Bus Fuel Consumption Problem: An in-depth Analysis and Prediction

Thanh-Nam Doan, Le Tuan Phan, Mina Sartipi
Center for Urban Informatics and Progress
University of Tennessee at Chattanooga
Chattanooga, TN
thanh-nam-doan@utc.edu,bbz181@mocs.utc.edu,mina-sartipi@utc.edu

ABSTRACT

The difficulty of fuel prediction for bus transit systems is an important research problem for both academia and industry, and an accurate predictive model has several applications e.g. urban planning, emission reduction, anomaly detection, smart city development, etc. However, this problem is complicated because of the relationships between multiple driving factors such as weather, spatial feature, traffic, etc. Predicting energy use of specific buses needs to depend on their sequential consumption nature and external driving factors. In this paper, we propose a fuel consumption prediction model based on recurrent neural network called **FuelPred**. Our model easily captures the sequential nature of fuel consumption due to the effectiveness of a recurrent neural network. FuelPred integrates various features of modeling prediction. Our extensive experiment shows that (i) FuelPred outperforms several baselines; and (ii) these features are necessary to well predict the fuel consumption.

CCS CONCEPTS

• **Networks**; • **Mathematics of computing** → **Probabilistic representations**;

KEYWORDS

recurrent neural networks, neural networks, fuel prediction, data mining

ACM Reference Format:

Thanh-Nam Doan, Le Tuan Phan, Mina Sartipi. 2020. Bus Fuel Consumption Problem: An in-depth Analysis and Prediction. In *UrbComp '20: The 9th SIGKDD International Workshop on Urban Computing, August 24th, 2020, Virtual Conference.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Vehicle emission has been a problem since World War II and has been researched for 70 years by transportation experts. Despite progress, emissions still cause a significant portion of air pollution.The U.S Environmental Protection Agency (EPA) estimates that the largest contributor of U.S. greenhouse gas emission comes from transportation, which is approximately 29% of total U.S. greenhouse gas emissions. Reducing vehicle emissions is a crucial task toward a greener earth. Emission reduction is not only valuable for environment, it is also economically beneficial. The EPA reported that every $1 spent on emission reduction saves $9 from healthcare, productivity, consumer saving, etc. For these reasons, public transit and, specifically, buses, [24] are an essential are of focus to solve the problem. We can reduce up to 61.3 million tones of CO2 emission when the ridership increases 25% in 2050 [9].

Due to public buses' crucial role in emission reduction, there is a need to predict the fuel consumption of buses in real time. Solving this problem can bring many benefits to many people; especially public transit agencies. With an accurate prediction model, public transit agencies can have a precise plan to replace buses that contribute to pollution with environmentally friendly buses. Such detailed plan can minimize the budget and maximize the efficiency to create greener environment. Second, bus transit systems usually cover large part of its city, so buses can be used as dynamic sensors to detect any anomalous activities. For example, we can use bus data to identify traffic jams and road damages.

While there are many research works on the energy use of buses, it remains to be a challenging problem. Public transit has been evolving to multimodel system. For example, the city of Chattanooga, TN offers a service named Car-a-van to serve the disabled in addition to existing public transit. Secondly, the energy use of public transit and buses is very complicated. Public transit is only more energy efficient than individual vehicles in highly-populated urban areas such as New York City, San Francisco–Oakland, Portland, Honolulu while in sparser areas, the opposite is true [26]. Thirdly, classical research works use data from surveys or simulations. This type of data is useful for doing research on a macro-level, but the lack of granular data leads to limiting the findings further. Fortunately, with the rise of technology, we now are able to collect fuel consumption data in a nearly real-time manner. Modern techniques such as deep learning and neural networks help us to oversee complicated issues such as the energy use of buses. There are not many works that take advantage of the new wave of data and techniques to study the fuel consumption problem.

Recently, deep learning has become a mainstream technique for researchers because of its effectiveness in unsupervised and supervised tasks. Natural language processing and speech recognition are well-known examples. Despite their high computational cost, these techniques yield high accuracy in prediction and are more capable of handling the non-linear nature of sequential data. Given fuel consumption prediction challenges, it is a great chance to investigate a deep learning or recurrent neural network method

UrbComp '20, August 24th, 2020, Virtual Conference

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

to generate better prediction results and incorporate more features relevant to fuel consumption behaviors.

Research Objectives. In this paper, we devote our efforts and resources to study the problem of fuel consumption of buses especially, ones using diesel. We also propose a neural network model to predict the energy use of buses. The main finding and contributions of this work can be summarized as follows:

- We will study the high resolution data of an entire city i.e. Chattanooga. Several features such as weather, temporal or spatial properties have been studied to illustrate their impact on fuel consumption of buses. With our analysis, we can understand the importance of these features as they relate to fuel consumption. To the best of our knowledge, our work is the first providing an extensive empirical analysis of these features to the energy use of buses.
- Utilizing Long-Short Term Memory (LSTM), we will integrate these above features into the fuel prediction model. The result is our proposed model named FuelPred. Our solution minimizes the number of assumptions and leverages the representation power of neural networks [17] to make its prediction.
- Through extensive experiments, we will show the importance of the above features to predict the fuel consumption of buses. The experiments are conducted upon the entire bus transportation system of Chattanooga. The results clearly illustrate the outperformance of our proposed method over several baselines. Moreover, many aspects of our model are also studies in our experiments.

