NJ Transit Delay Analysis and Suggestions for Improvement

Daniel Capone, Jason Winkler, Maria Golovco, Michael Cohan, Nan Yun, Yibin Song
Drexel University M.S. Business Analytics

Abstract

Public transportation's socioeconomic importance will continually increase as we move towards a greener world. The benchmark for public transportation, especially trains, is safety, reliability, and punctuality. A consistent breach of these rigid expectations, notably tardiness, manifests itself in an enclave full of disgruntled commuters. Our hypothesis is that poor weather conditions obstruct timeliness. Therefore, we are exploring the possibility of predicting train delays, based on future weather forecasts (next 3 to 5 days), by evaluating the relationship between train delays in the New Jersey Transit Rail system and historical weather data. This is the only the beginning of train delay prediction analysis. We are fully aware that weather is not the only cause of train delays, but at this stage of our analysis we believe this topic is much more researchable than other delay factors due to the latter's data limitations. By doing this study, we hope to open the door for others to perform future train delay analysis, as well as offer direction these people and those who manage public transportation services.

Keywords*: Train Delay Analysis, Weather, Predictive Models*

Background Information

NJ Transit is the second largest statewide public transit system in the United States. During an average weekday, rail ridership is roughly 144 thousand passengers throughout the system. We will refer to the NJ Transit Rail System as NJT for the duration of the paper. NJT is comprised of 12 lines and 165 stations. All eastbound trains end at New York Penn Station or Hoboken Station. NJT operates on a 27-hour schedule and implements its weekend schedule for the weekend and major holidays. Severe Weather Level 1 or Severe Weather Level 2 schedules can be enacted for extreme weather events such as a Nor'easter. During the winter of 2018, NJT utilized its severe weather schedule on two occasions.

Problem Identification

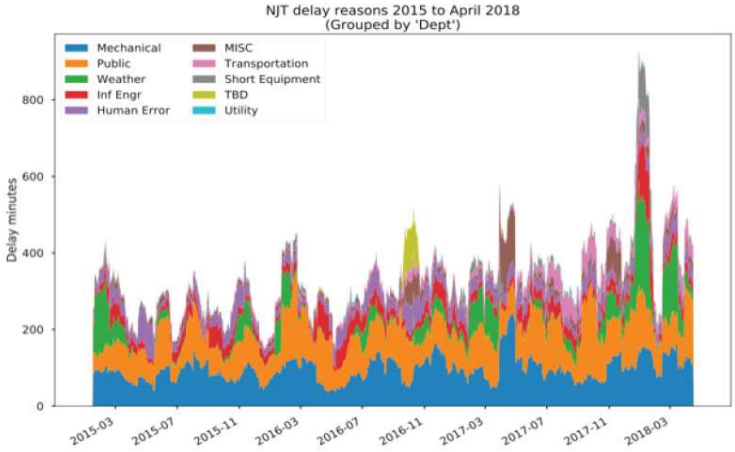
Over the past decade, NJT has developed a terrible reputation for its operating environment. Train delays are the number one issue that plagues the mass transit provider. Based on our background research, there are many reasons for NJT's current operating issues. Macroeconomic factors include:

1. Political Issues: During Governor Christie's tenure, his administration slashed NJT's operating budget, derailing its 5-year plan to upgrade its trains and infrastructure.
2. Natural Disaster: Category 3 hurricane, Superstorm Sandy, caused about $400 million worth of damage to NJT's infrastructure. NJT has not fully recovered.

Our paper aims to analyze the feasibility of delay predictions, so that better delay predictions can help NJT improve notification services. We found the following information from a consulting report (North Highland Worldwide Consulting 2018):

Due to data limitations, we chose to do our delay analysis based on the third greatest cause-weather. With that in mind, we know our predicting accuracy based purely on weather could be relatively low. However, as

we pointed out before, we are not focusing on the prediction accuracy. Instead, we show the possibility of predicting delays with ideas for future analysis when more data is available.



