TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
2022, Vol. 2676(12) 627–640

DOI: 10.1177/03611981221096442
journals.sagepub.com/home/trr
SAGE

Modeling System-Wide Urban Rail Transit Energy Consumption: A Case Study of Boston

Zhuo Han[1], Eric Gonzales[1], Eleni Christofa[1], and Jimi Oke[1]

Abstract

Rapid transit systems are critical components of urban public transportation networks in their impact, not only on personal mobility but also on the energy and environmental costs associated with network operations. To facilitate effective planning for current and future needs, a framework is required that provides important consumption metrics and also explains the various contributors to energy consumption, along with their interactions. To address this gap, we estimated models that utilized operational and ridership data for the Massachusetts Bay Transportation Authority's rapid transit system, as well as ambient temperature, to accurately predict system-wide electricity consumption. The models were trained with data from 2019 and tested with data from 2020. The estimated multiple linear regression (MLR) and random forest (RF) models explained 93% and 95% of the variance in the data set, respectively. The MLR model provided predictions with a root mean squared error (RMSE) of 2.7 MWh and mean absolute percentage error (MAPE) of 4.68%, while the RF model resulted in an RMSE of 2.94 MWh and MAPE of 5.01%. We also investigated the impacts of COVID-19 on the transit system by exploring the effects on ridership, energy consumption, cost, and train movement metrics before and during the pandemic. We find that the models are robust and perform well, even with the significant disruptions associated with the COVID-19 pandemic.

Keywords

data and data science, big data, multimodal analysis, public transportation, line, urban

Urban rapid transit systems typically use electric trains to move heavy and light rail vehicles with high passenger capacity. Large transit systems use hundreds of gigawatt-hours of electricity each year to carry millions of people. The accurate prediction of the energy consumption of rail vehicles is thus important for planning operations, managing fleets, making energy purchasing agreements, and understanding the environmental impacts of a system. Although it is well known that train speed and acceleration, ridership, weather conditions, and several other factors can affect the energy consumption of trains, the individual contributions of each of these factors is not fully understood. Therefore, a model that accurately predicts the energy consumption of rapid transit systems would be valuable for managing transit operations so as to limit energy-related expenditures while serving passenger needs.

As a case in point, the Massachusetts Bay Transportation Authority (MBTA) operates the public transportation network that serves the Boston metropolitan area—the fourth busiest in the United States by passenger ridership (*1*). It consists of a light rail line (Green Line) and three heavy rail lines (Red, Orange, and Blue Lines) as shown on the system map Figure 1. The MBTA currently has meters at electric substations throughout the system but no direct measurements of electricity consumption by trains. The energy consumption of the rapid transit system is significant, with the MBTA spending an average of $38 million on 422 GWh of electricity annually between 2009 and 2020. This includes demand for vehicle traction power, signal systems, and station operations.

[1]Department of Civil and Environmental Engineering, University of Massachusetts Amherst, Amherst, MA

Corresponding Author:
Zhuo Han, zhuohan@umass.edu

Figure 1. Schematic map of the rapid transit network of the Massachusetts Bay Transportation Authority. The lines considered for this study are: Red, Blue, Orange (heavy rail) and Green (light rail).
Image source: https://www.mbta.com/schedules/subway.

This paper thus addresses the gap in understanding the factors that contribute to energy consumption in an urban rapid transit system and how they interact. To this end, this research investigates the system-level relationship between energy consumption and train movement, using the MBTA rail transit system as a case study, via an interpretable modeling framework. During the course of the study, the COVID-19 pandemic had a profound impact on rapid transit ridership and operations, which led to greater variation in ridership and train operations than would normally be expected. These changes prompted the added objective of evaluating the impact of operations during the pandemic on system-wide energy consumption.

The rest of the paper is organized as follows. First, we provide a review of relevant research efforts in urban rail energy modeling, highlighting key findings and current gaps. Following this, we describe the case study network

and data sources, providing an exploratory analysis of energy consumption and operations data from the MBTA along with general insights about temporal trends and the relationships between the data. In the Methods section, we describe the research framework, which includes an automated process for constructing detailed trajectories for each vehicle's movement—speed and acceleration—which are important determinants of the energy required for tractive power. We also describe the multiple linear regression (MLR) and random forest (RF) models estimated to relate explanatory variables to the system-wide energy consumption of the MBTA rapid transit system. In the subsequent Results and Discussion section, we analyzed these models to identify the factors with the strongest effect on energy consumption. The performance of the models was tested with training data from 2019. The models were then used to make predictions on observations from 2020. The performance revealed the robustness of the models, as the COVID-19 pandemic led to significant reductions in rapid transit ridership and train operations. Yet, the models were able to accurately capture these effects on energy consumption. We conclude the paper by summarizing our findings, their implications, and future research directions.

Literature Review

Urban rail transit is most promising as a sustainable mode of transportation, as it contributes the least to transportation energy demand. Historically, several efforts have aimed to accurately model the energy consumption of train vehicles and proposed various approaches for reducing consumption. Notably, Wang and Rakha (*2*) created a train energy framework, which takes instantaneous regenerative braking efficiency into account, and demonstrated that a 20% reduction in energy consumption was possible in the Chicago network. Others have developed simulation, optimization, or machine learning approaches to find energy reduction pathways (*3*–*9*). In all of these instances, energy reductions were achieved by solving for optimal vehicle trajectory patterns or schedule changes. In real-world cases, however, these solutions are not often feasible to implement.

A few other researchers have attempted to pinpoint other sources of transit system-wide energy consumption besides train operations themselves. For instance, Wang et al. (*10*) demonstrated that lighting systems in underground stations are a significant contributor to overall energy consumption. González-Gil et al. (*11*) showed that energy gains of over 25% are possible via a combination of efficient driving strategies, smart metering, and power management, along with small-scale renewable generation. Leung and Lee (*12*) estimated a model that indicated station design and weather as important factors for energy consumption in a rail transit system.

The physical characteristics of railroads and train routes also have significant effects on energy consumption. Some research efforts thus proposed models for selecting the optimal vertical alignment or transit route design to reduce system energy usage and associated costs. Kang et al. (*13*) developed an optimization model for choosing the best train route. The selected route not only maximized the net benefits but also satisfied geometric design constraints. Comparisons and analysis of different vertical alignment designs have been conducted to find the most energy-efficient solution for railroads (*14*, *15*) and highways (*16*). In this study, however, the network geometry is fixed. Thus, the model for system-wide energy consumption is estimated based on the MBTA's existing infrastructure.

These developments notwithstanding, there remains a gap in the literature on system-wide models that are not only highly interpretable and but also high-performing in their predictive capability. Such models can be extremely useful for planners in understanding how various aspects of the transit system affect energy consumption, beyond train movements, as well as in providing practical pathways to reduce energy consumption, and, consequently, costs, both of which are of great importance. The research presented in this paper addresses this gap by estimating models based on the rail network that serves the Boston area. We demonstrate that our approach is readily deployable on transit systems to yield similarly useful insights.

Data

The goal of the study is to determine the key factors affecting system-wide energy consumption via an interpretable model that can also reliably predict electricity consumption based on readily observable factors. Thus, we use data obtained for the MBTA rapid transit system in Boston, Massachusetts, and surrounding communities (Figure 1). The following subsections provide descriptions of the data on energy consumption, energy costs, ridership, train movements, and weather, all of which were used in the modeling process.

