High-Dimensional Data-Driven Energy Optimization for Multi-Modal Transit Agencies

Philip Pugliese, Principal Investigator
Chattanooga Area Regional Transportation Authority
1617 Wilcox Blvd
Chattanooga, TN 37406
E-mail: philippugliese@gocarta.org

Abhishek Dubey, Principal Investigator
Vanderbilt University
1025 16th Ave., S., Suite 402
Nashville, TN 37235
E-mail: abhishek.dubey@Vanderbilt.Edu

Aron Laszka, Principal Investigator
University of Houston
3551 Cullen Blvd., Room 501
Houston, TX 77204
E-mail: alaszka@uh.edu

Yuche Chen, Principal Investigator
University of South Carolina 300 Main St., Room C214 Columbia, SC 29209
E-mail: chenyuc@cec.sc.edu

Michael Wilbur, Graduate Student
ScopeLab, Institute for Software Integrated Systems,
Vanderbilt University
Nashville TN

Afiya Ayman, Graduate Student
University of Houston
Houston TX

Amutheezan Sivagnanam, Graduate Student
University of Houston
Houston TX

Ruxiao Sun, Graduate Student
University of South Carolina
Columbia, SC

Fred Eisele, System Architect
Institute for Software Integrated Systems,
Vanderbilt University
Nashville TN

Project Report November 2020

Overview

Transportation accounts for 28% of the total energy use in the United States[1] and as such, it is responsible for immense environmental impact, including urban air pollution and greenhouse gas emissions, and may pose a severe threat to energy security. As we encourage mode shift from personal vehicles to public transit, it is important to consider that public transit systems still require substantial amounts of energy; for example, public bus transit services in the U.S. are responsible for at least 19.7 million metric tons of CO2 emission annually[2]. As such it is absolutely crucial that we study the bottlenecks to energy efficiency in public transit and develop new algorithms that can help the public transit agencies, especially those that are still operating mixed fleets, which may consist of Electric vehicles (EVs), hybrids (HEVs), and internal combustion engine vehicles (ICEVs), optimize the operations by deciding which vehicles are assigned to serving which transit trips.

Since the advantage of EVs over ICEVs varies depending on the route and time of day (e.g., the advantage of EVs is higher in slower traffic with frequent stops, and lower on highways), the assignment can have a significant effect on energy use and, hence, environmental impact. The Chattanooga Area Regional Transportation Authority (CARTA), in collaboration with academic partners at Vanderbilt University, the University of Houston, the University of South Carolina in addition to the Chattanooga Department of Transportation and the East Tennessee Clean Fuels Coalition, is developing mechanisms to precisely solve this problem. The key aspect of the project is the development of accurate energy consumption predictors developed using high resolution telemetry gathered from the fleet and use these models within a real-time operation and network guidance system.

Our approach is to use continuous monitoring sensors on the complete mix of CARTA transit buses and to develop predictors and optimization mechanisms using the data. These required specific activities detailed further later in the article (a) Acquire high-resolution (updated every minute) spatio-temporal telemetry data from CARTA vehicles and exogenous data sources, such as traffic and weather; (b) Develop an efficient framework to store and process the operational data and external data, including street and elevation maps; (c) Create multi-scale energy predictors using the real-world data; (d) develop guidance and network optimization algorithms that use the energy predictors and real-world data to optimize operations; and (e) codify the results in to visualization dashboards and simulators that can be used by other agencies for knowledge transferences.