Data Description, Evaluation and Processing

To begin our process, we had to extract the data, then structure it so it could be merged on as many relevant attributes as we could find. We wanted to start with a large, but logical, set of attributes. Then, narrow those attributes down as our models became more accurate. Our initial dataset, from Kaggle, held a year's worth of delay information about NJT trains. The data was broken up by month. A sample of our data set shows fields such as Train ID, stations along the trip, and expected versus actual arrival times. Each row of data held information around a train and it's start/stop point on a line.

The main dataset had roughly 3 million rows. In order to do our prediction analysis, we needed to reduce the data amount then reorganize the data so each train per day only represent one row in our final dataset. Based on our experience, we selected Qliksense as the tool we would use to aggerate and reduce the data.

Before we condensed the data and merge rows, we did the following steps to add more attributes:

1. **Add in weather data:** The Qlik Data Market houses a variety of sets, from weather to population statistics. Due to time constraints, we calculated the averages between New York City, New York and Philadelphia, Pennsylvania, then joined the weather to our primary dataset by date.
2. **Create our own attributes:** From dates, we extrapolated bank holidays, seasons, and weekday. Times were used to determine flags for rush hour. We researched what ranges in certain weather attributes lead to categorizations of 'extreme' or 'mild' weather. We also calculated flagged delays as delay time greater than or equal to 6 minutes, which was based on how NJT categorizes delays and the fact that around 95% of our data came from delays between the interval of 1-8 minutes. This flag will be used later when creating dependent variables.

After adding the variables for future analysis, we condensed our data and created dependent variable in three different ways:

1. **Use Average Delay Time:** We calculated the average delay times from the delay times between each station. Trains with average delay time over 6 mins are classified as delayed and marked as 1.

2. **Use Median Delay Time:** we used the same logic as above. Instead of using the average delay times for all stations, we used the median data to represent the train after condensing. Trains with average delay time over 6 mins are classified as delay and marked as 1.
3. **Delay Percentage:** as mentioned above, we flagged each original row as delayed, or not, based on 6 mins as a threshold. For example, if one train has 13 rows of data, and more than 50% of rows (7 + rows) are marked as delay, then once condensed, this train will be classified as delayed with 1.

After all the processing, we successfully reduced the newly created dataset from over 3 million rows to 241,653 rows with 25 variables in total (including 3 dependent variables). Above is a global representation of our data which looks at delays by train while incorporating weather. The graph shows that all trains have many delays across the whole year. This in a way confirms that mechanical issues are the foremost reason for delays, but this does not eliminate the role weather may play in a delay. Additionally, by taking a closer look at delays and the specific days where delays were highest, there is some overlap between weather metrics. In the following section, we will explore the relationship between delay and weather using statistical methodologies.

Model Creation and Testing

Logistic regression is a very common method for classification. The calculation result will be the probability of classifying the target item as 1 (binomial). By selecting the appropriate threshold, a classification decision can be made. Since we are aiming to predict delays, a delay probability made the most sense. Therefore, we explored logistic regression in depth and used it as our main method. In addition, support vector machine (SVM) and KNN were also used to explore and verify the logistic regression result.

Logistic Regression

1. Variables Selection

Variables were selected with the intent to improve the accuracy of our model. We started with temporal variables, non-temporal variables related to trains operations and capacity, weather variables, and delay aggregation variables. Most of the variables selected have been kept in the model, as explained in the model building section. Chi-Square Tests were performed on marginal decreases in deviation of the model as new variables were added. Several variables, such as AvgDewPointF, were not kept in the model because of redundancy reasons (more on collinearity shortly).

Some statistically significant variables (most of the temporal variables) were removed from the model. However, our processing resources were not enough to transform the original 3 million records. This prevented us from transforming the data at a more granular level, which in turn prevented us from being able to make conclusive statements about particular variables.

On top of individual model tuning, automatic methods of variables selection, such as stepwise selection with "backward", "forward", and "both" directions, were used. Neither of the methods seem to yield dramatically different results, with a slightly lower AIC (Akaike information criterion) for "both". Therefore, AIC was the preferred selection method where possible. There will be more details to follow under model selection.

