Predicting Transportation Carbon Emission with Urban Big Data

LU Xiangyong, OTA Kaoru, DONG Mianxiong, YU Chen, JIN Hai

IEEE Transactions on Sustainable Computing

Volume 2
Number 4
Page range 333-344
Year 2017
URL http://hdl.handle.net/10258/00009651

doi: info:doi/10.1109/TSUSC.2017.2728805
Predicting Transportation Carbon Emission with Urban Big Data

Xiangyong Lu, Kaoru Ota, Member, IEEE, Mianxiong Dong, Member, IEEE, Chen Yu, Member, IEEE, and Hai Jin, Senior Member, IEEE

Abstract—Transportation carbon emission is a significant contributor to the increase of greenhouse gases, which directly threatens the change of climate and human health. Under the pressure of the environment, it is very important to master the information of transportation carbon emission in real time. In the traditional way, we get the information of the transportation carbon emission by calculating the combustion of fossil fuel in the transportation sector. However, it is very difficult to obtain the real-time and accurate fossil fuel combustion in the transportation field. In this paper, we predict the real-time and fine-grained transportation carbon emission information in the whole city, based on the spatio-temporal datasets we observed in the city, that is taxi GPS data, transportation carbon emission data, road networks, points of interests (POIs) and meteorological data. We propose a three-layer perceptron neural network (3-layer PNN) to learn the characteristics of collected data and infer the transportation carbon emission. We evaluate our method with extensive experiments based on five real data sources obtained in Zhuhai, China. The results show that our method has advantages over the well-known three machine learning methods (Gaussian Naive Bayes, Linear Regression, Logistic Regression) and two deep learning methods (Stacked Denoising Autoencoder, Deep Belief Networks).

Index Terms—Transportation carbon emission, urban big data, multilayer perceptron neural network, real-time prediction.

1 INTRODUCTION

Transportation carbon emission is the main source of greenhouse gases (GHG). Between 2000 and 2010, carbon emission from transportation sector contributed about 11% to the total annual anthropogenic GHG emissions increase. The Intergovernmental Panel on Climate Change (IPCC) estimates that in the absence of effective emission reduction policies, the baseline global GHG emissions will increase anywhere from 25 to 90 percent between the years 2000 and 2030. With the increases of global GHG emissions, the average global temperatures will continue to rise. The increases in global temperatures will most likely cause our planet’s climate change in ways that will have significant long-term effects on human health and the environment. Given the importance of devising efficient emission reduction strategies, it is essential for policy makers to obtain the real-time and fine-grained information about carbon emission according to local conditions.

Unfortunately, even in the same city, transportation carbon emission differs in different places and relies on multiple factors, such as road traffic, human mobility and structure of road network. For instance, in Zhuhai, we divide the city into disjoint and uniform grids, as depicted in Figure 1(a). It is clear that each grid has different spatial distribution of the road networks and POIs, even though they border each other, as shown in Figure 1(b). The green line segments stand for road segments, the blue dots represent the POI. Figure 2(a) clearly exposes that the transportation carbon emission of different grids is diverse at 5pm on 9/26/2015, even if they are closed to each other. For example, the grid G26 and grid G37 are adjacent, the transportation carbon emission of them is very different. Figure 2(b) further reveals that carbon emission from transportation in an area changes with time of day. The adjacent areas have different time variation curves, such as G18 and G26. These results indicate that the transportation carbon emission of one region close correlate with its spatial and temporal characteristics.

(a) Grids in Zhuhai city
(b) The density of road networks and POI

Fig. 1. The spatial geographical information of Zhuhai city

In this paper, we take advantage of the excellent feature learning ability of multi-layer network architectures. We propose a three-layer perceptron neural network (3-
layer PNN) model to infer the real-time and fine-grained transportation carbon emission in each region, based on the heterogeneous spatio-temporal data sources in the city, such as meteorology, road traffic, human mobility, structure of road networks and POIs. Although, in previous studies, the environment scientists and governmental agencies have proposed some greenhouse gas emission calculation methods and models, such as top-down and bottom-up models [1]. Carbon emission from transportation is usually estimated by separating from the total urban carbon emission from fossil fuel combustion. It is calculated as a simple product of the following factors: fuel consumption, the carbon coefficient of a particular fuel and the percent of fuel that is combusted [2]. However, different degrees of uncertainty will be introduced, when the fossil fuel combustion is measured in different ways. The carbon emission coefficients vary with region and country. Besides, these methods and models often work with limited sets of factors and empirical parameters, which are often not applicable to other regions with different environments. In addition, it is difficult to obtain the real time fossil fuel combustion in the transportation field.