The key contributions of our project are as follows: (a) we developed and demonstrated an efficient big data infrastructure of managing real-time telemetry data from transit vehicles at a resolution of 1 Hz and merging it in real-time with other transit related data including occupancy

[1] EIA, "U.S. Energy Information Administration: Use of energy explained – energy use for transportation (2018)," https://www.eia.gov/energyexplained/use-of-energy/transportation.php, Accessed: May 31st, 2020, 2018.
[2] Office of Transportation and Air Quality, "Fast facts: U.S. transportation sector greenhouse gas emissions 1990–2017," Tech. Rep. EPA-420-F-19-047, June 2019. [Online]. Available: https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100WUHR.pdf

statistics and trip level statistics encoded in real-time General Transit Feed Specifications (GTFS) (b) we have developed large datasets that can be shared with the community to highlight the key features and covariates that effect transit energy performance (c) We have developed machine learning models to be able to predict energy prediction of a future trip assignment depending upon the weather and expected traffic congestion at both macro and micro resolutions. Macro resolution focuses on trip level statistics and micro resolution focuses on vehicular dynamics. Lastly, (d) we have developed optimization algorithms that can use our data infrastructure and the machine learning models to help the transit agencies decide on an energy optimal trip assignment roster. Most of these works have appeared in peer reviewed articles. A list is available at the website of our project[3].

Note that our project improves the state of art because as described by our recent paper[4], most attention in the literature is focused on passenger cars. In contrast, our project is focused on energy estimation of mixed transit vehicles that includes diesel, electric and hybrid buses. Methodologies of existing models can be roughly classified into rule-based and data-driven. Rule-based models adopt a "white-box" approach that follows some fundamental physics laws and mimic the dynamics and interactions of various vehicle/powertrain components to estimate energy consumption. Data-driven models draw on a "black-box" approach so that users do not need to understand the physical process of electricity generation and consumption, or even the principles governing vehicle dynamics and powertrain operation but rely on the exploration of statistical relationship between inputs and energy outputs with certain assumptions or statistical techniques. Among the data-driven approaches, regular linear or multiple linear regression models are the most common approaches in electric passenger car energy prediction models. Limited studies have adopted machine-learning based methods, e.g., ANN, etc. (As indicated by our results, we have found that neural networks are better suited for learning these models).

We now discuss the key highlights and results from our work.

Acquiring High Resolution Data Telemetry from Transit Vehicles

CARTA operates a mix of vehicle types, including gasoline powered vans, diesel and diesel-hybrid buses, battery-electric shuttles, and battery-electric buses, with production dates ranging from 1998 to 2016. CARTA provides service with 17 fixed routes, 3 neighborhood demand-response routes, 2 downtown circulator routes, and a complementary ADA paratransit service. The team configured operating data associated with vehicle routes, passenger counts, bus operators, and baseline performance for analysis. CARTA selected and installed a telematics kit produced by ViriCiti LLC on each CARTA fleet vehicle, to provide a real-time data stream at a minimum 1 Hz resolution of all available vehicle operating parameters, as well as GPS positioning

[3] https://smarttransit.ai/publications/

[4] Chen, Y., Wu, G., Sun, R., Dubey, A., Laszka, A., and Pugliese, P., "A Review and Outlook of Energy Consumption Estimation Models for Electric Vehicles". https://arxiv.org/abs/2003.12873

for each record. In total, we have already obtained around 32.3 million data points for electric buses and 29.8 million data points for diesel buses.

In addition, we collect static GIS elevation data from the Tennessee Geographic Information Council[5]. From this source, we download high-resolution digital elevation models (DEMs), derived from LIDAR elevation imaging, with a vertical accuracy of approximately 10 cm. We join the DEMs for Chattanooga into a single DEM file, which we then use to determine the elevation of any location within the geographical region of our project. We also collect weather data from multiple weather stations in Chattanooga at 5-minute intervals using the DarkSky API. This data includes real-time temperature, humidity, air pressure, wind speed, wind direction, and precipitation. In addition, we collect traffic data at 1-minute intervals using the HERE API, which provides speed recordings for segments of major roads, which provides data in the form of timestamped speed recordings from selected roads. Every road segment is identified by a unique Traffic Message Channel identifier (TMC ID). Each TMC ID is also associated with a list of latitude and longitude coordinates, which describe the geometry of the road segment.