2. Class Bias and Data Sampling

A common issue in LR is unbalanced classes. Since unusual events have lower rates, the models predicted in favor of the majority class, the "zeroes". There were +/-15% "ones" and +/-85% "zeroes" in our data after aggregation. Proportions differ based on the aggregation method used. A test of proportions was performed to ensure there are statistically significant differences between each of the three proportions.

Certain variables had a strong correlation. For example, Avg Temperature and Avg Dew Point had clear multicollinearity warnings, observable in unstable coefficients and changes in signs. We performed a VIF multicollinearity test which concluded variables such as Avg Dew Point needed to be removed from the model.

3. Model building:

We used two methods to try to correct the imbalance in class.

- lowering the probability cut-off in assigning an outcome to the positive or negative class (optimal cut-off to maximize AUC was selected after each run)
- correcting the imbalance in class with "controlled" oversampling, under-sampling, both over-and-under-sampling as well as synthetic samples creation. These were all run with R package ROSE.

4. Results

The table shows the result of our logistic regression. As we can see, the accuracy is not very high, around 75% for all three different methods of creating dependent variables. However, this is expected since weather is not the main cause of delays.

	Null Model	Significant predictors	Stepwise (Both Direction)
1. Use Average Time	AIC: 156,021	Weekday1: -0.11 DailyPrecipIn: 0.21 LineShared: 0.41 NYPennStart: 0.22 Holidays: -0.15	AIC: 148,958 AUC: 65.1% Accuracy: 78.0%
2. Use Median Time	AIC: 163,647	Weekday1: -0.14 DailyPrecipIn: 0.16 LineShared: 0.5 NYPennStart: 0.29 Holidays: -0.15	AIC: 155,891 AUC: 66.1% Accuracy: 76.0%
3. Use Delay Percentage	AIC:170,078	Weekday1: -0.11 DailyPrecipIn: 0.21 LineShared: 0.41 NYPennStart: 0.22 Holidays: -0.15	AIC: 162,101 AUC: 65.5% Accuracy: 74.0%

5. Model Validation

We follow the steps below to validate the model.

- Split data in 75% vs 25% based on the Class Delay (3 splits since there were 3 methods of aggregation).
- Validate results against the test sample/unknown to the model.
- Use confusion matrix to calculate performance metrics such as Accuracy, Precision and Recall.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is another classification technique. It has received considerable attention because it is based on statistical learning theory and has shown promising empirical results in practical applications. Even though the mathematical theory behind SVM method is relatively complicated, the result is typically straight forward. A successful SVM analysis can identify an optimal hyperplane which, for a two-dimensional space, divides a plane in two parts where each class lay in either side. In addition to clear analysis results, two other reasons we wanted to give SVM a try is that SVM works well with high-dimensional data and avoids the curse of dimensionality problem.

1. Variable Selection

Based on variable exploration from logistic regression and the SVM theory, we selected more appropriate variables for this analysis. The variables for the two methods are not a one-to-one match. However, we did try to keep similar variables in both analyses, so results are comparable. Most of the variables we selected are weather related, such as average humidity, average pressure, average temperature, average wind speed,

and daily snowfall. This was done for two reasons- first, weather is our research target; second they are numerical variables, which work well for SVM.

2. SVM Analysis

1) Data processing

Since we want to use 75 percent of all data to train the SVM model and 25 percent to test the model, we spilt dataset into training and testing sets randomly. Training set ends up having 181,241 observations and testing set having 60,414 observations. Also, due to distance calculation, we standardized our data, so that variables in the nature of large scale won't dominate the distance calculation result.

2) Model Creation

Due to the large quantity of data, we were not able to complete the analysis in R. Instead, we utilized Python to complete our SVM analysis. In the first step, we imported 'sklearn' library. To reduce the running time in python, the parameter 'iteration' is assigned to 10,000. The parameter 'gamma' is auto. By applying SVC formula from 'sklearn', we trained SVM model in training set with different target values (assign 0 or 1 by average delay time, median delay time and delay percentage).