Energy Consumption

Energy consumption data from 2008 to 2020 were extracted from spreadsheets provided by MBTA. Monthly energy use from these years is shown in Figure 2 via a boxplot. This figure reveals that the greatest monthly energy consumption occurs consistently during the winter months, presumably as a result of heating needs. Peak usage is observed in January (average of

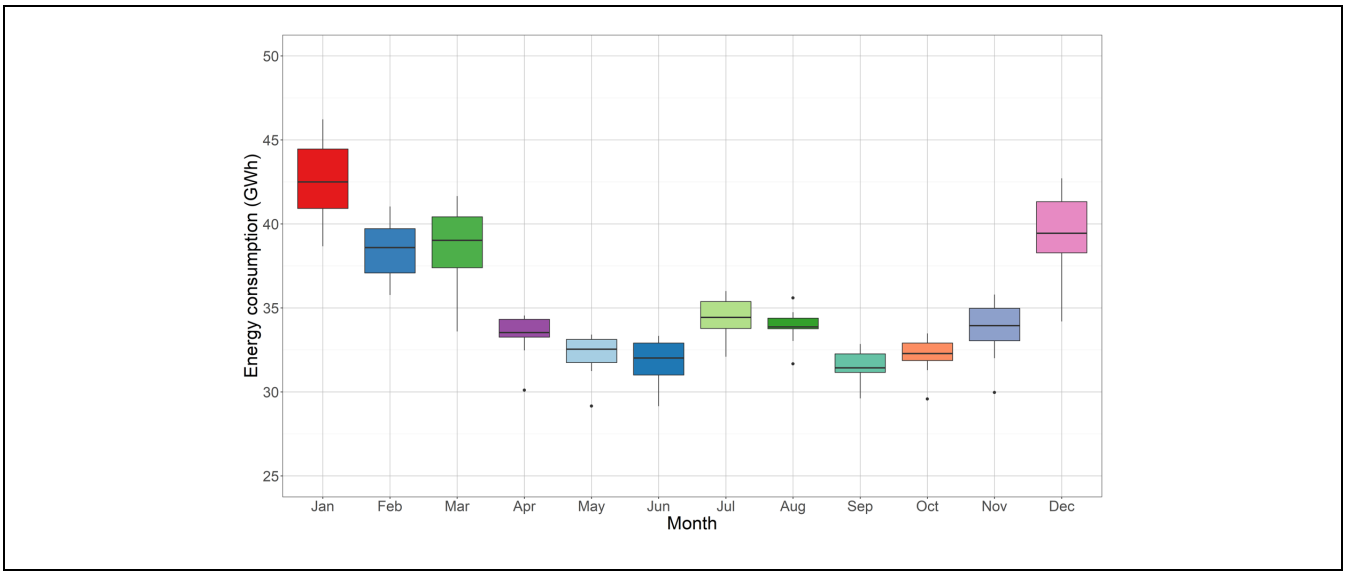

Figure 2. Boxplot of Massachusetts Bay Transportation Authority rapid transit monthly energy usage, 2008–2020.
Note: The height of each box indicates the interquartile range of observations in that month. Horizontal lines in the boxes indicate the median.

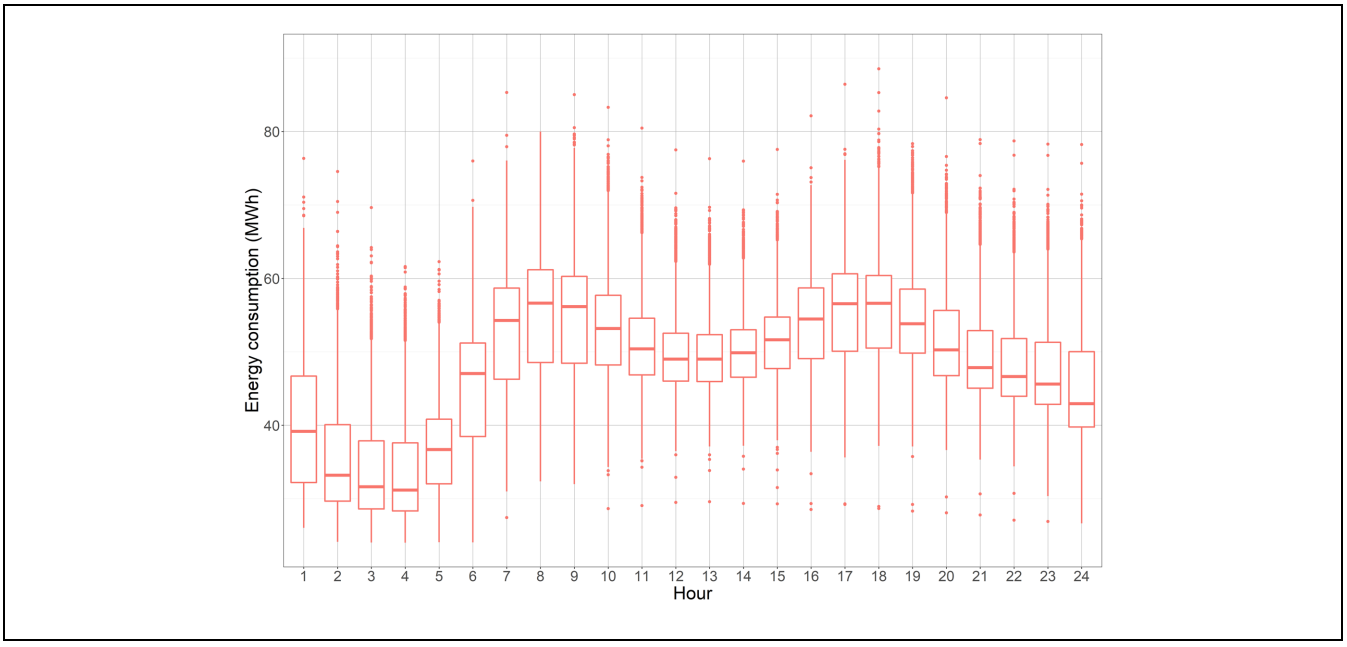

Figure 3. Boxplot of Massachusetts Bay Transportation Authority rapid transit hourly energy consumption.

43 GWh). From April through November, energy consumption is visibly lower, presumably because of reduced third-rail heating needs. A smaller peak, however, is observed in July, which is typically the hottest month of the year during which the demand for air conditioning is greatest. Thus, weather appears to be a significant explanatory variable for energy consumption.

The hourly energy distribution, also based on observations from 2008 to 2020, is shown in Figure 3. The boxplot indicates that the peak consumption period occurs from 7 to 9 a.m. and from 4 to 6 p.m., which are the hours associated with the most frequent train operations. For reference, the average hourly peak energy consumption in 2019 was 53 MWh and average overnight hourly