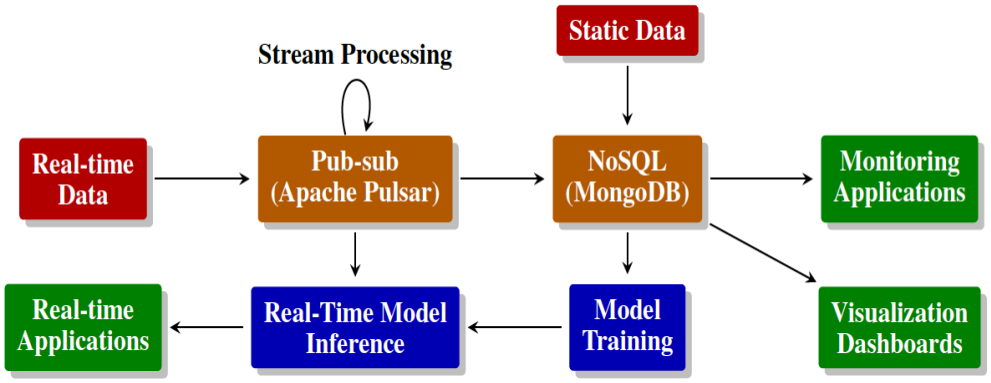
The table showing the complete datasets available to us is shown below.

Data	Source	Frequency	Scope	Features	Schema/Format
Diesel vehicles	ViriCiti and Clever Devices	1 Hz	50 vehicles	GPS, fuel-level, fuel rate, odometer, trip ID, driver ID	ViriCiti SDK and Clever API
Electric vehicles	ViriCiti and Clever Devices	1 Hz	3 vehicles	GPS, charging status, battery current, voltage, state of charge, odometer	ViriCiti SDK and Clever API
Hybrid vehicles	ViriCiti and Clever Devices	1 Hz	7 vehicles	GPS, fuel-level, fuel rate, odometer, trip ID, driver ID	ViriCiti SDK and Clever API
Traffic	HERE and INRIX	1 Hz	Chattanooga region	TMC ID, free-flow speed, current speed, jam factor, confidence	Traffic Message Channel (TMC)
Road network	OpenStreetMap	Static	Chattanooga region	Road network map, network graph	OpenStreetMap (OSM)
Weather	DarkSky	0.1 Hz	Chattanooga region	Temperature, wind speed, precipitation, humidity, visibility	DarkSky API

[5] Tennessee Department of Finance and Administration. (2019) Elevation data. [Online]. Available: https://www.tn.gov/finance/sts-gis/gis/data.html

Elevation	Tennessee GIC	Static	Chattanooga region	Location, elevation	GIS - Digital Elevation Models
Fixed-line transit schedules	CARTA	Static	Chattanooga region	Scheduled trips and trip times, routes, stops	General Transit Feed Specification (GTFS)
Video Feeds	CARTA	30 Frames/Second	All fixed line vehicles	Video frames	Image
APC Ridership	CARTA	1 Hz	All fixed line vehicles	Passenger boarding count per stop	Transit authority specific

An Efficient Framework to Store and Process Operational Data



Data architecture overview - real time data is streamed to an Apache Pulsar cluster consisting of 5 broker/bookie nodes and 5 zookeeper nodes running on-site in VMWare. A MongoDB cluster running in Google Cloud reads from the Pulsar cluster, continuously updating its data view and adding spatial indexing for monitoring and dashboard applications.