3) Results

When we use average delay time to assign dependent variable, we got 79% accuracy, 52% AUC and 467 true negatives. When we use median delay time, the results show 78% accuracy, 54% AUC and 990 true negatives. Using the last method of assigning dependent variable, we achieved 78% accuracy, 54% AUC and 1,025 true negatives. Since we are aiming at predicting delay, the highest true negative is the best, because true negative means we predict delay correctly. However, if we look at the model result from all three models, the results are relatively stable. Please see the table above for result comparison and the table below shows confusion matrix.

Use Average Time	Use Median Time	Use Delay Percentage
AUC: 52.0% Accuracy: 79.0% True Positive:50,488 True Negative:467	AUC: 54.0% Accuracy: 78.0% True Positive:49,040 True Negative:990	AUC: 54.0% Accuracy: 78.0% True Positive: 49,126 True Negative:1,025

	Use Averge Time		Use Median Time		Use Delay Percentage	
	Predicted: No Delay	Predicted: Delay	Predicted: No Delay	Predicted: Delay	Predicted: No Delay	Predicted: Delay
Actual: No Delay	50488	593	49040	1238	49126	1270
Actual: Delay	8866	467	9146	990	8993	1025

3. Model discussion

Considering time and computing power limitation, we were not able to utilize K-fold cross-validation. In K-fold cross-validation, data would be spilt into k subsamples. A single subsample would be retained as the validation data for testing the model, and the remaining k − 1 subsamples would be used as training data. The SVM model will be trained several times. The amount of time it will take is unpredictable and we decided to skip this process. This is one of the limitations we have for our analysis, which will be discussed more later.

KNN Method

After exploring logistic regression and support vector machine, we decided to try another way of classification, the K-Nearest Neighbors (KNN) methodology. KNN is a non-parametric lazy learning algorithm, because it does not make assumptions about the data distribution and it keeps all training data without doing any generalization.

1. Variable Selection

KNN classifies new cases based on a similarity measure (e.g. distance functions). In addition, we want to keep the variables for different methodologies relative similar, so that the analysis results are comparable. Therefore, we used the same variables as SVM for KNN analysis, which are shown in the chart before.

2. KNN Analysis

1) Data processing

In order to keep the consistency of our analysis, we also split data by 75/25 for KNN analysis. 75% of the data are used as training dataset, while the remaining 25% are used as testing dataset. Standardization is also applied to numerical variables.

2) Model Creation

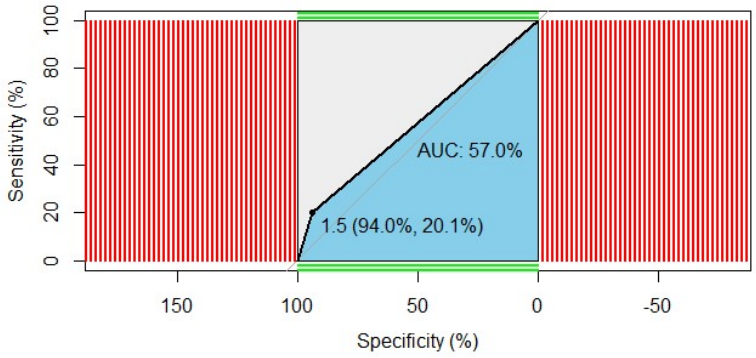
KNN analysis model is developed in R with the function "knn". We selected K=1 as a start point, then we picked K=5 as a second scenario for comparison purpose. Both K=1 and K=5 models are generated for the dependent variables we created using three different methodologies.

3) Results

The results of our KNN analysis are very similar for K=1 and K=5 scenario. Below is the chart summarizing the model results.