Paper Outline. The rest of our paper is organized by the following structure. Section 2 provides a literature survey of some previous works that are related to our research while Section 3 shows the data collection and analysis processes. The two Sections 4 and 5 present the prediction model and experiment result, respectively. Finally, Section 6 concludes our findings and also discusses some future directions.

2 RELATED WORKS

The previous related works can be divided into two categories: machine learning methods and its applications to solve problems in public transportation.

2.1 Machine learning methods

Machine learning is a branch of artificial intelligence that offers a system able to find statistical patterns from data and improve its performance without being explicitly defined. These methods have been successfully applied to solve many problems such as computer vision and natural language processing [6]. Recently, neural network, with its ability to approximate any non-linear function, [17] has become a mainstream technique for machine learning research. To model sequential data, Long-Short term memory (LSTM) [12] models prove it is superior to other classical methods in various applications. The method is armed with other state-of-the-art techniques such as attention mechanism [3] or convolutional neural network [27] to improve its performances.

2.2 Machine Learning for public transportation research

Due to the recent breakthrough of machine learning and neural network in various fields, the research community uses them as new techniques to analyze public transportation. Moreover, with the rise of internet of thing devices, various types of data from transportation can be easily crawled so it accelerates the adoption of machine learning in the transportation research community. The community has a wide range of research topics in this field.

2.2.1 Traffic Analysis. Chiang *et. al.* [7] used the information of public transportation to predict congestion cascades in urban areas. The authors [16] extended the work further and deployed it into web dashboard for monitoring traffic congestion in Singapore. Instead of using a Bayesian model like Chiang *et. al.*, Basak *et. al.* [4] considered the sequential nature of public transportation data by using LSTM [12] to predict the traffic congestion. Hoang *et. al.* [11] explored social network information to understand events related to public transportation such as waiting and missing buses. Specifically, their work employed natural language processing techniques to analyze public tweets of Twitter users in Singapore. Now, their model helps them to study the current status of Singapore public transit. Arabghalizi and Labrinidis [1] analyzed the public transit in city of Pittsburgh in Pennsylvania to build a framework to predict how full the bus is.

2.2.2 Speed Optimization. Stovall *et. al.* [22] proposed a framework for processing urban informatics. In their system, they applied computer vision technique such as YOLO [21] to track and count vehicles in urban environment. This method can be extended to predict the speed of vehicle. Sun *et. al.* [23] built an attention neural network [3] upon an LSTM model to predict the speed of buses in Singapore. In their model, they considered external features such as weather, temporal information to increase the prediction performance. Wei *et. al.* [25] proposed deep reinforcement learning-based method *IntelliLight* to control traffic lights in an urban area. Through their extensive experiments using real data, their method can increase the speed of public transit.

2.2.3 Energy Consumption. This topic has strong relation with emission study which has been received a lot of attention from transportation researchers. Vincent and Jerram [24] studied several scenarios in a medium-sized US city and found that using bus rapid transit was the preferred choice since it offered lowest CO2 emission per passenger. The work used some assumptions to estimate the CO2 emission because of technology limitations. Without granular measurements, we cannot provide accurate solutions. Ercan *et. al.* [9] used a simulation to find that if public transportation ridership increases by 9%, CO2 emissions can be reduced up to 766,000 tonnes annually by 2050. Ayman *et. al.* [2] is one of the most recent work that uses machine learning to predict the fuel consumption of a bus fleet. They employed the power of linear regression, decision tree, and multi-layer perceptron with various features for their prediction. However, the drawback of this work is that they did not incorporate the sequential nature of bus movement into their prediction model. This limits their accuracy in predicting fuel consumption. The main difference between this work and ours is that the authors focused on predicting energy usage in route-level, while we target the individual vehicle which is more granular and harder to handle.

3 DATASET & EMPIRICAL ANALYSIS

In this section, we present our data crawling process as well as provide the overview of our dataset. Later, we show the analysis of multiple features affecting fuel consumption of a public transportation fleet.

3.1 Dataset

We collect data from sensor devices deployed on all buses in Chattanooga, Tennessee. The API is provided by the bus management agency of the city i.e. Chattanooga Area Regional Transportation Authority (CARTA). For nearly 50 years, CARTA's mission is to provide a reliable multimodal transit system for people of the City of Chattanooga and its surrounding areas. Each year, CARTA serves more than 3.1 million trips.

The range of collected data is 35 days from Feb 6th to March 11th 2020. We do not collect more recent data because COVID-19 changes the behaviors of bus riders. The entire bus fleet contains 28 vehicles and data points are sampled every 10 seconds. The longest sequence contains 14140 data points. Figure 1 provides an example of a particular bus trajectory and its fuel consumption in Chattanooga for one day. Each data point is represented as a tuple of timestamp, location, fuel consumption. We also apply some heuristic methods to filter noise in the dataset. For example, sequences shorter than three are classified as noise and we filter them out from our dataset. The final total number of sequences is 786.

3.2 Empirical Analysis

From the collected data, we conduct several empirical studies to analyze the fuel consumption of bus transportation under different features: (i) weather, (ii) temporal information, (iii) spatial feature and (iv) individual bus characteristic.