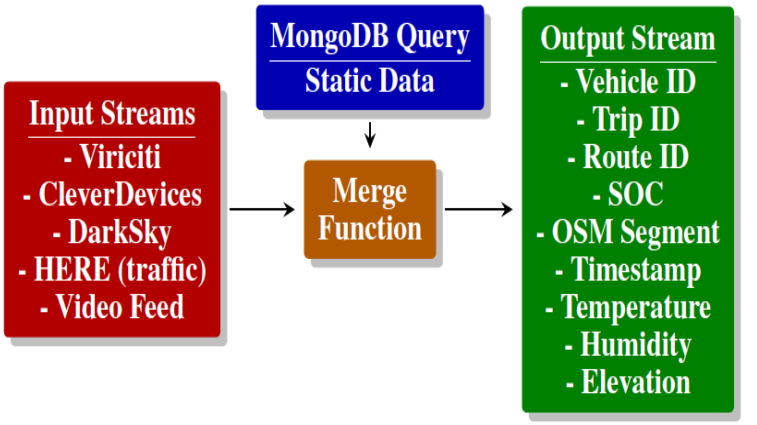
Given the volume and the rate of the data being collected, we had to design a new data architecture for the project. The purpose of this architecture is to store the data streams in a way that provides easy access for offline model training and updates as well as real-time access for system monitoring prediction. This architecture consists of a publish-subscribe cluster implemented with Apache Pulsar, which stores topic-labeled sensor streams, and a MongoDB database backend. An overview of the data architecture is provided in the figure above.

This architecture solves two challenges. The first challenge is the persistent storage of the high-velocity, high volume data streams. The second challenge is that the data is highly unstructured and irregular and different data streams have to be synchronized and joined efficiently. With this architecture, we stream each data source to a topic-based publish-subscribe (pub-sub) layer that persistently stores each data stream as a separate topic. Further, we used a three-tiered naming

convention for topic labeling. The first tier represents the name of the data tenant and all authentication and access are managed at this level. The second tier is the data category, i.e., vehicle telemetry, traffic, weather, etc. The third tier is the topic name, which represents the data source or provider, such as ViriCiti, HERE, or DarkSky. For ViriCiti, the fleet name is appended to the topic name to separate electric, diesel, and hybrid vehicles. The tenant, category, and topic names together form a topic, which downstream applications can use to access the data streams. We persistently store all messages on each topic in an append only ledger. Therefore, the topic can be used to read data in near real-time or to playback previous data streams to synchronize new downstream applications. All replication is handled at the ledger level, which allows downstream storage and applications to adapt and expand without concern for data resiliency. For this system we used Apache Pulsar[6] due to its native support for authentication and access at the tenant level and high throughput. We run Pulsar on-site on a VMWare cluster.

We include two methods for long term, structured access to the data streams. First, Pulsar includes support for Presto SQL which is a distributed SQL query engine for big data. Presto SQL integrates with the Pulsar data stores to provide an SQL interface on top of the Pulsar topics. This is useful for analytics teams comfortable with SQL, however as it is designed for large scale batch queries and does not support geospatial indexing it is not optimal for user-centric applications such as visualization dashboards. Therefore, we implemented a downstream MongoDB cluster running in Google Cloud. MongoDB was chosen for its native support of geospatial, r-tree indexing which optimizes our system for aggregate geospatial queries for monitoring and visualization applications discussed later in the report.

As our framework has expanded, we are running numerous streaming join functions within Pulsar. An example is provided in figure 2, which outputs a data stream that is used for our energy prediction models and energy dashboard. The input is the telemetry data from ViriCiti, route, trip and driver data from Clever Devices, weather from DarkSky, traffic from HERE and the video feeds. Additionally, our predictive models rely on road level information from OSM. As this data is static the latest OSM network is stored in a MongoDB collection which the function queries each evening to keep up to date. These data sources are merged at 1 second time windows, which is the resolution required by the predictive models.