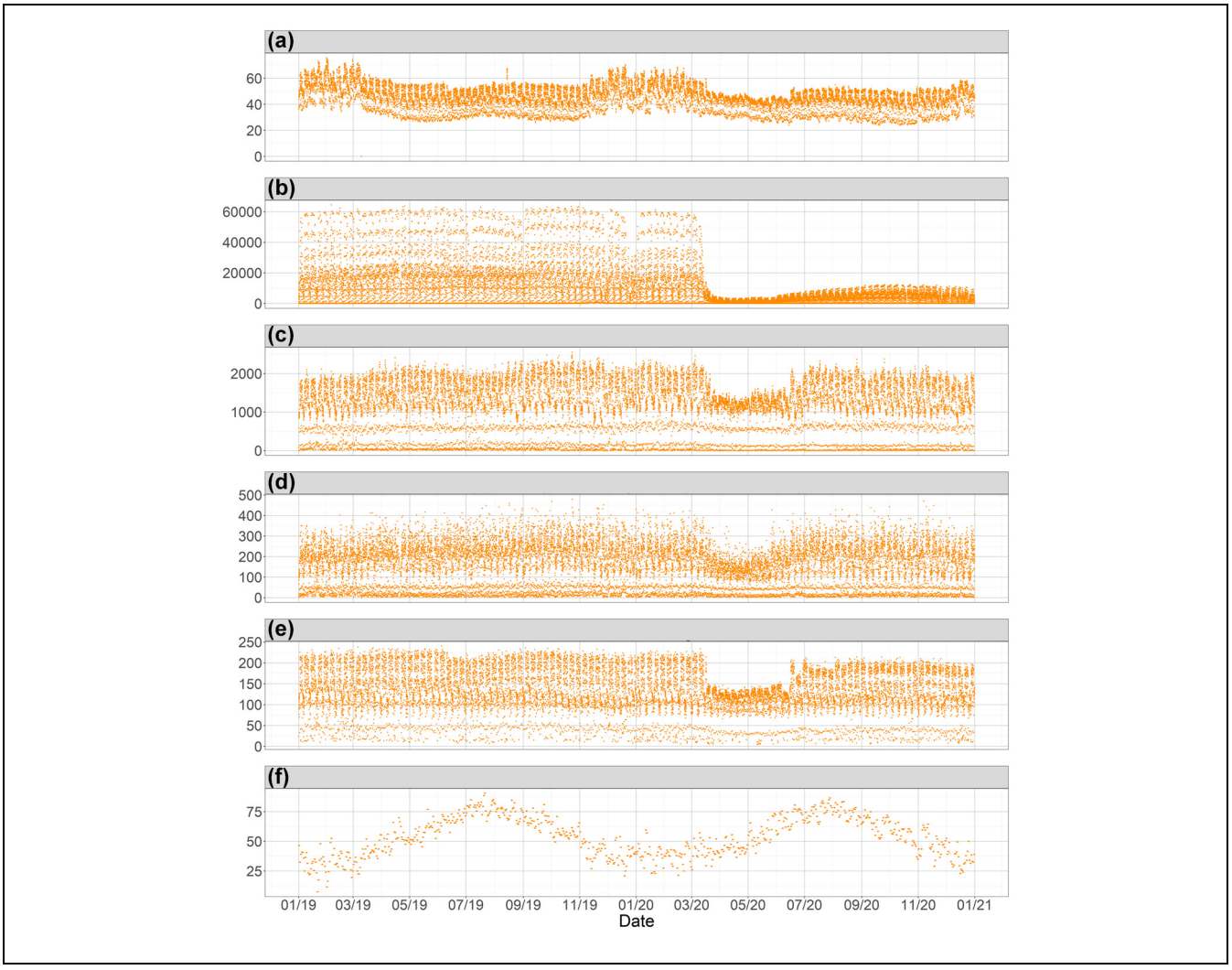

Figure 4. Time series plots of energy, ridership, vehicle-miles, vehicle-hours, and operating trains (at hour resolution) and average daily temperature, from January 2019 through December 2020 for the Massachusetts Bay Transportation Authority (MBTA) rapid transit system: (*a*) hourly energy consumption (MWh), (*b*) hourly system ridership, (*c*) hourly operating distance (vehicle-miles), (*d*) hourly operating time (vehicle-hours), (e) hourly operating trains, and (*f*) average daily temperature (°F).
Data sources: (*a*)–(e) MBTA; (*f*) National Oceanic and Atmospheric Administration (https://www.noaa.gov/weather).

energy peak was 75 MWh. The "overnight" period corresponds to the hours during which there is minimal train movement and no ridership, as the MBTA rapid transit system closes from a period after midnight until about 5 a.m. The lowest energy consumption occurs between 2 and 4 a.m., when the system is closed, yet there is a baseline average energy consumption of 34 MWh during this period. The overall average hourly energy consumption from 2008 to 2020 is 48 MWh.

The time series of hourly energy consumption for the years 2019 and 2020 is shown in Figure 4*a*. Energy consumption declined sharply in March 2020, coinciding with the beginning of the pandemic and the associated lockdown policies.

Ridership

System-wide hourly ridership data were obtained from the MBTA Research Database for 2019 and 2020. The time series is shown in Figure 4*b*. The ridership is sourced from farecard tap-in data that logs the location and time that each passenger enters the system. These data provide reliable counts of total ridership in aggregate. Since the MBTA's rapid transit vehicles are not equipped with automated passenger counters (APC) and passengers do not tap out of the system, the passenger loads on individual rapid transit lines or on specific vehicles can only be inferred through analysis of patterns of farecard use and vehicle tracking data through the MBTA's Origin-Destination-Transfer (ODX) model (*17*). During 2019,

ridership followed a typical trend with high values during weekday peaks and lower values during off-peak hours and on weekends. Beginning in March 2020 a steep decline in ridership is detected as a result of the COVID-19 pandemic and consequent lockdown policies. As of the end of 2020, ridership had only begun to increase again but remained well below pre-pandemic levels.

Train Location

Train location data were obtained from the MBTA Research Database, which includes comprehensive records of time-stamped locations of every vehicle in the system. Because of the size of the tables, the data were downloaded and processed for 2019 and 2020 only. The number of unique trains running and operating distance as well as operating time in each hour were calculated by analyzing the sequences of observations for each train ID. These measures are shown in Figure 4*c–e*. We observe that these three measurements have a similar trend from 2019 to 2020, decreasing as a result of targeted service reductions in response to COVID-19 lockdown policies starting in March 2020 and ending in July 2020.

Weather

Average daily temperatures in Boston for the years 2019 through 2020 were obtained from the National Oceanic and Atmospheric Administration. Figure 4*f* shows the time series of the temperatures from 2019 to 2020. As expected, there is a clear seasonal pattern in the data. The lowest temperatures are recorded in January (average: 34°F) and the highest in July (average: 77°F). In a given year, we see that weather impacts on energy consumption can be approximated by a quadratic effect. This is because of the increased energy needs both for cooling (as temperatures rise) and for heating (as temperatures drop).

Methods

Trajectory Computation and Energy Modeling Framework

We developed an integrated framework to (a) obtain and process data from a variety of sources, (b) compute train trajectories (distance, time, speed, and acceleration), (c) generate model input variables at the hourly level, (d) estimate planning metrics (such as energy per vehicle-mile, energy per vehicle-hour), and (e) model and predict hourly energy consumption. The framework is depicted in Figure 5.