Weather feature: The effect of weather on energy consumption has been studied in Ayman *et. al.* [2]. However, in their work, the empirical analysis of weather information does not provide the reasons of using this data to study energy consumption. Moreover, we argue that data such as temperature and wind speed are difficult to utilize since their degree of change is hard to interpret.

Our original dataset does not contain weather information. Therefore, we gathered this information using the Python API from DarkSky which compiles data from multiple different stations. From DarkSky, Chattanooga is divided into a grid and the weather information of each grid is reported. For our analysis, each weather data record is a tuple of timestamp, grid center location and weather condition e.g. snow, rain, foggy; we ignore other information such as temperature, humidity, wind speed.

To find weather conditions for a particular fuel data point, we follow the steps below:

- Firstly, we calculate the distance between the fuel data point to every grid center location. We select the corresponding grid for that data point as the closest grid.
- Secondly, we find the weather record belonging to the corresponding grid with the closest timestamp the fuel data point. The weather condition of area is assigned to the fuel data point. In this step, we assume the weather condition is unique for all points in each grid.
- Finally, to ensure the correctness of our process, we randomly sample a subset of data points and manually examine the weather information.

There are six weather conditions reported by DarkSky: rain, cloudy, foggy, snow, clear and normal. For each weather condition, we calculate the fuel consumption of the buses and use box plots to display the result in Figure 2. From the figure, we observe that under different weather conditions, the fuel consumption of bus transportation is affected. For example, under snow conditions, the fuel consumption is lower than other conditions. We argue that the reasons are (i) snow is severe weather in southern states like Tennessee so people avoid driving and going outside, it leads to less traffic then buses can keep stable speed during operation; (ii) due to the skewness of our dataset, we observe fewer data points of snow condition than others so it can favor the fuel consumption in this condition.

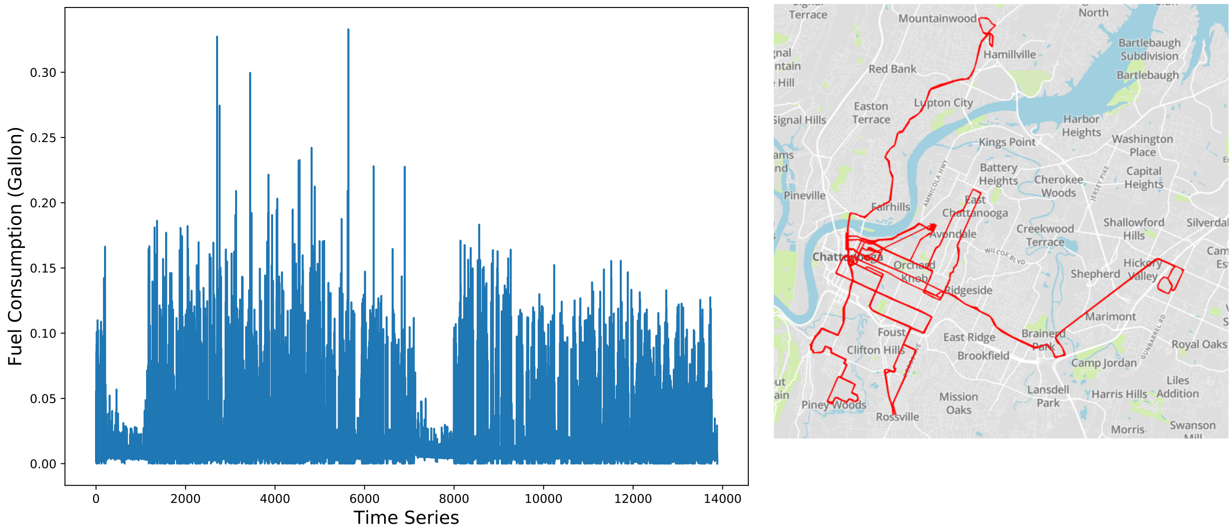
Since fuel consumption is different in various weather conditions, we need to use weather as a feature for our prediction.

Temporal feature: The temporal feature plays a crucial role in urban activities [28]. It usually takes more energy to travel in rush hours than normal hours. For this reason, we use real data to further explore its effect on the problem of energy consumption of public transit.

For each day of week, we calculate the average fuel consumption of all buses for each hour. Figure 3 shows the fuel usage of bus transportation according to the day of week. From the figure, we first observe that different days have different patterns. For example, weekends (i.e. Sunday and Saturday) are more stable than other weekdays. Secondly, at the start of day, fuel consumption is significantly larger. We argue that buses need to consume more energy to warm up their engines at the beginning of the day. Additionally, the demand for public transit in the morning is higher than other parts of the day.

Spatial feature: The effect of spatial information has significant impact on human activities [8, 14] but its importance for energy consumption has limited studies in previous works. Intuitively, some locations make vehicles consume more than others. There are multiple reason behind this phenomenon. For example, vehicles digest more in urban areas than rural areas, or going uphill requires more energy than usual. Due to this complex scenario, we simplify our analysis to a two dimensional map and display the average fuel consumption across the map of Chattanooga. First, we divide the city by using the grid of k by k. In this analysis, we use $k = 100$. Second, we calculate the average fuel usage in each cell of the grid. Finally, we normalize the average fuel usage into the range (0, 1) by softmax function [10]. The normalization helps to increase the contrast of our display.