Due to the considerable amount and diversity of the raw urban data, it is not advisable to use them directly to train our model. For achieving excellent prediction performance, we firstly extract five kinds of feature datasets \((F_t, F_{Mo}, F_p, F_{RN}, F_{Me})\) based on the relationships between the datasets and transportation carbon emission. Moreover, combined with the selected features, we train our model to learn the neural network parameters and use the trained model to predict the future transportation carbon emission. Furthermore, we evaluate our method using the real world datasets.

Our work presents a multilayer neural network prediction model to infer the regional transportation carbon emission, leveraging the multiple urban data. The challenges of our approach lie in three aspects. The first is to identify the efficient features from the heterogeneous data sources. The second is the calculation of transportation carbon emission in each area based on the “top-down” method proposed by the IPCC. The third is how to construct the multilayer neural network to get the better performance. The main contributions of this paper include:

1. We devise a multilayer perceptron neural network prediction model, which learns the model parameters leveraging diverse feature datasets. This prediction model could accurately predict the regional transportation carbon emission in the future period of time.
2. We identify five kinds of feature datasets \((F_t, F_{Mo}, F_p, F_{RN}, F_{Me})\), based on the multiple urban data. These feature datasets can not only be applied to our prediction model, but also could be used in the prediction of other automobile exhaust emissions.
3. We evaluate our method based on the multiple datasets in Zhuhai city, that is meteorological data, road networks, POIs, transportation carbon emission data and the GPS trajectories generated by over 3,000 taxicabs in Zhuhai city from August 1st to October 14th, 2015. And we justify the advantages of our method over the other three machine learning algorithms and two deep learning algorithms.

The rest of the paper is organized as follows. We first present the related work in section 2. Then we formally make some preliminary definitions and describe the framework of our system in section 3. We introduce the feature extraction in section 4. Feature learning and inference are discussed in section 5. Experimental evaluation is shown in section 6. Finally, we conclude our work in section 7.

2 RELATED WORK

2.1 Greenhouse Gases Estimation and Prediction

Greenhouse gases provide us with a hospitable living environment by trapping some of the sun’s natural heat. The main sources of GHGs consist of the electricity production, transportation and industry. With the advancement of urbanization and industrialization, GHGs concentrations in the atmosphere continue to increase. Under this background, researches on the statistics and projections of GHGs attract many scientists and research institutions around the world. There are two major approaches to estimate the GHGs emissions, that are “top-down” and “bottom-up” models proposed by the IPCC [1]. The “top-down” approach firstly calculates the total carbon emission from the total fuel consumption in the city. The total carbon emission then is apportioned to each economic sector, such as transportation sector. As carbon emission varies in locations non-linearly, this model would not capture the accurate estimate of the carbon emission according to local conditions. The “bottom-up” approach directly calculates the fuel consumption from each economic sector. Then the amount of carbon emission is calculated according to the carbon emission parameters of the fuel.

However, there are many uncertainties in the calculation of these two models. The uncertainties are mainly from the empirical assumptions and the uncertainty in the primary data: fuel consumption, carbon content, and oxidation factors [3]. When the total carbon emission is allocated to the individual sectors, additional uncertainties are always being introduced. Compared with the previous work in this area, in this paper we calculate the exact amount of carbon emission from the transportation sector, utilizing the multiple local data sources to increase the reliability of input data. What’s more, the increasing numbers of sensors in the city provide us with rich source data. A wide variety of
urban data will not only help us to understand the dynamics of the city, but also help us to optimize the urban structure.

2.2 Urban Computing
With the increasing number of city data, researches on urban computing become more attractive. Urban computing is described in detail in paper [4]. Zheng [5] proposed a semi-supervised learning approach based on a co-training framework to infer the urban air quality, based on the air quality data and a variety of data sources. Karamshuk [6] used the supervised regression model and supervised learning model to optimize the placement of the retail stores. He extracted the features by studying the predictive power of various features on the popularity of retail stores based on the dataset collected from Foursquare in New York. Nevertheless, Zhang [7] combined the supervised and unsupervised machine learning techniques to identify the traffic of zero-day applications in a network. Compared with their interesting and influential work, in this paper we devise a multilayer perceptron neural network model to predict the transportation carbon emission, leveraging multiple urban data sources. The biggest difference between Zheng’s research and mine is that the characteristics of the training methods are different. He divides the feature datasets into two parts, and uses them to train different models. However, we utilize the all feature datasets to train a 3-layer PNN model. Compared with Karamshuk’s research, our training datasets are much more diverse.