We queried the MBTA Research Database to obtain heavy and light rail location data, along with tap-in ridership data, for a given month and year. Location data were used to compute trajectories and other train-operation variables. We also obtained daily average temperature data from the National Oceanic and Atmospheric Administration (NOAA) and spreadsheets for hourly energy consumption and monthly expenditures on electricity from MBTA. All of these variables were then integrated into a combined table and aggregated at the desired temporal level (hourly). Using this integrated table, we created a dashboard that depicts

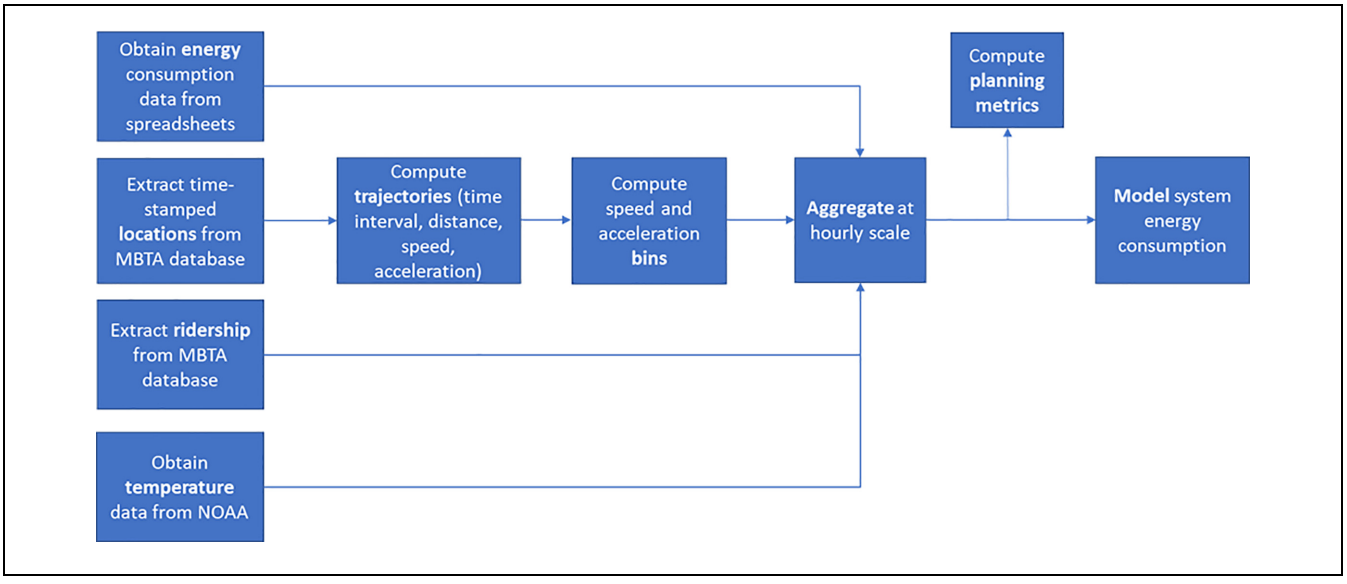

Figure 5. Flowchart of modeling framework indicating data inputs, processing, and model outputs.
Note: MBTA = Massachusetts Bay Transportation Authority; NOAA = National Oceanic and Atmospheric Administration.

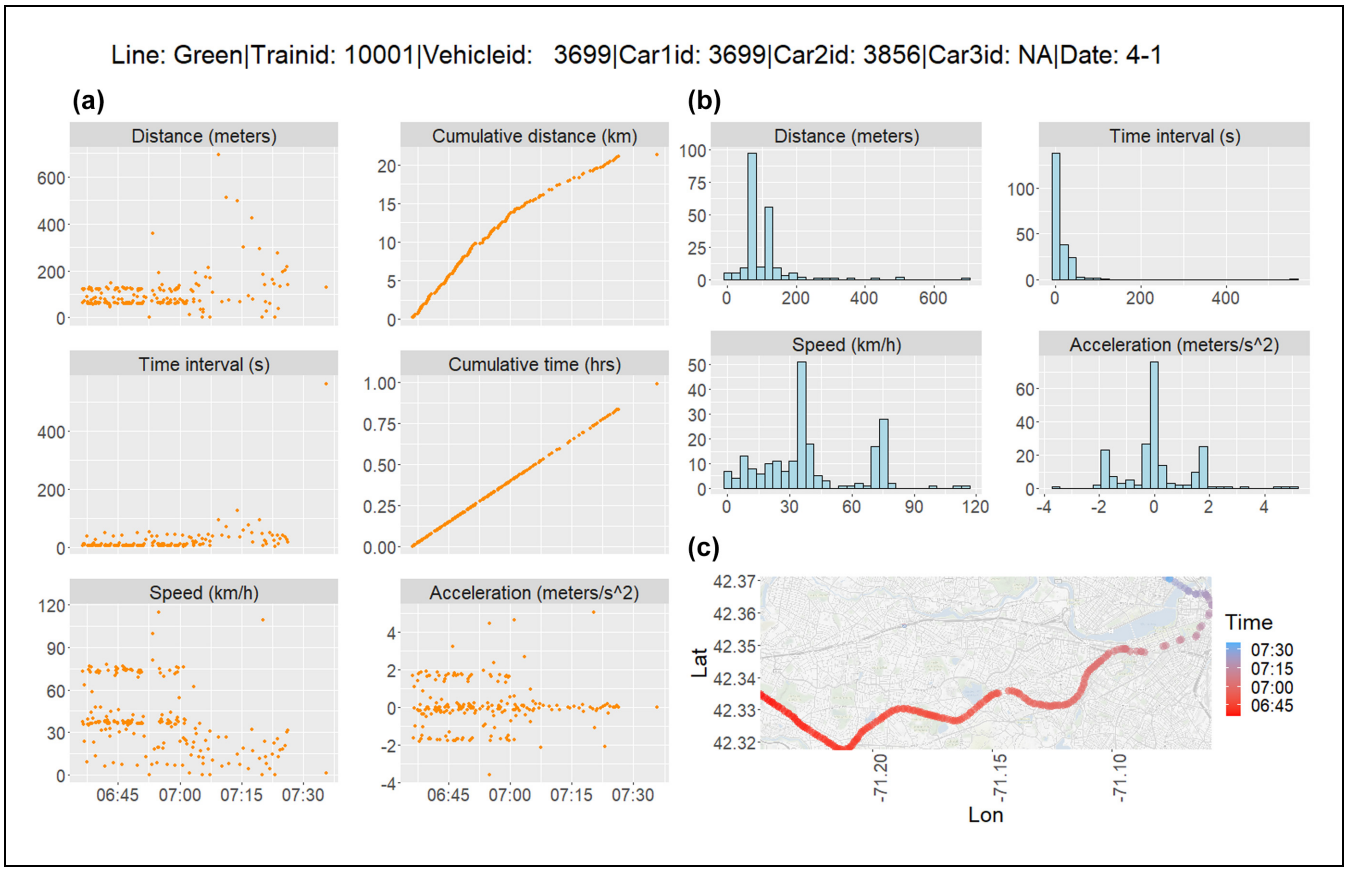

Figure 6. Trajectory dashboard of Green Line 10001 on April 1, 2019 including: (*a*) time series and (*b*) histograms of vehicle distance traveled, time interval, and calculated speed per time interval, and (*c*) map tracking the train location.

time trends and distributions of these data as well as the trajectory per train line; see Figure 6. The metrics of interest are: the distance traveled between observations, the time interval between observations, as well as speeds and accelerations inferred from the location and time data. The dashboard provides a visualizing platform for tracking train operation and helps us estimate the metrics. In addition, these integrated tables were used to compute planning metrics for energy, cost, and emissions. We then used these metrics as training and validation data for the estimated system energy model.

The distance, time interval, speed, and acceleration for each unique train in a given day were computed using location data. Together, these measures constitute the trajectory of a given train. An example of a trajectory from the Green Line is shown in Figure 6. Based on the coordinates in the raw trajectory table, the distance traveled between each two time series records and the corresponding time interval duration were computed. Using these data, the speed, acceleration, cumulative distance, and cumulative time could be calculated accordingly. All of these measurements are visualized by time series in Figure 6*a*. Figure 6*b* shows the distribution of distance, time interval, speed, and acceleration observations in the form of histograms. These distribution data are important for verifying that all of the speed and acceleration measurements are plausible for the corresponding line, otherwise errors would need to be identified and addressed. Finally, Figure 6*c* presents the sequence of location (latitude, longitude) observations for a specific train on a map. The color of each point corresponds to the time of day to indicate the direction that observed trains traveled over the course of the entire day. The trajectory in Figure 6*c* is traveling from west to east because the earlier times start with red and later times end with blue.