Figure 4 shows the normalized average fuel consumption in Chattanooga. The darker the color, the more fuel the bus fleet needs to consume in this grid cell. From the figure, we observe that the fuel consumption of buses is not evenly distributed. In some places such as downtown (middle of the figure) and shopping malls (east of the city), buses consume more fuel than other places. The reason is that the ridership demand in these areas is higher



(a) Fuel Consumption

(b) Trajectory

Figure 1: The fuel consumption (left) and the corresponding trajectory (right) of a particular bus route in Chattanooga, TN in February 6th 2020

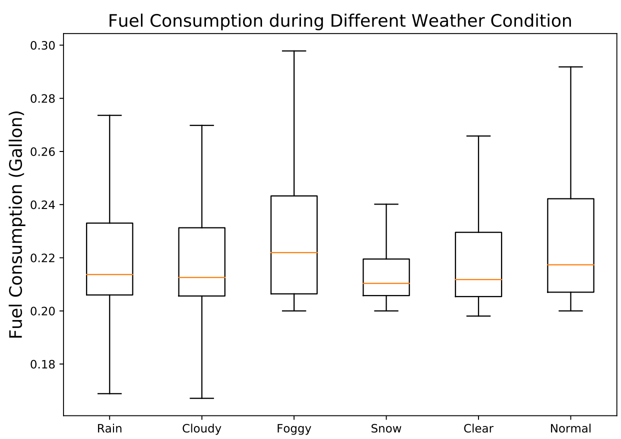


Figure 2: Box plots of fuel consumption of bus vehicles in Chattanooga under different weather conditions.

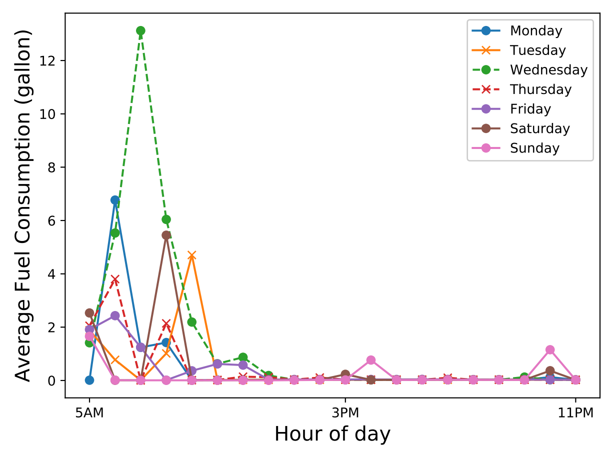


Figure 3: Fuel consumption of bus vehicles in Chattanooga under different time of days.

so buses use more energy to serve its riders. Traffic jams in these areas are more common, which contributes to the fuel usage of bus transportation. For different values of k, we still obtain the same pattern. It indicates that the spatial information of the bus is essential to model the fuel usage of bus system.

Individual bus characteristic: For each bus, we aggregate its energy consumption over hours of day. Then, we calculate the pair t-test for every pair of vehicles. Particularly, the *null hypothesis* is that the energy usage of a pair of vehicles are statistically similar while the *alternative hypothesis* considers that both vehicles have a different energy pattern. The *p-values* of 80% pairs of vehicles are less than 0.05 while the *p-values* of the remaining pairs are from 0.05 to 0.1. Therefore, we conclude that most of vehicles have their own pattern of consuming energy.

The above phenomenon can be explained that despite being manufactured by machines, each bus still has its own characteristics which distinguish itself from others. Moreover, these characteristics can profoundly affect its energy use. For example, each manufacturer has its production procedure so its products are distinct from others. The maintenance also influences the fuel consumption of buses. Due to the complexity of bus characteristics, it is impossible to handpick and model each characteristic individually. For this reason, we use a latent vector to capture all characteristics of a particular bus e.g. made year, manufacturer and such vector is learned automatically by neural network (see later sections for more details).

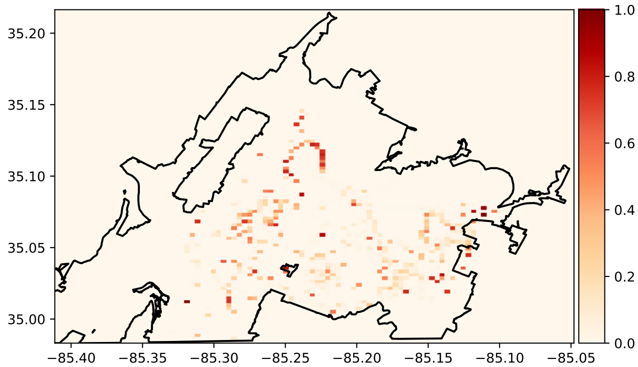


Figure 4: Normalized average fuel consumption of bus transportation in Chattanooga. The darker the color the more energy use the area.

Conclusion: The above analysis shows that fuel consumption is a complex problem since it is under the influence of multiple features especially weather, spatial, temporal features and individual bus characteristic. Hence, there is a need to incorporate these features to our fuel consumption prediction model for gaining more accurate predictive performance.