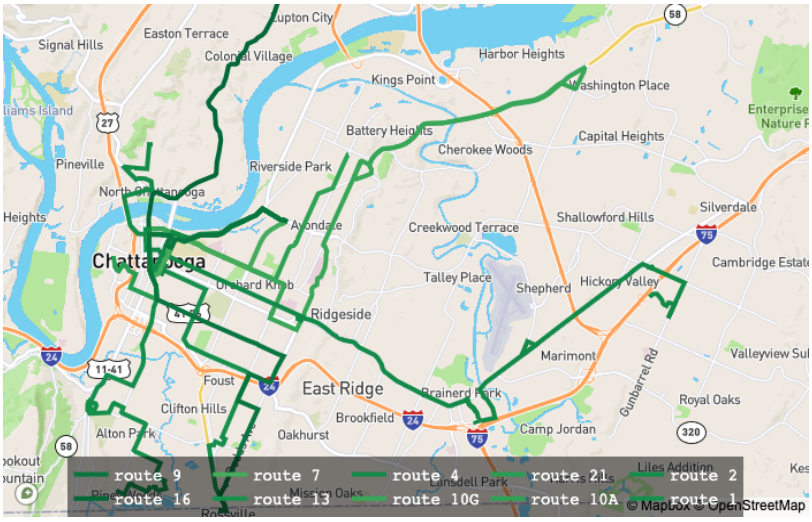


An example stream data join. Real-time telemetry and routing data from Clever Devices and ViriCiti are combined with weather from DarkSky, traffic from HERE and the video feed. The output stream includes all fields from these sources, as well as static data from OSM, GTFS and elevation. The output stream is updated periodically in real-time

[6] https://pulsar.apache.org/

Data Analysis Dashboards

To help in analysis of the big data collected and being collected as part of the project we have developed data dashboards through which the users can query based on time, fleet and route. The data is presented to the user over the map of Chattanooga as shown in the figure on the right and as a series of statistical visualizations, one of which is energy per fleet as shown in the figure as well. This dashboard is used by the data management team and CARTA to monitor the performance of the CARTA fleets over time and is available to the public[7]. Additionally, we developed a ridership dashboard to visualize occupancy of the vehicles throughout the bus transit network. The presentation of the occupancy dashboard is similar to the energy dashboard and is available to the public as well.[8]



The average energy statistics across route are shown over the map. The figure above shows the average statistics for 2020 across the region for electric fleet.

The dashboard also allows the visualization of daily averages for the whole fleet.

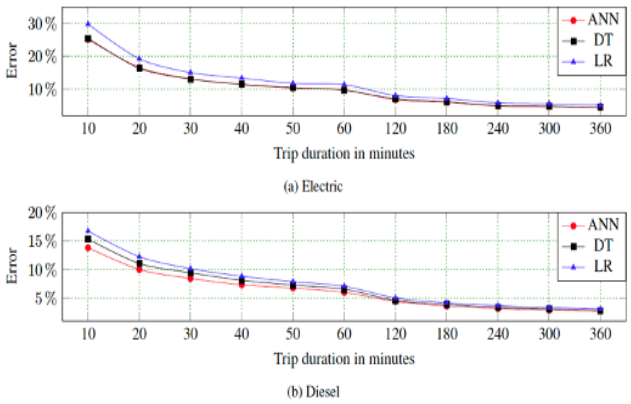
Machine Learning Models for Energy Usage Estimation

One of the first machine learning models we built was a macro-scale energy predictor[9] that can provide planning foresight by estimating energy consumption at the level of route segments. To develop this model, we used the following features for EVs: timestamp, GPS-based position (latitude and longitude), battery current (A), battery voltage (V), battery state of charge (%), and

[7] https://smarttransit.ai/energydashboard/
[8] https://smarttransit.ai/cartadashboard/
[9] A. Ayman, et.al., Data-Driven Prediction and Optimization of Energy Use for Transit Fleets of Electric and ICE Vehicles, ACM Transactions of Internet Technology, 2020.

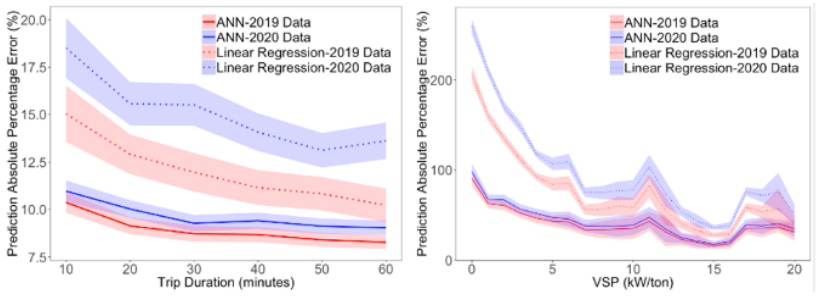
charging cable status (0 or 1). For diesel and hybrid vehicles, instead of battery data, the team collected fuel level (%) and the fuel used, in gallons. We had to remove all data points that were recorded when the vehicle was in the garage or was charging (for EVs). Next, the team calculated energy consumption by integrating the product of the measured current and voltage values and verified that these consumption values coincided with changes in state of charge. For diesel and hybrid vehicles, the team performed similar steps with fuel used.