Speed and Acceleration Binning

The energy consumption of trains fundamentally depends not only on their mass, but also on their speeds and rates of speed change (i.e., acceleration). Given the high-level system model objective, we can capture the contribution of train movement to energy consumption by observing how much time is spent at various speeds and accelerations. This allows for fewer variables

Table 1. Speed and Acceleration Bins

Bin number	Percentile range	Speed (mph)	Acceleration (m/s^2)
1	[0, 16.7)	[0, 3.8)	[−5.0, −0.4)
2	[16.7, 33.3)	[3.8, 11.0)	[−0.4, 0)
3	[33.3, 50)	[11.0, 15.5)	[−0.1, 0)
4	[50, 66.7)	[15.5, 22.6)	[0, 0.1)
5	[66.7, 88.3)	[22.6, 30.1)	[0.1, 0.7)
6	[88.3, 100]	[30.1, 104]	[0.7, 5.0]

(depending on how many intervals are used) and, therefore, less uncertainty in the model parameters.

To facilitate this estimation, we created equal-probability (quantile) bins for speed and acceleration. We tested the efficacy of different bin numbers, with the objectives of parsimony and interpretability. Ultimately, we selected six bins for both speed and acceleration. These bins are summarized in Table 1. Each bin is initially used as an indicator variable (1 or 0) for each train trajectory observation. Each indicator variable is then multiplied by the time duration of the corresponding interval in which it is observed and summed up to represent the total duration of time in a given bin.

Based on the fundamental physics of train energy consumption (*2*), both the speed and acceleration of a vehicle are contributors to the energy demand. Thus, we further compute speed–acceleration bin-time variables using each of the bins. This results in 36 combinations of speed and acceleration bins. Each interval observation is then assigned an indicator corresponding to the matching speed–acceleration bin it represents. When the data are aggregated at the hour level, the results of the speed–acceleration interaction variables denote the total time (vehicle-hours) collectively spent at a given speed interval, acceleration interval, or joint speed–acceleration interval.

For convenience, we use the notation "S[*X*]A[*Y*]" to represent the speed–acceleration interaction time variable. For example, S1A1 represents the speed–acceleration interaction time for speed bin 1 [0, 3.8 mph) and acceleration bin 1 [−5.0, −0.4 m/s^2).

Summary of Variables

Following the data processing and speed–acceleration binning, the interval observations in the data set were aggregated at the hour level to reduce the number of observations. To determine how each of these variables interacts and affects energy consumption, we estimated two models based on the observations. The set of variables considered are summarized in Table 2.

Multiple Linear Regression

An MLR approach was used to predict hourly energy consumption as a function of the before mentioned explanatory variables. MLR has the advantage of being highly interpretable because of its simplicity, without sacrificing performance. First, we used the Lasso technique to extract the most relevant variables. The Lasso performs this based on a regularization approach that shrinks irrelevant coefficients to zero. Following this, we then estimated a few MLR models based on variable subsets obtained by removing further correlated variables. Multicollinearity was identified via correlation coefficients and variable inflation factors. The training set consisted of 80% of the hourly observations in 2019, while the remaining 20% was reserved for validation. The final MLR model was selected by considering fitness and validation statistics, namely, adjusted coefficient of variation (R^2), root mean squared error (RMSE) and mean absolute percentage error (MAPE). Observations from the year 2020 were used to test the performance of the final model, in addition to analyzing the impacts of COVID-19.

Random Forests

We also used an RF model to predict energy consumption. RF is an ensemble learning approach that estimates

Table 2. Explanatory Variables Considered for Hourly Energy Modeling

Variable (units)	Description (in given hour)
A[*Y*] (hours)	Total operating time spent in equal-probability acceleration bin *Y*
Average hourly speed (mph)	Average of operating train speeds
Monthly dummy	Month-specific indicators
Number of trains	Unique operating vehicle count
Operating distance (vehicle-miles)	Sum of distances covered by operating vehicles
Operating time (vehicle-hours)	Sum of total moving time of operating vehicles
Ridership	Tap-in ridership counts
S[*X*] (h)	Total operating time in an equal-probability speed bin *X*
S[*X*]A[*Y*] (hours)	Total operating time jointly spent in speed bin S[*X*] and acceleration bin A[*Y*]
Temperature (°F)	Average daily temperature

Table 3. Summary of Model Goodness-of-Fit Metrics with Validation Metrics for the Random Forests Model Based on the Out-of-Bag Estimates

		Validation		Test	
Model	Training	RMSE (MWh)	MAPE (%)	RMSE (MWh)	MAPE (%)
Multiple linear regression	R^2: .93	2.19	3.32	2.70	4.68
Random forests	PVE: .95	1.78	NA	2.94	5.01

Note: RMSE = root mean square error; MAPE = mean absolute percentage error; PVE = proportion of variance explained; NA = not available.

multiple regression trees based on respective bootstrap samples of the data (*18–20*). It mitigates noise and bias by using a random fixed-size subset of variables at each branching (node-splitting) step of the tree-partitioning process. The two hyperparameters the modeler must select are the number of estimators (trees) and size of the subset of random features to be considered for tree partitioning. The goodness of fit of an RF model can be determined based on error metrics computed on the out-of-bag (OOB) sample. OOB observations are those that are not present in any of the bootstrap samples. Thus, the OOB metrics serve as an estimate of the validation error of the model. Ensemble models, such as RF, are less interpretable than parametric approaches. However, in RF, variable importances can be computed from the tree-partitioning process. These importances can rank the relevance of each of the explanatory variables to the dependent variable.

In this application, the best hyperparameters for the RF model were 500 estimators (trees) and a random splitting-variable subset size of 40. The entire set of 2019 hourly observations were used for training the model (noting that about 37% of these observations are expected to be in the OOB sample). We reserved observations from the year 2020 for predictive performance testing.

Results and Discussion

Model Fitness and Performance

We gauge the fitness of the MLR model based on the adjusted coefficient of determination (R^2), which indicates the proportion of variance explained (PVE) by the model. The RF model is evaluated by a similar PVE metric. In determining the best parameterization for each model, we assessed the RMSE and the MAPE. For the MLR model, these were computed on 20% of observations in 2019 used for validation. In the RF model, validation metrics were computed from the OOB sample observations, which were not used by the model in training. Finally, we compared the performance of both models on the test set of observations from the year 2020. Based on the prediction over the year, we obtained an RMSE of 2.7 MWh and a MAPE of 4.68% from the MLR model, while the RF model resulted in an RMSE of 2.94 MWh and a MAPE of 5.01%. The model fitness and performance metrics are summarized in Table 3.

Figure 7 shows time series of observed and predicted system-wide energy consumption. The plots indicate that the prediction errors from May to December are greater than earlier in the year, indicating that there are some temporal effects the model does not fully capture. The average residual of the MLR model is −0.31 MWh, while that of the RF model is −1.28 MWh, further indicating the slightly better performance of the MLR model. Nevertheless, both models perform well in testing and are clearly robust to the COVID-19 disruptions that occurred within 2020, as predictive performance did not show any significant degradation in spite of the system changes that occurred after March 2020.