4 MODEL DESCRIPTION

In this section, we formally present the problem that we aim to study. Then, we will briefly introduce LSTM and describe our FuelPred model which integrates features to enhance the capacity of LSTM to solve the stated problem. Finally, we discuss the loss function for our model and parameter learning process.

4.1 Problem Statement

The input is the sequence of fuel consumption $X = (x_1, x_2, ..., x_{t-1})$ where $x_i \in \mathbb{R}; i \in [1, t-1]$ and for fuel consumption data point x_{t-1}, we also have the associated attributes $\{y_1^{t-1}, y_2^{t-1}, ..., y_N^{t-1}\}$ where N is the number of feature. The objective is to predict the fuel consumption at time t by learning a nonlinear relation from the input and the corresponding associated attributes:

$$\hat{x}_t = F(x_1, ..., x_{t-1}, y_1^{t-1}, y_2^{t-1}, ..., y_N^{t-1}) \tag{1}$$

where F is the non-linear relation that we would like to learn. In this work, we do not consider the historical sequence of attributes for simplicity.

4.2 Preliminary

We use Long-Short-Term Memory (LSTM) [12] method to handle the fuel consumption sequences. The method learns to update the hidden state by using the input and the previous hidden state. The formula below can be used to capture the update process

$$h_t = g(h_{t-1}, x_t) \tag{2}$$

where h_t and h_{t-1} denote the hidden states of LSTM at time t and $t-1$ respectively. $x_t \in X$ is the element of sequence X at time t. The goal of hidden state h_t is to “memorize" latent characteristics of the input sequence up to timestep t. In traditional RNN, function g has a very simple structure such as tanh function. Such simple structures struggle to handle long-term dependencies [5] because of gradient vanish, non-robustness to input noise. To overcome these problems above, LSTM introduces four types of gates and memory cell whose interactions are described by the following equations

$$\begin{aligned}
f_t &= \sigma(\Theta_f x_t + \Omega_f h_{t-1} + b_f) \\
n_t &= \sigma(\Theta_n x_t + \Omega_n h_{t-1} + b_n) \\
o_t &= \sigma(\Theta_o x_t + \Omega_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + n_t \circ \sigma'(\Theta_c x_t + \Omega_c h_{t-1} + b_c) \\
g(h_{t-1}, x_t) &= o_t \circ \sigma'(c_t)
\end{aligned} \tag{3}$$

where f_j, n_j and o_j are corresponding to forget, input, and output gates' activation vectors respectively; c_j is the cell state vector; Θ, Ω, and b are weight and bias parameter; $\sigma(\cdot)$ is sigmoid function; and $\sigma'(\cdot)$ is hyperbolic tangent function [18]. Notation $\circ$ denotes Hadamard product.

To predict the fuel consumption at time $t+1$, we use hidden vector h_t as input for neural network.

$$\hat{x}_{t+1} = \hat{F}(h_t|\theta) \tag{4}$$

where $\hat{F}$ denotes the neural network structure and θ denotes the set of parameters of $\hat{F}$.

4.3 FuelPred Model

The vanilla LSTM in previous section is not able to handle supporting features associated to data point x_t. In this section, we propose a model named FuelPred which is able to incorporate the features into a LSTM structure.

Suppose that we have N features, each feature n is represented by a $N_n \times s_n$ matrix X_n where N_n denotes the number of attributes and s_n is the size of embedding vector of feature n respectively. Note that for feature n, its size of embedding vector s_n can be different from $s_{n'}$ the one of feature n' where $n \neq n'$. To retrieve the representative vector of attribute a_n, we can compute $X_n^T y_n$ where y_n is one-hot vector corresponding to attribute a_n.

In the FuelPred model, instead of using raw x_t as input for LSTM, we combine x_t with its associated attributes of features. Recall that for input x_t, we have the list of its associated attributes $\{y_1^t, y_2^t, ..., y_N^t\}$ of N features. Each attribute is encoded by one-hot vector so from the list of attributes, we derive the list of embedding vector by multiplying each one-hot attribute vector to its corresponding feature matrix i.e. $X_i^T y_i^t$ where $i \in [1, N]$. Then, we concatenate all embedding vectors of the attributes as one supportive vector i.e. $y_t = [X_1^T y_1^t; X_2^T y_2^t; ...; X_N^T y_N^t]$. The order of concatenation is not important since neural network is able to handle that case; however, for consistency, the order in our concatenation step above is similar to the order of features. Finally, the supporting vector y_t is concatenated to the raw input x_t to create a new input vector $\bar{x}_t = [x_t; y_t]$. We then apply the new input vector to LSTM framework. For this reason, Equation 2 changes to

$$\bar{h}_t = g(\bar{h}_{t-1}, \bar{x}_t) \tag{5}$$

The concatenation operator is used in this step because it has some positive effects as below

- It does not assume the length of embedding vector is equal to the length of hidden vector. This relaxation helps us to use the suitable size of embedding vector which can be different from the length of hidden vector. For this reason, the model can achieve better performance.
- The concatenation does not assume any interaction between embedding and hidden vectors and the interaction is learned through neural network architecture. It is different from a matrix factorization method which assumes that the kth element of user latent vector must interact with the corresponding element of item latent factor.
- Moreover, the concatenation operator can be extended to unlimited number of features in theory. Classical collaborative filtering methods model the direct interaction of two considering features and it is difficult for the mechanism to handle more than that number of features. With concatenation, we can ease this limitation.