Use Average Time		Use Median Time		Use Delay Percentage	
K=1	***K=5***	***K=1***	***K=5***	***K=1***	***K=5***
AUC: 58.7% Accuracy: 74.9%	AUC: 58.1% Accuracy: 78.0%	AUC: 58.8% Accuracy: 73.4%	AUC: 58.5% Accuracy: 76.5%	AUC: 58.2% Accuracy: 77.7%	AUC: 57.0% Accuracy: 80.8%

As we can see all the six models show very similar result. Using Delay Percentage as dependent variable and K=5 as an example, we created the ROC chart. Please see the graph below:



4) Model Discussion

KNN analysis is very computationally expensive, because it stores all of the training data. It took less time than SVM method. For SVM, each model takes about 3 hours to produce result, however, for the six models we created using KNN method, each takes about 15 mins to generate. We also tried to test the model with cross-validation. Unfortunately, our computing power reached upper limit. We tried the function "knn.cv" in R, and the R code was running overnight without producing any result.

Model Results Comparison

After utilizing all three methods, we put all the model results in the same table to compare. Please see the table below:

	Logistic Regression	Support Vector Machine (SVM)	KNN Method (k=5)
1. Use Average Time	Accuracy: 78.0% AUC: 65.1%	Accuracy: 79.0% AUC: 52.0%	Accuracy: 78.0% AUC: 58.1%
2. Use Median Time	Accuracy: 76.0% AUC: 66.1%	Accuracy: 78.0%: AUC: 54.0%	Accuracy: 76.5% AUC: 58.5%
3. Use Delay Percentage	Accuracy: 74.0% AUC: 65.5%	Accuracy: 78.0% AUC: 54.0%	Accuracy: 80.8% AUC: 57.1%

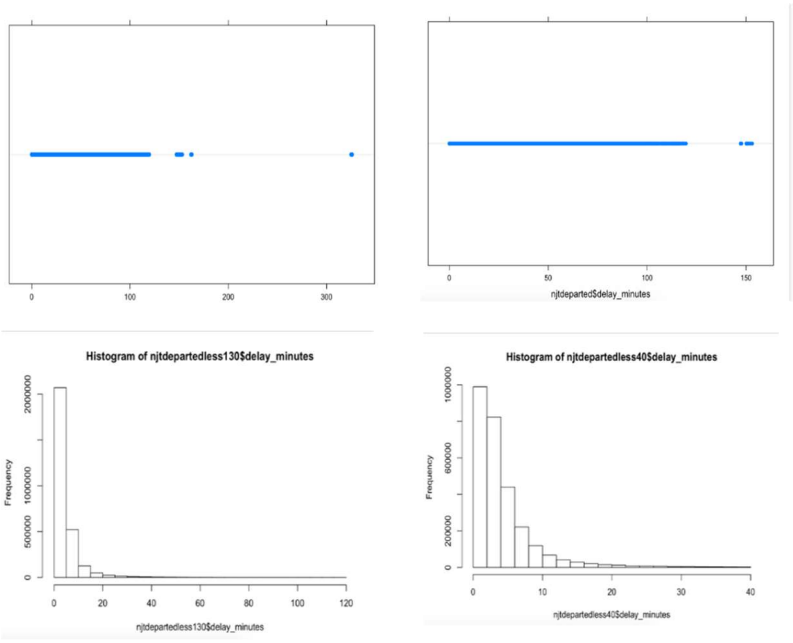
As we can see, all three methodologies produce very similar results. Accuracies are in the range of 75% to 80%, and AUC is in the range of 50% to 70%. In our situation of imbalanced data, we need to consider recall, precision and F- measure in addition to accuracy. We decided to only compare accuracy in this situation due to its functional business purpose. Based on the nature of the transit industry, customers are more sensitive to whether you can accurately predict delays. The cost of false positive and false negative is relatively low. From the table, we can tell the overall model accuracies are not very high. As we pointed out above, weather is a relatively small reason for train delays, therefore the low prediction accuracy is not a surprise to us. For the same reason, AUC is not very high. In order to increase both prediction accuracy and AUC, we will need data related to the other possible causes for a delay, such as mechanical and public factors, which are considered the top two delay reasons.