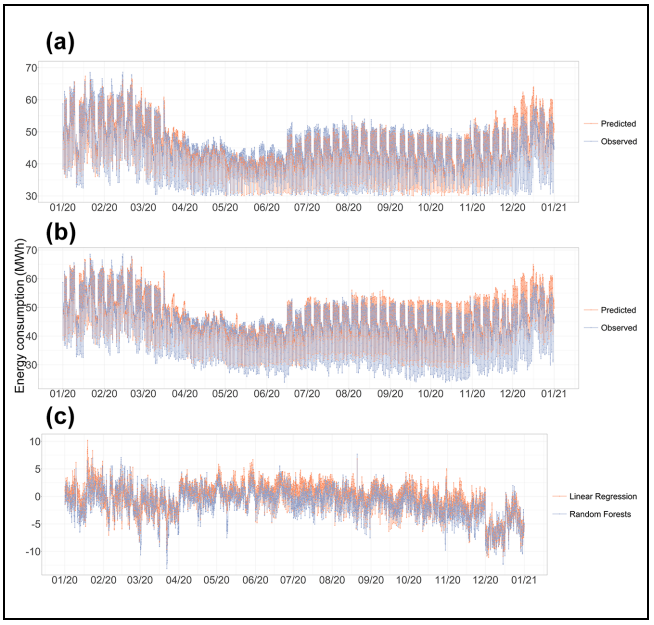

Figure 7. Predictive performance of (*a*) multiple linear regression and (*b*) random forest models on 2020 energy consumption for the MBTA. Residuals for both models are shown in (*c*).

Table 4. Variables, Estimated Coefficients, *p*-Values (Based on the Multiple Linear Regression Model), 2019 Average Values and Contribution (Product of Coefficient and Average Value) to Hourly Energy Consumption

Category	Variable	Coefficient	*p*-Value	2019 avg	Contrib (MWh)
Non-movement	Intercept	58.48	2×10^{-16}	na	58.48
	Temperature (°F)	−0.81	2×10^{-16}	53.5	−43.3
	Temperature ^{2} (°F ^{2})	0.01	2×10^{-16}	3,155.2	31.6
	Number of trains	0.06	2×10^{-16}	129.8	7.8
	Ridership	1.49×10^{-5}	3.42×10^{-5}	17,299.1	0.3
Monthly dummy	February	1.67	2×10^{-16}	na	1.67
	March	−0.52	1.18×10^{-5}	na	−0.52
	April	−3.92	2×10^{-16}	na	−3.92
	May	−5.49	2×10^{-16}	na	−5.49
	June	−4.86	2×10^{-16}	na	−4.86
	July	−4.5	2×10^{-16}	na	−4.5
	August	−3.37	2×10^{-16}	na	−3.37
	September	−4.41	2×10^{-16}	na	−4.41
	October	−5.44	2×10^{-16}	na	−5.44
	November	−3.56	2×10^{-16}	na	−3.56
	December	1.28	2.67×10^{-15}	na	1.28
Speed–acceleration interaction time (h)	S1A1	−0.38	2.33×10^{-5}	0.92	−0.35
	S2A1	−0.23	0.04	1.99	−0.46
	S3A1	0.59	3.16×10^{-7}	1.623	0.96
	S4A1	−0.19	0.13	1.617	−0.31
	S5A1	−0.81	7.23×10^{-6}	1.28	−1.04
	S1A2	0.13	6.29×10^{-8}	11.27	1.47
	S2A2	0.14	8.15×10^{-4}	9.42	1.32
	S3A2	1.23	2×10^{-16}	1.57	1.93
	S5A2	0.99	3.74×10^{-5}	0.82	0.81
	S2A3	0.03	6.14×10^{-4}	10.61	0.32
	S3A3	0.27	1.61×10^{-12}	2.007	0.54
	S4A3	0.52	3.22×10^{-8}	1.07	0.56
	S5A3	0.59	2.2×10^{-5}	1.11	0.65
	S6A3	−1.81	6.93×10^{-12}	0.59	−1.07
	S2A4	0.02	4.89×10^{-4}	12.58	0.25
	S4A4	−0.08	7.72×10^{-3}	3.49	−0.28
	S5A4	0.27	0.06	3.12	0.84
	S6A4	0.88	4.57×10^{-6}	0.951	0.84
	S2A5	0.63	6.91×10^{-5}	0.954	0.6
	S3A5	−0.74	6.23×10^{-9}	1.92	−1.42
	S4A5	−0.21	0.04	3.12	−0.66
	S6A5	1.01	1.11×10^{-9}	2.11	2.13
	S3A6	2.02	1.4×10^{-4}	0.23	0.46
	S5A6	−2.1	2×10^{-16}	1.46	−3.07
	S6A6	0.46	0.01	2.59	1.19

Note: avg = average; contrib = contribution; na = not applicable.

Key Factors Driving System Energy Consumption

The coefficients estimated from the MLR model facilitate a physical quantification of the average effect of relevant variables on energy consumption (summarized in Table 4). The table also shows the average value for each of the variables from 2019 to provide a better sense of the magnitude of each variable and its relative contribution to the hourly energy consumption. For reference, the average hourly energy consumption in 2019 was 47.3 MWh. Thus, we see that the intercept alone contributes a sizable net positive effect of 58.5 MWh. This could be a result of baseline operations, such as station lighting and climate control. Temperature is next (with the linear and quadratic portions included to capture the contributions of colder and hotter temperatures). In comparison, the net effect of ridership is quite small at 0.3 MWh. This indicates that ridership is less important in predicting energy consumption than are the train-operation- and climate-control-related energy demands on the system.

Since it was not possible to explicitly capture the effects of third-rail heating, and also because of unknown interactions with temperature, monthly dummy variables were included in the model to capture these variations.

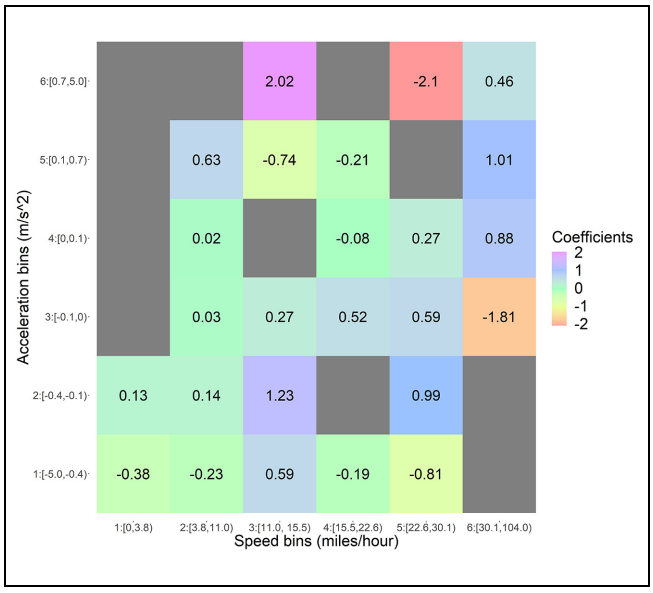

Figure 8. Hourly energy consumption coefficients of speed and acceleration bin interaction terms. The interaction signifies the amount of time spent within a given speed interval and the corresponding acceleration interval. Negative values indicate the net energy saving.

January is the baseline month. On average, there is a larger reduction in energy usage as the months proceed from March through November. Compared with the January baseline, the greatest savings are seen in May and October, when cooling and heating demands, respectively, are the lowest.

We also analyzed how the energy varies with the speed–acceleration interaction variables. The coefficients of these interaction variables are included in Table 4 and also visualized using a matrix heatmap in Figure 8. Nearly all of the interaction terms involving acceleration bin 1 (A1) have a negative coefficient except for S3A1. The negative sign indicates a net energy saving and is reflective of those trains in the system with regenerative braking capabilities (not all of them have this technology). One potential reason for the negative coefficients in positive acceleration bins (e.g., S3A5, S5A6) might be unobserved interactions that are currently not captured in the model. In contrast, a positive sign indicates a net energy consumption. We observe that the average effect of S6A5 on energy consumption is 2.13 MWh. This represents the greatest contribution to energy consumption among all positive interaction terms. These results indicate that trains spent most of their accelerating time operating within speed bin 6 (S6) and acceleration bin 5 (A5). The overall average effect of the interaction terms based on 2019 measurements is 0.94 MWh.