Yang [8] modeled the spatial and temporal activity preference separately and then used a principle way to combine them for the preference inference. Fan [9] proposed a predicting-by-clustering framework to predict crowd behavior at a citywide level based on human mobility big data. Shimosaka [10] proposed a low-rank bilinear Poisson regression model to predict the urban dynamics, utilizing the one year’s worth of mobility records. Li [11] proposed a smart city infrastructure to serve people, based on the multisensors data. Sun [12] promoted the concept of smart and connected communities, based on the Internet of Things Technologies and big data analytics. Yang [13] proposed a Wifi-based, real-time monitoring of a carbon monoxide system for application in the construction industry. Wang [14] provided an overview of water cyber-physical systems for sustainability. Compared to their studies, the present paper predict the transportation carbon emission leveraging the spatial-temporal urban data, such as meteorology, road traffic, structure of road networks and POIs. The training datasets are different between their work and mine. we extend the application of multi-layer neural networks to the field of carbon emission analytics.

2.3 Traffic optimization
With the increasing deployment of sensors on the roads and vehicles, it will help us get more urban traffic data. And with the help of these data, a series of researches on city traffic have been done recently, including traffic forecasting, congestion management, vehicle distribution and vehicle routing. Abadi [15] implemented an optimization methodology to predict the flows of a traffic network in San Fransico, based on the real-time and estimated traffic data. Wang [16] presented an interactive system for visual analysis of urban traffic congestion, based on GPS trajectories. Zhang [17] proposed a data-driven system to sense the refueling behavior and citywide petrol consumption in real time, based on a trajectory dataset, POI dataset and road network dataset. Chen [18] proposed an intelligent particle swarm optimization algorithm to develop short-term traffic flow predictors, based on the dataset captured by on-road sensors. Chen [19] adopted the regression-and-ranking methodologies to predict the potential bike trip demand, utilizing the highly variant urban open data. Compared with them, we concentrate on inferring the transportation carbon emission, mainly utilizing the traffic related data. Our work can be seen as an extension of traffic to the environment.

3 Overview
3.1 Preliminary
Definition 1: Transportation carbon emission. The transportation carbon emission is a measure of the total amount of carbon dioxide emissions emitted through the combustion of fossil fuels, which are directly and indirectly caused by the traffic. The IPCC guidelines for national greenhouse gas inventories [1] gives two calculation methods, respectively, “top-down” and “bottom-up” approach. The “top-down” approach is formulated in Equation 1 as follows:

$$ W = \sum_{i \in N} \sum_{j \in N} K_{i,j} \cdot n_{i,j} \cdot s_{i,j} \cdot e_{i,j}, $$

where $W$ represents the total amount of transportation carbon emission. $K_{i,j}$ represents the carbon dioxide emission coefficient of $j$ type of fuel consumed by $i$ vehicle type. $i$ is a kind of vehicle types. $j$ is a type of fuel. $n$ is the number of vehicles. $s$ represents the transport mileage of vehicles. $e$ stands for the average intensity of consumption per unit mileage.

The “bottom-up” approach is formulated in Equation 2 as follows:

$$ W = \sum_{j \in N} (K_j \cdot E_j - M) \cdot \varphi_j \cdot \frac{44}{12}, $$

where $K_j$ represents the carbon dioxide emission coefficient of $j$ type of fuel. $E_j$ represents the amount of carbon emission of $j$ type of fuel. $M$ represents the amount of carbon fixation. $\varphi_j$ stands for the share of fuel oxidation.

Definition 2: Trajectory. A spatial trajectory $T$ is a sequence of GPS points that are time-ordered spatial points, $T: p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n$. Each GPS point $p_i$ consists of a longitude, a latitude and a timestamp.

Definition 3: POI. A POI is a specific point location in the physical world, consisting of a name, category, longitude, latitude. It is a place where a lot of people can find interesting and useful.