Discussion of Limitations

Given that NJT is still scratching the walls of the abyss caused by two terms of neglect and financial evisceration by the Christie administration. We think the new, public-facing and focused campaign on commuter sentiment, is a good way to improve customer satisfaction. At a 5/31/2019 Board Operations and Customer Service Committee Meeting that was open to members of public, data was used to support efforts made towards the outreach campaign (NJT 2019). Data transparency, which decreased alongside capacity, capability, and even NJT's own visibility in to how the rail system, is operating in real time (Tate 2017). In turn, our limitations for this study were plagued by data acquisition struggles.

The challenging part in working with the NJT dataset was its size given various train lines combined with several daily journeys where each journey would have variating number of sequences. In order to bring the data to a manageable size, aggregations have been made based on the 3 parameters as explained above. However, these transformations might have negatively impacted both model's precision and accuracy.
The variable distribution which was used as a basis of the Class calculation (Minutes of Delay) was strongly right skewed in addition to having several strong outliers. From the graphs below we can demonstrate how variable's Delay Minutes range is narrowed down by just applying some basic filtering.

If we were to calculate the outliers for njtdeparted$delay_minutes variable based on the classical rule of IQR-/+1.5*1st or 3rd Quartiles, we would have to consider all the delays > 11.67 minutes outliers. Given this wide range of variability in delay minutes and considering the constraints we worked under, using the NJT business rule where delays are classified as such when >6 minutes was the best choice from the available options.



Another interesting finding prior to data aggregation was that stop_sequence was a statistically significant coefficient when regressed against delay minutes, meaning the "later" stops, will have more delays systematically. We tried to account for this effect by adding a variable "Stop" counting the number of stops for each train complete journey.

Although with these limitations, we view our work as a starting point for future researchers. For meaningful advances to be made on the overall impact caused by the 14% of delays due to weather (North Highland Worldwide Consulting 2018), we recommend obtaining four key areas data, at the train-station level, that we did not incorporate: granular intraday weather, accurate rail ridership data from each station, peak/off-peak weekday figures, and the compounding effect that weather delays cause both intraday and inter-day. Longer term, we posit a much greater risk exist taking into account NJT's troubled history.

Future Implications

New Jersey is feeling the effects of climate change which may be increasing the percentage of delays contributed to weather. For instance, in 2012, New Jersey was ravaged by superstorm Sandy. As previously mentioned, the storm caused $400 millions worth of damage and incapacitated NJT. It took several days for NJT to start limited train service to and from New York Penn Station. An event like Superstorm Sandy is rare, but it did show the cracks in the system.

The state of New Jersey is experiencing a change in its weather pattern. New Jersey has experienced record temperatures since 1985, and 9 of the 10 warmest summers in the last 20 years. In addition, 2018 was the wettest year on record and was 18 inches above average – yearly precipitation records have been kept since 1895 ("How will climate change impact New Jersey", February 2019).

The variety of increased weather events will bring new challenges for NJT and its riders. Each type of weather event needs to be handled in a distinct way. For example, weather events such as a prolonged heat waves put additional strain on NJT's infrastructure. Prolonged heat waves can cause mechanical issues on individual trains and the tracks they ride on. In order to alleviate the heat-related stress, NJT can invest in expansion joints, anchors, and ties to secure the track and prevent buckling during prolong stretches of 90 degrees or above days (Resilience of NJT (NJT) Assets to Climate Impacts, June 2012). The ever-changing landscape of New Jersey's weather places a new emphasis on the need for a model to accurately predict

delays based on forecasted and real-time weather, and the compounding effect weather events have on other delay causes.

Train Delay Prediction Related

As more data becomes available, the model will expand to become more accurate about predicting delays caused by or contributed from weather related reasons. As our model grows, we would like to predict delays in the future similar to a local news broadcast of its 5-day weather forecast. In addition, we believe our model can provide real-time predictions during extreme weather events, such as flash-floods due to heavy rain or snow fall above the anticipated forecast. This type of information can help NJT better prepare company resources to minimize delays during extreme weather events.