The RF model also ranks the variables based on their relevance to energy consumption (see Figure 9). Generally, we find that the importance rankings agree with the coefficients of the linear model. The number of operating trains was the variable with the greatest impact on energy consumption. In addition, average daily

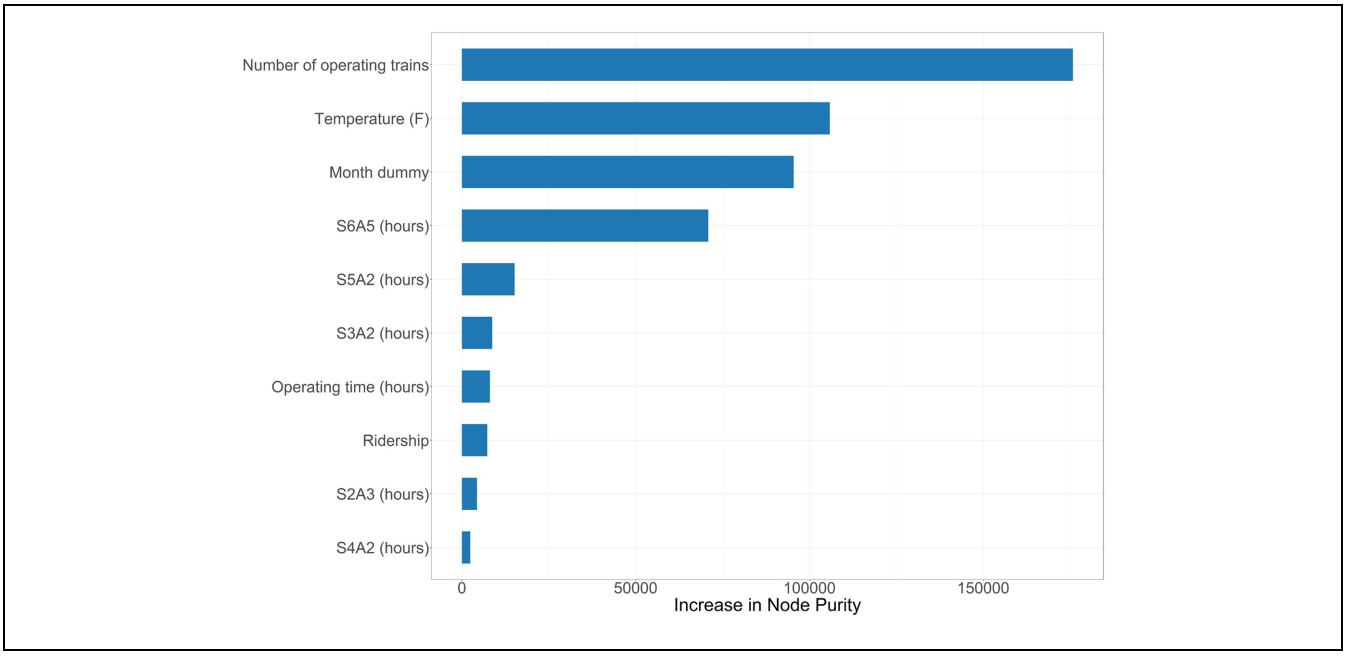

Figure 9. Top 10 important variables selected by the random forest model. The importance is indicated by the magnitude of the "increase in node purity" score.

temperature, ridership, operating time, monthly factors and some train movements (S6A5, S5A2, S3A2) are also very important for energy consumption. This model further confirmed that trains spent most of the time operating in speed bin 6 [30.1, 104 mph) and acceleration bin 5 [0.1, 0.7 m/s^2).

COVID-19 Impact Analysis

COVID-19 was first reported in December 2019 and rapidly spread all over the world, reaching pandemic proportions in January 2020 (*21*). It was declared a US national emergency in March 2020 (*22*). Subsequently, all non-essential travel was curtailed. Public places were closed and working or schooling from home was mandated in many areas. Given the infectiousness of the disease, social distancing was imposed in many locales and this significantly reduced transit usage across the US (*23*). In response to these events, and to cut costs as revenues declined, the MBTA reduced service on the Red, Orange, and Green Lines by 20%, while reducing Blue Line service by 5% beginning March 14, 2020. By July 2020, regular service was largely restored, even though ridership remained depressed throughout the rest of the year.

Key metrics for the MBTA urban transit system in 2019 and 2020 are compared in Table 5. Ridership decreased by 66% from 150.3 million tap-ins in 2019 to 51.1 million tap-ins in 2020. This ridership decline is also shown in Figure 4*b*. Compared with other metrics, ridership had the greatest change (66% decline in 2020). It did not begin to climb to prior levels until the end of 2020. The energy consumption in 2020 decreased by 7.6%, contributing to a cost decrease of 13.6%. The service reductions resulted in declines in operating distance (vehicle-miles) and operating times (vehicle-hours) of 5.7% and 13.3%, respectively. Regular MBTA rapid transit operations resumed in July 2020.

Table 5. Summary of COVID-19 Impacts on MBTA with Key Metrics Shown for the Years 2019 and 2020 and the Corresponding Percentage Change

Metric	2019	2020	% change
Cost ($\times 10^6$ \$)	16.2	14	−13.6
Energy consumption (GWh)	410.9	379.8	−7.6
Energy per vehicle-mile (kWh/mi)	39.2	38.6	−1.5
Energy per vehicle-hour (kWh/h)	270.3	281.5	+ 4.1
Ridership ($\times 10^6$)	150.3	51.1	−66.0
Vehicle-miles ($\times 10^6$)	10.5	9.9	−5.7
Vehicle-hours ($\times 10^6$)	1.5	1.3	−13.3

Conclusion

Using data from train movement and operations, ridership, and ambient temperature, we estimated models to accurately predict and explain system-wide electricity consumption in an urban rail transit network. Our case study was the MBTA network, which serves the Boston metropolitan area. Notably, we developed an integrated framework for data processing and trajectory computation. We also propose a trajectory dashboard that can visualize train trajectory variables (distance, time, speed, acceleration) and their distributions, as well as map the trajectories in real-time.

We estimated a high-performance energy consumption MLR model with an R^2 of 0.93 and RF with PVE of 0.95. On testing with 2020 data, these two models produced errors of less than 5.1%. The models provide insights into the driving factors of system-wide energy consumption, while also showing potential as a decision-support tool for future planning. These key drivers were identified as: temperature, baseline energy consumption by facility, and train operations. Ridership was found to have a very small impact on energy consumption in our framework. Importantly, our model predictions held up under COVID-19 disruptions (reduced ridership and train operations).

One limitation of the current approach is that movement variables are aggregated without accounting for rail type, that is, heavy or light rail. Given the clear differences in speeds and vehicle mass between these two types, the models could be improved by computing these variables (e.g., bin times, train numbers) as type-specific or line-specific. Ongoing research efforts include equipping individual trains with accelerometers to calibrate physical models of electric train energy consumption using their high-resolution data. These models can then be upscaled for better system-wide energy exploration, and potentially shed more light on the contribution of ridership and other operational factors.