Definition 4: Road network. A road network $RN$ is a network structure formed by road segments. Each road segment $r$ connects each others and has different functions and locations.

Definition 5: Grid. We divide a city into disjoint and uniform grids (e.g., $5km \times 5km$ in the experiments) as illustrated in Figure 1(a). Each grid $g$ has its carbon emission $g,W$, which
stands for the total amount of carbon emission in the grid \( g \). The coordinate of each grid center is used as the coordinate of the grid. We assuming the meteorological of the city is uniform in one day.

**Definition 6: Neuron.** A multiple-input neuron is shown in Figure 3. The individual inputs \( P_1, P_2, ..., P_R \) are each weighted by corresponding elements \( w_{11}, w_{12}, ..., w_{1R} \). The \( b \) is a bias. The \( f \) represents the activation function. The \( a \) represents the output of the neuron. The neuron can be formulated as the Equation 3.

\[
a = f(n) = f(\sum_{i=0}^{R} (P_i \cdot W_{1i}) + b), \quad R \in N
\]  

**3.2 Framework**

Our aim is to train a multilayer perceptron neural network to infer the future information of transportation carbon emission throughout the city. Primarily, we collect and filter the relevant source data according to the prior knowledge. Additionally, we preprocess the raw data and extract the features that are closely related to the transportation carbon emission. For obtaining the labeled data, we calculate the amount of transportation carbon emission in each grid based on the processed trajectory datasets. Furthermore, using the extracted features and labeled data, we train the proposed 3-layer-PNN model to learn the model parameters. Accordingly, based on the trained neural network model, we predict the transportation carbon emission of each grid during the future period of time. Our framework consists of four main parts, as shown in Figure 4.

**Collecting data sources**: In this part, we select four real-world datasets, respectively, the taxi trajectories, road networks, POIs and meteorological data. We get the spatial trajectories generated by over 3,000 taxicabs in Zhuhai city during a period of nearly three months. The POI dataset consists of nearly 40,000 POIs in this city. The road network dataset contains about 3,027 road segments.

**Preprocessing datasets**: Each trajectory is extracted from the raw trajectory datasets and mapped onto the road networks to improve the quality of the data. Then we utilize the “top-down” approach to calculate the amount of transportation carbon emission of each grid using the cleaned trajectory data. Accordingly, we obtain the transportation carbon emission dataset. Then we extract the diverse spatio-temporal features that are closely related to transportation carbon emission, based on the correlation and the effect on the prediction accuracy after the fusion with other features. In the process of extraction, we try to keep the original attributes of the source data. Finally, prior to training, we regularize these feature datasets. Detailed in section 4.

**Feature learning**: We feed the extracted features into our 3-layer-PNN model to train the neural network structure and optimize the learning parameters. The transportation carbon emission data is used as the labeled data for the supervised learning. Detailed in section 5.

**Inference**: Based on the trained network structure and the learning parameters, we use the logistic regression method to infer the future transportation carbon emission for each grid once every hour. See section 5 for details.

**Problem statement**

Given a collection of grids \( G = \{g_1, g_2, ..., g_n\} \), where \( g_i \cdot W \) is the carbon emission of the grid \( g_i \), a road network \( RN \) crossing \( G \), a POI located in \( G \), a trajectory dataset \( T \) passing \( G \), and a record of meteorological data in \( G \), we intend to infer the \( g_i \cdot W \) in the future.

**4 Feature Extraction**

**4.1 Traffic Features: \( F_T \)**

The traffic features are the main characteristics to infer the transportation carbon emission. Usually, the gasoline-fueled vehicles have three operating mode conditions, that is cold start, hot start, and hot stabilized. Under different condition, the vehicles have significantly different emission rates. The vehicle speed and travel length are also the important features in the study of transportation carbon emission. Accordingly, we select five relative features for each grid, that is the number of start operations, expectation of speeds , standard deviation of speeds, the number of vehicles and the travel length. Figure 5 shows the correlation matrix between the first four features and transportation carbon emission. These features are extracted from the GPS trajectories generated by vehicles traversing the grid in the past hour.