It is important for future researchers to do a deep dive into the role weather plays in the multitude of reasons for NJT delays. Understanding the effect weather has on delays, such as delays stemming from mechanical issues, can help NJT better inform its customers of a potential delay or a change in the duration of an existing delay. The need for a better alert system for customers is essential to improve the customer experience. Currently, customers receive an alert if there is a delay longer than 15 minutes. As the model improves, NJT can develop an alert messaging system/ mobile app that would allow for consumer customization. We envision a NJT rider being able to choose to receive alerts about an individual train line, how long a delay is, or if a delay has increased or decreased in length.

Data Share and Other Business Influence

NJ Transit can partner with local or national businesses to better align delay knowledge with business strategies to capitalize on accurately knowing when NJT passengers will arrive at a train-station. The establishment of an information sharing system with local taxi companies or national ridesharing companies, like Uber or Lyft, would allow drivers to be in an ideal position to pick-up customers, especially during a weather event. In addition, NJT could share the information with restaurants surrounding a train-station, so restaurants can prepare to serve the influx of NJT customer.

Advancing the model and proper implementation of it can make NJT the gold standard in commuter travel. By utilizing our recommendations/suggestions, NJT can improve operations and dramatically improve its relationship with their customers. Additionally, the model and recommendations could be adapted to help businesses in the transportation sector, especially for companies that rely on weather forecast to make business decisions.

Acknowledgements

We appreciate the guidance from our advisor, Dr. Chuanren Liu, Assistant Professor at the Decision Sciences and MIS department of Drexel University.

References

Fitzsimmons, Emma G, and Patrick McGeehan. 2016. "Neglect Brings a Steep Decline for N.J. Transit: [Metropolitan Desk]." *New York Times*, October 14: 3.

Higgs, Larry. 2017. *NJ.com.* November 6. Accessed 06 08, 2019. https://www.nj.com/traffic/2017/11/coming_this_winter_simpler_nj_transit_storm_schedules.html.

Kiefer, Eric. 2018. *Patch.* February 21. Accessed May 21, 2018. https://patch.com/new-jersey/livingston/here-are-new-jersey-transit-s-most-least-used-train-stations.

Levin, Alan. 2012. *Bloomberg.com.* December 6. Accessed June 10, 2019. https://www.bloomberg.com/news/articles/2012-12-06/nj-transit-had-400-million-in-hurricane-sandy-damage.

NJT. 2019. "BOARD OPERATIONS AND CUSTOMER SERVICE." *njtransit.com.* May 31. Accessed June 8, 2019. https://www.njtransit.com/AdminTemp/Customer_Service_Committee_Meeting_Agenda.pdf.

—. 2019. *NJTransit.com.* Accessed June 09, 2019. https://www.njtransit.com/tm/tm_servlet.srv?hdnPageAction=CorpInfoTo.

—. 2018. *NJTransit.com.* December 05. Accessed 06 09, 2019. https://www.njtransit.com/tm/tm_servlet.srv?hdnPageAction=PressReleaseTo&PRESS_RELEASE_ID=3257.

North Highland Worldwide Consulting. 2018. *STATE OF NEW JERSEY Department of Transportation Strategic, Financial, & Operational Assessment of NJT.* Assessment, North Highland Worldwide Consulting.

plansmart nj & URS. 2011. *NJT Transit Score Guidebook .* Guidebook, Trenton, NJ: plansmart nj.

State of New Jersey Governor Phil Murphy. 2018. "State of New Jersey Governor Phil Murphy News and Events Press Releases 2018." *NJ.gov.* December 20. Accessed June 8, 2019. https://nj.gov/governor/news/news/562018/approved/20181220a.shtml.

Tate, Curtis. 2017. *NorthJersey.com.* September 26. Accessed June 8, 2018. https://www.northjersey.com/story/news/watchdog/2017/09/26/gold-standard-no-more-nj-transits-train-operations-display-board-going-dark/672811001/.