We can utilize LSTM framework (Equation 3) for function g and use $\bar{h}_t$ for prediction (Equation 4).

4.4 Loss Function & Parameter Learning

We compare the target value and the predicted energy use by using mean square error as loss function [10]. To avoid overfitting, we employ a Dropout regularization technique. Specifically, Dropout discards some components as well as their corresponding structures with probability p at each step of training process. The idea of the technique is to restrict the neural network structure from co-adapting too much.

For parameter learning, we use Adam optimization algorithm [15] to find the optimal values of model parameters that are associated to minimal value of the loss function.

5 EXPERIMENT

In this section, we show the experiment settings and the fuel prediction performance of our FuelPred model and its variants compared with some baselines. We also discuss the prediction performance under different settings.

5.1 Experiment Setup

Proposed Models: Below is the list of our proposed models which are evaluated in the experiments:

- FuelPred (T): FuelPred model with temporal feature. There are seven days per week and 24 hours per day so the number of attributes of temporal feature is 168.
- FuelPred (W): FuelPred model with weather feature. We consider six weather attributes: rain, cloudy, foggy, snow, clear and normal. Each weather attribute is associated to an embedding vector.
- FuelPred (S): FuelPred model with spatial feature. There are 3168 streets in Chattanooga and each street is considered an attribute of spatial feature. For each fuel data point, we use its location to find the corresponding street.
- FuelPred (B): FuelPred model with individual bus characteristic. The number of attribute of bus feature is 28 which is equal to the number of buses in the city fleet. Each bus has an embedding vector that captures its latent characteristics.
- FuelPred (W/S): FuelPred model with weather and spatial features. We combine these two features together because they can be classified as external conditions that affect fuel consumption behavior of buses.

Baselines: To demonstrate the performance of our methods, we compare them with the below baselines:

- Global Mean: We use the mean of all training data points for predicting values in test set.
- Bus mean: The prediction on the testset is done according to the average fuel consumption of the corresponding bus calculated by training data.
- Multi Layer Perceptron (MLP): We use the classical multi layer perceptron method [10] to predict the value in test set. The input of MLP is the concatenation of the embedding vectors of bus, weather, street and temporal features. This baseline is used to illustrate the importance of formulating fuel consumption as sequential model.
- Vanilla LSTM (v-LSTM): We use the sequence of fuel consumption without using any other features i.e. weather, temporal or spatial features.

Evaluation Metrics: We adopt two popular error metrics, mean absolute error (MAE) and root mean square error (RMSE) to evaluate the performances of our methods and the baselines. These two metrics are widely used in sequential prediction problems [19, 20]. The smaller the value of MAE and RMSE, the more accurate the model is. In general, RMSE penalizes more on the large errors and less on smaller ones than MAE does. The two metrics are defined by the two formulas below:

$$
\begin{aligned}
RMSE &= \sqrt{\frac{1}{N_{test}} \sum_{x \in T_{test}} \sum_{t=1}^{L_x} (x_t - \hat{x}_t)^2} \\
MAE &= \frac{1}{N_{test}} \sum_{x \in T_{test}} \sum_{t=1}^{L_x} |x_t - \hat{x}_t|
\end{aligned} \tag{6}
$$

where N_{test} is the number of total data points of all fuel consumption data on test set T_{test} and L_x is the length of a particular sequence x on test set T_{test}.

Parameter Setting: The dimension of each latent feature of each variant of FuelPred model is 50. The hidden layer has the size of 50. The number of epochs for our experiment is 500. We use 0.001 as learning rate for Adam algorithm [15] to optimize the loss function. We randomly split 80% of sequences for training and the rest 20% is used for testing. To initialize the hidden vector of LSTM, we use uniform distribution within the range 0 and 1.

5.2 Experiment Result

5.2.1 Prediction Task. Table 1 shows the prediction performances of variant of FuelPred as well as the baselines. From the table, we obverse that.

- Firstly, the table points out that neural networks-based methods have better performance than the naive ones. For instance, the prediction improvement of MLP over Global Mean is around 15.2% while the one between v-LSTM and Global Mean is 23%. It shows that the fuel prediction problem is much more complicated and it can contain some non-linear

Table 1: Prediction performance of our models and the baselines. The smaller the value, the better the model is. The best performance is highlighted.

Method	MAE	RMSE
Global Mean	2.07	8.11
Bus Mean	1.92	7.23
MLP	1.7971	6.78
v-LSTM	1.6728	5.32
FuelPred (T)	1.237	5.27
FuelPred (B)	1.32	5.30
FuelPred (W)	1.35	5.309
FuelPred (S)	1.248	5.279
FuelPred (W/S)	**1.126**	**5.136**

correlations. Luckily, the neural network has the ability to approximate non-linear relationships and helps us to reveal and make use of these relations.