In addition, we plan to further develop the trajectory analysis tool into an online trajectory-energy dashboard. The dashboard can be used for real-time monitoring to pinpoint areas or vehicles in network with significant changes in energy use patterns. While the estimated models have shed more light on the relevant variables for energy consumption, they would ultimately be most impactful as high-level decision-support tools that can guide planning efforts in light of future budgetary constraints or in response to disruptive events. By implementing learning procedures to map these low-level movement and operations variables to high-level planning metrics, generative processes can be estimated to produce valid synthetic data in response to any proposed policy response. Thus, future strategies can be readily assessed for energy and cost impacts, yielding sustainable decision pathways.

Acknowledgments

We would like to thank Sean Donaghy of the Massachusetts Bay Transportation Authority (MBTA) for his support and collaboration. We also thank Dr Lily (Hongyan) Oliver and Michael Flanary, both of the Massachusetts Department of Transportation, for their inputs and support. Finally, we appreciate the administrative support of University of Massachusetts Transportation Center staff Kimberley Foster, Matt Mann, and Michelle Augustinowicz, during the course of the research project.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: EC, EG, JO; data collection: JO, ZH; analysis and interpretation of results: EC, EG, JO, ZH; draft manuscript preparation: EC, EG, JO, ZH. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was undertaken as part of the Massachusetts Department of Transportation Research Program, funded by the Federal Highway Administration (FHWA) State Planning and Research (SPR) funds via the Massachusetts Department of Transportation.

Data Accessibility Statement

All the code used in generating the results and figures in this paper is publicly available at https://github.com/narslab/mbta-rail-system-energy. A sample data set is also included for reproducing the methods.

ORCID iDs

Zhuo Han https://orcid.org/0000-0003-2787-9281
Eric Gonzales https://orcid.org/0000-0002-4389-242X
Eleni Christofa https://orcid.org/0000-0002-8740-5558
Jimi Oke https://orcid.org/0000-0001-6610-445X

References

1. APTA. *Ridership Report*. Technical Report. American Public Transportation Association, Washington, DC, 2021.
2. Wang, J., and H. A. Rakha. Electric Train Energy Consumption Modeling. *Applied Energy*, Vol. 193, 2017, pp. 346–355. https://doi.org/10.1016/j.apenergy.2017.02.058.
3. Higuera, A., A. Claudio, L. Vela, L. Hernandez, and J. Valdez. Energy Performance Analysis in an Electrical Subway Traction System. *IEEE Latin America Transactions*, Vol. 14, No. 2, 2016, pp. 729–736. https://doi.org/10.1109/TLA.2016.7437216.
4. Mao, F., Z. Mao, and K. Yu. The Modeling and Simulation of DC Traction Power Supply Network for Urban Rail Transit Based on Simulink. *Journal of Physics: Conference Series*, Vol. 1087, 2018, p. 042058. https://doi.org/10.1088/1742-6596/1087/4/042058.
5. Ruigang, S., Y. Tianchen, Y. Jian, and H. Hao. Simulation of Braking Energy Recovery for the Metro Vehicles Based on the Traction Experiment System. *Simulation*, Vol. 93, No. 12, 2017, pp. 1099-1112. https://doi.org/10.1177/0037549717726146.
6. Modi, S., J. Bhattacharya, and P. Basak. Estimation of Energy Consumption of Electric Vehicles Using Deep Convolutional Neural Network to Reduce Driver's Range Anxiety. *ISA Transactions*, Vol. 98, 2020, pp. 454–470. https://doi.org/10.1016/j.isatra.2019.08.055.
7. Malikopoulos, A., and J. P. Aguilar. *Optimization of Driving Styles for Fuel Economy Improvement*. Technical Report. National Transportation Research Center, Oak Ridge National Lab (ORNL), TN, January, 2012.
8. Oettich, S., T. Albrecht, and S. Scholz. Improvements of Energy Efficiency of Urban Rapid Rail Systems. *Proc., Urban Transport X*, Dresden, Germany, 2004, p. 10.
9. Li, Y., Q. He, X. Luo, Y. Zhang, and L. Dong. Calculation of Life-Cycle Greenhouse Gas Emissions of Urban Rail Transit Systems: A Case Study of Shanghai Metro. *Resources, Conservation and Recycling*, Vol. 128, 2018, pp. 451–457. https://doi.org/10.1016/j.resconrec.2016.03.007.
10. Wang, J., A. Ghanem, H. Rakha, and J. Du. A Rail Transit Simulation System for Multi-Modal Energy-Efficient Routing Applications. *International Journal of Sustainable Transportation*, Vol. 15, No. 3, 2020, pp. 1–16. https://doi.org/10.1080/15568318.2020.1718809.
11. González-Gil, A., R. Palacin, P. Batty, and J. P. Powell. A Systems Approach to Reduce Urban Rail Energy Consumption. *Energy Conversion and Management*, Vol. 80, 2014, pp. 509–524. https://doi.org/10.1016/j.enconman.2014.01.060.
12. Leung, P. C. M., and E. W. M. Lee. Estimation of Electrical Power Consumption in Subway Station Design by Intelligent Approach. *Applied Energy*, Vol. 101, 2013, pp. 634–643. https://doi.org/10.1016/j.apenergy.2012.07.017.
13. Kang, M.-W., M. K. Jha, and R. Buddharaju. Rail Transit Route Optimization Model for Rail Infrastructure Planning and Design: Case Study of Saint Andrews, Scotland. *Journal of Transportation Engineering*, Vol. 140, No. 1, 2014, pp. 1–11. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000445.
14. Kim, M. E., N. Marković, and E. Kim. A Vertical Railroad Alignment Design with Construction and Operating Costs. *Journal of Transportation Engineering, Part A: Systems*, Vol. 145, No. 10, 2019, p. 04019043. https://doi.org/10.1061/JTEPBS.0000269.

15. Kim, M. E., P. Schonfeld, and E. Kim. Comparison of Vertical Alignments for Rail Transit. *Journal of Transportation Engineering*, Vol. 139, No. 2, 2013, pp. 230–238. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000476.
16. Kang, M.-W., S. Shariat, and M. K. Jha. New Highway Geometric Design Methods for Minimizing Vehicular Fuel Consumption and Improving Safety. *Transportation Research Part C: Emerging Technologies*, Vol. 31, 2013, pp. 99–111. https://doi.org/10.1016/j.trc.2013.03.002.
17. Sánchez-Martínez, G. E. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data: Dynamic Programming Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2652: 1–7.
18. Ho, T. K. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, 1998, pp. 832–844. https://doi.org/10.1109/34.709601.
19. Arikan, Y., and E. Çam. Assessment of Factors of Energy Consumption in Railways with the AHP Method. *The Online Journal of Science and Technology*, Vol. 8, No. 1, 2018, pp. 21–28.
20. Breiman, L. Random Forests. *Machine Learning*, Vol. 45, No. 1, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324.
21. Cheng, Z. J., and J. Shan. 2019 Novel Coronavirus: Where We Are and What We Know. *Infection*, Vol. 48, No. 2, 2020, pp. 1–9. https://doi.org/10.1007/s15010-020-01401-y.
22. Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak. March, 2020. https://www.federalregister.gov/documents/2020/03/18/2020-05794/declaring-a-national-emergency-concerning-the-novel-coronavirus-disease-covid-19-outbreak.
23. Liu, L., H. J. Miller, and J. Scheff. The Impacts of COVID-19 Pandemic on Public Transit Demand in the United States. *Plos One*, Vol. 15, No. 11, 2020, p. e0242476. https://doi.org/10.1371/journal.pone.0242476.

The authors are solely responsible for the facts, the accuracy of the data and analysis, and the views presented here.