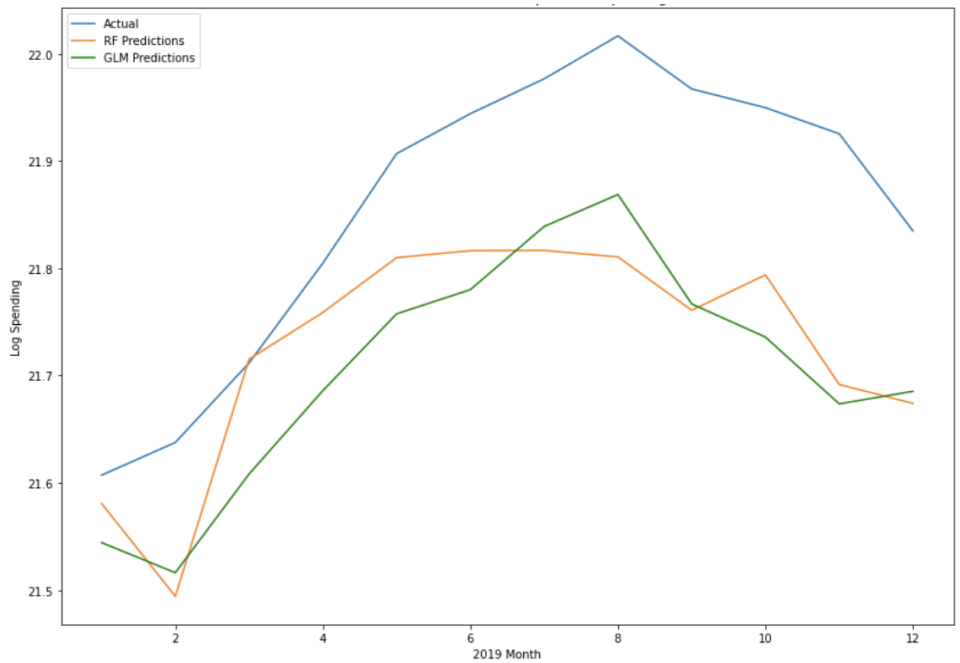


Figure 13: RMSE for GLM vs. Random Forest
A table summarizing the root mean squared error for both the GLM and random forest model. Includes the raw RMSE value in log units, and the RMSE converted to $ units.

	Bayesian Gamma GLM	Random Forest
Training RMSE (Log Units)	0.0904	0.0313
Test RMSE (Log Units)	0.1755	0.1484
Training RMSE (Dollar Units)	$194,200,157	$64,152,531
Test RMSE (Dollar Units)	$521,960,888	$453,046,908

Figure 14: Results of Matching and Derivation of the ATE

Plotted the distribution of outcome differences between matched pairs, revealing both that the average treatment effect is positive (23.490 million PRPM) but that it is very slight, as the mode of the distribution is around 0, which if it were the mean would indicate no causal effect.

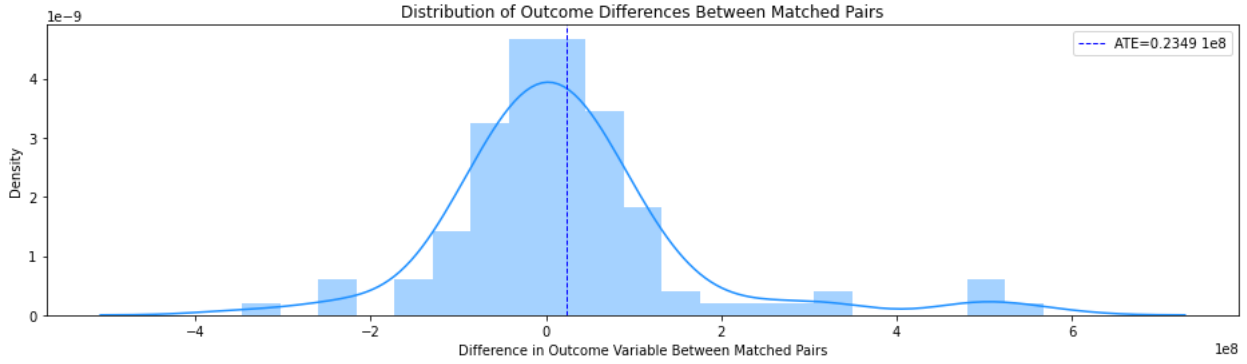


Figure 15: Distribution and Spread of Bootstrapped ATE Estimates

Plotted the distribution of the bootstrapped ATEs with mean ±1 SD shaded in, showing that negative ATE values are within one SD of the bootstrapped ATE mean. This non negligible probability of a negative ATE significantly dampens the strength of the causal effect we found.