We discovered that different neural network structures work best for electric and diesel vehicles. This is perhaps due to the way different features affect the powertrain of the vehicles. For electric vehicles, the best model has one input, two hidden, and one output layer. The input layer has one neuron for each predictor variable. The two hidden layers have 100 neurons and 80 neurons, respectively. For diesel, the best model required five hidden layers compared to the electrical vehicle model. The five hidden layers have 400, 200, 100, 50, and 25 neurons, respectively. We use sigmoid activation in all hidden layers and linear activation in the output layer. With these models we saw that the relative prediction error is generally lower for longer trips; this is expected as the individual errors of large numbers of samples cancel each other out with an unbiased prediction model. In addition, we compared the neural network models with other machine learning models including decision trees and linear regression and found that the mean error was least with the neural network models.



Energy prediction error plotted against the trip duration with neural networks (ANN), decision trees (DT) and linear regression (LR)

In addition to the macro energy models, which are applicable for route specific analysis, we have also developed micro models[10] that have been finely tuned for individual vehicles. These



(Left) Electric Vehicles: Mean and 95% confidence interval of absolute percentage errors of microscopic energy prediction for electric vehicles with trip duration for different models. (Right) Mean and 95% confidence interval of absolute percentage errors of microscopic energy prediction at with respect to the vehicle specific power.

[10] Ruxiao et.al Hybrid electric buses fuel consumption prediction based on real-world driving data. Accepted for Publication in Transportation Research Part D. Available at https://smarttransit.ai/files/microprediction2020.pdf

models are essential for estimating energy consumption under various traffic control and operational strategies. Thus, they are widely used by researchers and transportation practitioners in evaluating benefits and comparing traffic control and operational strategies.

Energy Optimal Trip to Vehicle Assignment

Based on the energy prediction models, the team set up an optimization problem that focuses on minimizing fuel and electricity use by assigning vehicles to transit trips and scheduling them for charging, while serving the existing fixed-route transit schedule in Chattanooga. The problem formulation is general and applies to any transit agency that has to provide fixed-route transit service using a mixed fleet. To solve the problem, the team introduced an integer program, a greedy algorithm, and a simulated annealing algorithm.

The team evaluated these algorithms on CARTA's transit routes using the macro-level energy predictors to evaluate the objective of total energy costs of the operations during the day. The results showed that the proposed algorithms are scalable and can reduce energy usage and, hence, environmental and operational costs. For CARTA, the proposed algorithms could save $48,910 in energy costs and 175 metric tons of CO2 emission annually.

Conclusion

The team has completed the tasks associated with this phase of the project, with the development of the vehicle telemetry system, data store and analysis framework, and initial testing of macro and micro level prediction models. The effectiveness of the developed energy usage estimation models and the developed optimization algorithms can be judged from the point that they can save $48,910 in energy costs and 175 metric tons of CO2 emission annual operations. We are actively working on improving the optimization algorithms and developing comprehensive graph neural networks to improve the prediction results. Note that the deployment of a vehicle telemetry system across legacy transit vehicles of various ages and powertrain systems was more challenging than anticipated. Additional development was necessary to convert older SAE J1708 communications to SAE J1939 format represented on newer vehicles and obtaining unique and proprietary data protocols from vehicle manufacturers often proved difficult.

Acknowledgement

This material is based upon work supported by the Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), under Award Number DE-EE0008467. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Specifically, neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe

privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. We also acknowledge the support provided through the cloud research credits by Google for this research.