**Number of start operations**: \( NumSta \). Given a spatial trajectory of a vehicle, we retrieve all the track points that fall in the affecting region of each grid \( p.l \in g.R \). The points contained in the trajectory are strictly increasing in time. We check the two adjacent points. If the speed of the former point \( p_l \) equals to zero, meanwhile the speed of the later point \( p_{l+1} \) is not...
equal to zero, we believe that there is a start event. We count the start action times $NumSta$ according to Equation 4.

$$Sat = Sat + 1, \quad p_i.v = 0, p_{i+1}.v \neq 0,$$

where $p_i, p_{i+1}$ fall in the same grid $g$.

**Expectation of speeds**: $E(v)$. We also retrieve the track points of the trajectories that fall in the affecting region of each grid. As the sampled data of the GPS device is discrete, we use the expectation of speeds to represent all the speeds of passing vehicles in $g.R$. We first design a formula to calculate the distance between two consecutive points, as Equation 5. We then calculate the total driving distance and the traveling time. Finally, we calculate the expectation of speeds according to Equation 6.

$$Dist(p_i, p_{i+1}) = \frac{p_i.v + p_{i+1}.v}{2} \cdot |p_{i+1}.t - p_i.t|,$$

$$E(v) = \frac{\sum Dist(p_i, p_{i+1})}{\sum |p_{i+1}.t - p_i.t|}, \quad p_i.\text{l}, p_{i+1}.\text{l} \in g.R$$

**Standard deviation of speeds**: $D(v)$. As the non-uniform driving has an important influence on the carbon emission of vehicles, we use the standard deviation to represent the discrete degree of the speeds of traveling vehicles in the $g.R$ in the past hour. We calculate this feature according to Equation 7.

$$D(v) = \sqrt{\frac{\sum [E(v) - p_i.v]^2 \cdot |p_{i+1}.t - p_i.t|}{\sum |p_{i+1}.t - p_i.t|}}, p_i.\text{l}, p_{i+1}.\text{l} \in g.R$$

**Number of vehicles**: $NumVeh$. We calculate the number of vehicles entering a grid by retrieving the vehicles’ travel trajectories. If some points of trajectories fall into the affecting region of the grid, we believe that the vehicle has access to the grid. The same vehicle may visit a grid several times in the past hour, while the number of vehicles in the grid is only increased by one.

**Travel length**: $Len$. Based on the Equation 5, we sum all the distance between two consecutive points to calculate the total travel length of all trajectories in the grid $g$. As shown in Equation 8.

$$Len = \sum Dist(p_i, p_{i+1}), \quad p_i.\text{l}, p_{i+1}.\text{l} \in g.R$$

In figure 5, each row or column indicates one feature. Each plot indicates the amount of transportation carbon emission. These features are extracted from a GPS trajectory dataset generated by over 3,000 taxicabs in Zhuhai. As the taxicabs travel accounts for about 4.5% of the total daily travel of residents in the city [20], the dataset is big enough to represent the traffic patterns there [5].

With the increase of $NumVeh$ and $NumSta$, the amount of transportation carbon emission increases significantly, as depicted in the third row and the fourth column. Vehicles are the main sources of transportation carbon emission which mainly comes from the combustion of fossil fuels. Therefore, normally the more emission sources are always producing more emissions. Just as we have known the vehicles will consume more fuel, when vehicles starts. If there are more start operations, the carbon emission should be greatly increased. When $E(v)$ is between $0km/h$ and $40km/h$, the carbon emission gradually increases and reaches the maximum value. When $E(v) > 40$, the larger $E(v)$ is, the more small amount of the carbon emission (e.g., the plot on the first column). It is surprising that not all vehicles emit less carbon emission when they are moving close to a constant speed, as showed in the first row. $D(v)$ and $E(v)$ are closely related to the analysis of transportation carbon emission. In the case of small $D(v)$, with the increase of $E(v)$ the transportation carbon emission will be reduced, i.e., in a traffic jam each vehicle has to move very slowly, the $D(v)$ is small but the carbon emission is relatively large. When $D(v)$ and $E(v)$ simultaneously increase, the carbon emission will gradually decrease. In fact, if there is no traffic jam, the speed of the vehicles on the road is varied and $D(v)$ will be maintained at a relatively large value.