- Secondly, we observe that the performance of v-LSTM is better than MLP. Specifically, v-LSTM improves 7.43% more than MLP method. It suggests that we need to model the fuel consumption as a sequence and LSTM is a suitable tool for such prediction.
- Thirdly, all variants of FuelPred outperform v-LSTM model. For instance, the prediction performance (MAE) of FuelPred (T) is more than 10% higher than the one of v-LSTM. From this result, we can conclude that all features are useful to predict the fuel consumption of buses. It strengthens our empirical analysis in Section 3.
- Fourthly, among variants of FuelPred, temporal features produce significant improvement compare to other features. Particularly, the RMSE of FuelPred (T) is 5.27 while the closest one among is 5.279 i.e. FuelPred (S).
- Lastly, the combination of two external features i.e., weather and spatial features, show significant improvement compared to other features alone. Moreover, FuelPred (W/S) has the highest performance of all other methods in both metrics. It indicates that these two features have some latent relationship which helps us to contribute to model the more accurate prediction method.

We further apply hypothesis testing to examine if our improvements are significantly better than the baselines. Specifically, the *null hypothesis* states that the performance of our methods and the baselines are statistically similar while the *alternative hypothesis* specifies the performance of our methods are significantly improved compared to the baselines. To achieve the goal, we apply the paired t-tests [13] to compare each FuelPred and v-LSTM. Since the *p-values* of all tests are less than 0.05, we conclude that the performance of each FuelPred is significantly and statistically better than v-LSTM. For that reason, it implies the importance of features such as weather to enhance the modeling of sequential nature of fuel consumption.

5.2.2 Importance of features in training process. The spatial, temporal, weather and bus features do not only benefit the prediction accuracy, they also play an important role in training process. To illustrate this point, we compare the value of loss function of multiple methods over epochs.

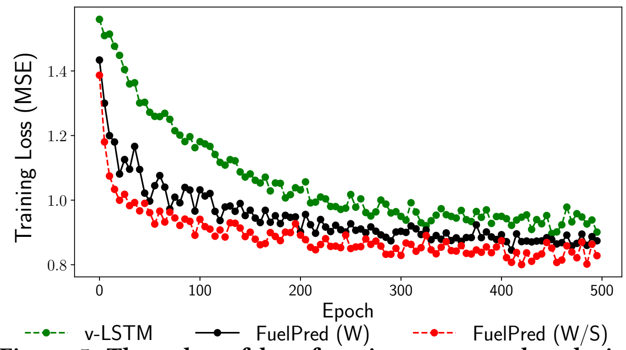


Figure 5: The value of loss function over epoches during training of v-LSTM, FuelPred (W) and FuelPred (W/S).

Figure 5 shows the above task for the three models: v-LSTM, FuelPred (W) and FuelPred (W/S). The naive methods i.e. global/bus mean and MLP are not depicted in the figure because of (i) their poor prediction performance and (ii) the lack of modeling sequential nature. From the figure, we observe that after 500 epochs, the three models seem to converge to their stationary points. Compared to v-LSTM, FuelPred (W) converges faster. Specifically, after more than 100 epochs, FuelPred (W) reaches its convergence point while v-LSTM takes more than 200 epochs. Due to the lack of space, we only depict the value of loss function of FuelPred (W). However, other models with single features i.e. FuelPred (T), FuelPred (B), FuelPred (S) have the same trend with FuelPred (W) which take less epochs to reach the convergence point than v-LSTM. Therefore, it suggests that our proposed features are actually useful not only for prediction, but also for helping us reduce training time. Between the two models FuelPred (W) and FuelPred (W/S), the latter model only takes around 70 epochs for convergence. After 500 epochs, the two models reach the similar training loss but the gap between FuelPred (W) and FuelPred (W/S) is smaller than the one of v-LSTM and FuelPred (W). Hence, we conclude that using spatial and weather features benefits the training process more than weather features exclusively.

5.2.3 Parameter Study Experiment. Next, we study the size of embedding vectors to the prediction performance. For not giving advantages to any features, we use the equal size of embedding vector for all features. We tune the size of embedding vector from 20 to 300 and use MAE as well as RMSE metrics to measure the prediction performance of our FuelPred models.

Table 2 presents the performance of FuelPred (S), FuelPred (W) and FuelPred (W/S) with various setting of the size of embedding vector. From the table, we observe that under different embedding sizes, FuelPred (W/S) consistently performs better than the other two models in the two metrics (except when embedding size is 20). Moreover, FuelPred (W/S) achieves the best performance when the embedding size is equal to 100. In general, increasing the size improves the prediction performance for all FuelPred models. However, the improvement when the size increases 100 to 200 is slower than the one in previous configuration. Increasing the size from 200 to 300 could not lead to the significant improvement in both

Table 2: Prediction performance of FuelPred with various embedding sizes. The smaller the value, the better the model is. The best performance for each model is highlighted.

	MAE			RMSE		
Embedding size	FuelPred (S)	FuelPred (W)	FuelPred (W/S)	FuelPred (S)	FuelPred (W)	FuelPred (W/S)
20	1.278	1.37	1.32	5.311	5.38	5.21
50	1.248	1.35	1.126	5.279	5.309	**5.136**
100	1.245	**1.3312**	**1.1129**	**5.272**	**5.3**	5.139
200	**1.241**	1.333	1.1145	5.3	5.301	5.139
300	1.272	1.351	1.13	5.33	5.38	5.22

metrics. It suggests that the models encounter the overfitting problem. Despite showing results for the above three methods, similar observations are also achieved from other variants of FuelPred, but due to space limitation, we omit their performances in this section.