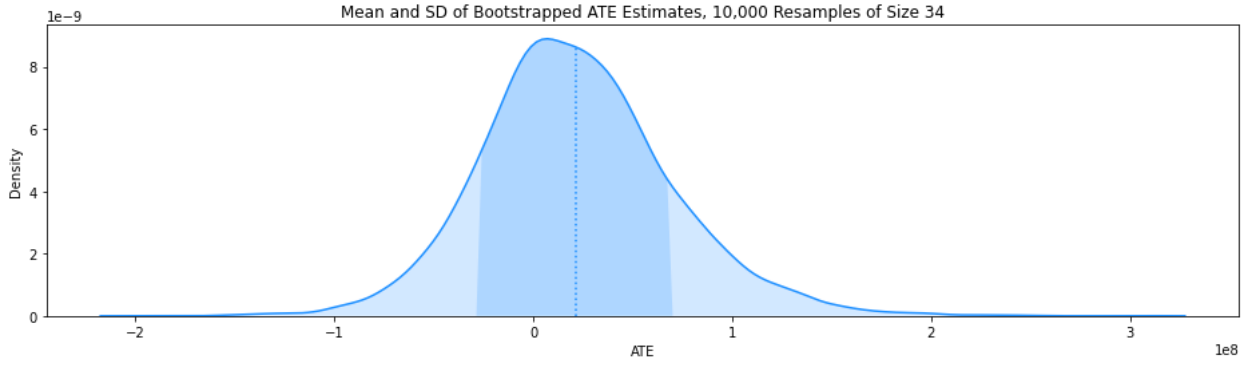
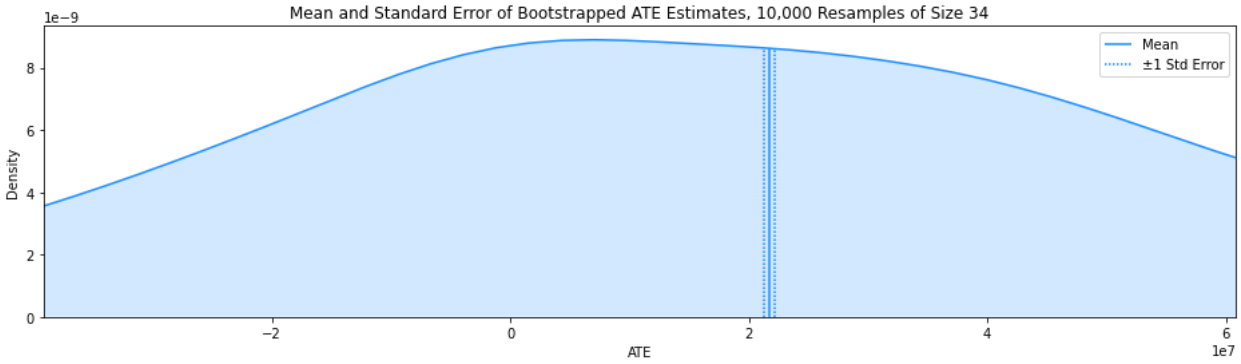


Figure 16: Standard Error of Bootstrapped ATE Estimates

Plotted the distribution of the bootstrapped ATEs with mean ±1 SE highlighted. The SE from 10,000 bootstrap ATE estimates is small, less than 500K PRPM, meaning the slight causal effect of ~23 million PRPM we found is still likely positive despite being so weak.



8 Sources

BTS Statistical Standards Manual October 2005. 2005, pp. 23-30.

Branch, Foreign Trade Data Dissemination. “Foreign Trade: Census Data.” *U.S. Trade with South and Central America*, 21 Apr. 2009, https://www.census.gov/foreign-trade/balance/c0009.html.

Manning, Willard. “Costs and Generalized Gamma Distributions .” *National Bureau of Economic Research*, 2003, https://www.nber.org/system/files/working_papers/t0293/t0293.pdf.

“U.S. Trade Representative Developing Country List .” *Home | United States Trade Representative*, http://www.ustr.gov/.

Pramuk, Jacob. “Democrats Splinter over How Best to Move Forward with Biden's Economic Agenda.” *CNBC*, CNBC, 17 Aug. 2021, https://www.cnbc.com/2021/08/17/house-democrats-disagree-over-infrastructure-bill-vote-strategy.html.

“Public Transportation Facts.” *American Public Transportation Association*, 7 July 2021, https://www.apta.com/news-publications/public-transportation-facts/.

Wilson, Martha C. “Transportation and Supply Chain Disruptions .” *Transportation Research Part E: Logistics and Transportation Review*, Pergamon, 19 Jan. 2006, https://www.sciencedirect.com/science/article/abs/pii/S1366554505000967.