6 CONCLUSION

In this work, we have investigated the problem of energy use of bus transportation. First, by using real data from the city of Chattanooga, we analyzed the impact of several features (e.g. weather/temporal/spatial features) to the problem. Second, we proposed a method named FuelPred that is based on LSTM to predict the energy consumption of buses. Third, we conducted several extensive experiments to compare the contribution of these features to prediction model. Through experiments, we conclude that FuelPred is better than the baselines and the features play significant roles in the training and testing processes. This paper is one of the first works that use real data to understand the fuel consumption of bus transportation.

There are several directions to extend this research further. In the current prediction model, we have not considered the historical sequence of the features' attributes. Previous works [20] have shown that considering driving series can actually improve the performance of sequential predictive model. We also assume that the skill of all bus drivers is similar, so including this factor may increase the performance of FuelPred model. Furthermore, with the rise of attention mechanisms [3] in neural networks, we can extend FuelPred to measure the in-correlation of features to energy consumption of buses.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Andrew Rodgers, Chattanooga Area Regional Transportation Authority (CARTA) and the Lyndhurst Foundation for their support of this project.

REFERENCES

[1] Tahereh Arabghalizi and Alexandros Labrinidis. 2019. How full will my next bus be? A Framework to Predict Bus Crowding Levels.. In *Proceedings of the ACM SIGKDD international workshop on urban computing (UrbComp '19)*.
[2] Afiya Ayman, Michael Wilbur, Amutheezan Sivagnanam, Philip Pugliese, Abhishek Dubey, and Aron Laszka. 2020. Data-Driven Prediction of Route-Level Energy Use for Mixed-Vehicle Transit Fleets. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*.
[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[4] Sanchita Basak, Abhishek Dubey, and Bruno P. Leao. 2019. Analyzing the Cascading Effect of Traffic Congestion Using LSTM Networks. In *IEEE Big Data*. Los Angeles, Ca.
[5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.
[6] Christopher M Bishop. 2006. *Pattern recognition and machine learning.* springer.
[7] Meng-Fen Chiang, Ee-Peng Lim, Wang-Chien Lee, and Agus Trisnajaya Kwee. 2017. BTCI: A new framework for identifying congestion cascades using bus trajectory data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 1133–1142.
[8] Thanh-Nam Doan and Ee-Peng Lim. 2019. Modeling location-based social network data with area attraction and neighborhood competition. *Data Mining and Knowledge Discovery* 33, 1 (2019), 58–95.
[9] Tolga Ercan, Nuri Cihat Onat, and Omer Tatari. 2016. Investigating carbon footprint reduction potential of public transportation in United States: A system dynamics approach. *Journal of cleaner production* 133 (2016), 1260–1276.
[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning.* MIT press.
[11] Thong Hoang, Pei Hua Cher, Philips Kokoh Prasetyo, and Ee-Peng Lim. 2016. Crowdsensing and analyzing micro-event tweets for public transportation insights. In *IEEE International Conference on Big Data*.
[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[13] Henry Hsu and Peter A. Lachenbruch. 2008. Paired t Test. (2008).
[14] Shan Jiang, Joseph Ferreira Jr, and Marta C Gonzalez. 2012. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD international workshop on urban computing (UrbComp '12)*. 95–102.
[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[16] Agus Trisnajaya Kwee, Meng-Fen Chiang, Philips Kokoh Prasetyo, and Ee-Peng Lim. 2018. Traffic-cascade: Mining and visualizing lifecycles of traffic congestion events using public bus trajectories. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1955–1958.
[17] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks* 6, 6 (1993), 861–867.
[18] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*.
[19] Jianmo Ni, Larry Muhlstein, and Julian McAuley. 2019. Modeling Heart Rate and Activity Data for Personalized Fitness Recommendation. In *The World Wide Web Conference*. 1343–1353.
[20] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. 2017. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. In *IJCAI*.
[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
[22] Jose Stovall, Austin Harris, Amanda O'Grady, and Mina Sartipi. 2019. Scalable Object Tracking in Smart Cities. In *2019 IEEE International Conference on Big Data (Big Data)*.
[23] Yidan Sun, Guiyuan Jiang, Siew-Kei Lam, Shicheng Chen, and Peilan He. 2019. Bus Travel Speed Prediction using Attention Network of Heterogeneous Correlation Features. In *SIAM International Conference on Data Mining,*.
[24] William Vincent and Lisa Callaghan Jerram. 2006. The potential for bus rapid transit to reduce transportation-related CO 2 emissions. *Journal of Public Transportation* 9, 3 (2006), 12.
[25] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2496–2505.
[26] Susan E Williams, Stacy Cagle Davis, and Robert Gary Boundy. 2017. *Transportation Energy Data Book: Edition 36.* Technical Report. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
[27] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.
[28] Ke Zhang, Qiuye Jin, Konstantinos Pelechrinis, and Theodoros Lappas. 2013. On the Importance of Temporal Dynamics in Modeling Urban Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing (UrbComp '13)*.