### 4.2 Mobility Features: $F_{Mo}$

Urban transportation is to meet the basic mobility-related needs of human. Digging out the characteristics of human mobility in different areas is necessary to the research on urban mining. Based on the characteristics of human mobility in a certain area, we can evaluate the popularity of this area. It contributes to the inference of transportation carbon emission. In our study, we select two human mobility-related features, that is the number of people arriving at ($numArr$) and leaving ($numLea$) a grid’s affecting region $g.R$ in the pass hour. We extract the two features from the dataset generated by vehicles traversing the grid in the past hour. Given trajectories in the dataset, we retrieve the pickup points($p_{up}$) and the corresponding drop-off points ($p_{off}$) falling in the grid. Obtaining these two features of each grid, we just need to traverse all trajectories once. We calculate the $numArr$ and $numLea$ using the Equation 9.

$$numArr = numArr + 1, p_{up}.l \notin g_i.R, p_{off}.l \in g_i.R$$

$$numLea = numLea + 1, p_{up}.l \in g_j.R, p_{off}.l \notin g_j.R$$

![Fig. 5. Correlation matrix between traffic features and Transportation carbon emission](image)
where \( g_i.R \) and \( g_j.R \) respectively represent the affecting region of the grid \( g_i \) and the grid \( g_j \). numArr indicates the number of people arriving at \( g_i.R \) and numLea indicates the number of people leaving \( g_j.R \). \( p_{up}.l \) and \( p_{off}.l \) respectively represent the location of the pickup points and the location of drop-off points.

In our study, we obtain the POI region’s popularity and regional functions. For instance, a POI region with a lot of people leaving \( g_i.R \) and \( p_{off}.l \) will contribute to the increase of the transportation carbon emission. This is also one reason why we need to separate them.

\begin{equation}
\rho(C_i) = |\{ I : I \in C_i, I.\text{loc} \in g.R \}| \tag{10}
\end{equation}

where \( I.\text{loc} \) is the location of the POI \( I \).

**TABLE 1**
The category of POIs

| C_1 : Vehicle Services (sales, repair) | C_7 : Hotels and Residences |
| C_2 : Food and Beverage | C_8 : Scenic Spots |
| C_3 : Shopping Services | C_9 : Culture and Education |
| C_4 : Life Services | C_10 : Infrastructure Services |
| C_5 : Sports and Leisure | C_11 : Financial Services |
| C_6 : Health Care | C_12 : Companies |

**4.4 Road-network Features:** \( F_{RN} \)

Road networks are very important for road traffic in each area and they also have great influence on the road traffic. The road networks in Zhuhai city are shown in Figure 1(b). In our study, we select the following two related features for each grid \( g \), that is the total length of highways \( f_{highway} \) and the other roads \( f_{road} \). These two features can be formulated by the Equation 11. Figure 7 demonstrates the correlation between transportation carbon emission and this two features, utilizing the road-network dataset extracted from the Auto Navi MAP. The transportation carbon emission is generated by vehicles traversing the grids in the past hour. As described in the figure 7, the transportation carbon emission in each grid is strongly correlated with \( f_{road} \). The longer \( f_{road} \), the more transportation carbon emission. On the surface, there is a small relationship between transportation carbon emission and \( f_{highway} \). However, it is clear that \( f_{road} \) and \( f_{highway} \) have very different relationships with transportation carbon emission. This is also one reason why we need to separate them.

\begin{equation}
\begin{aligned}
f_{highway} &= \sum \text{len}(r_{highway}), \quad r_{highway} \in g.R, \\
f_{road} &= \sum \text{len}(r_{other}), \quad r_{other} \in g.R 
\end{aligned} \tag{11}
\end{equation}

where \( \text{len}(r_{highway}) \) and \( \text{len}(r_{other}) \) are the length of the highway \( r_{highway} \) and the other road \( r_{other} \) respectively.

**4.5 Meteorological Features:** \( F_{Me} \)

The meteorology could potentially affect our travel plans or possibly cause us to choose short distance travel, the results will indirectly affect the amount of transportation carbon emission. Accordingly, in this section, our purpose is to identify what useful information from the meteorology can contribute to the analysis of transportation carbon emission. In the experiment, we identify two features for each grid: weather \( f_w \) and temperature \( f_t \), which can be formulated as the Equation 12. Figure 8 illustrates the relationship between
In this section we combine the extracted features to train our multilayer perceptron neural network and optimize the learning parameters. We aim to utilize the combination of multiple temporal and spatial features to improve the prediction. Nowadays some deep neural networks have been used in many fields such as image recognition and speech recognition, and have achieved very good results, but for our data characteristics, the performance of the deep learning algorithms is not very good. Based on the characteristics of extracted data, we propose a 3-layer perceptron neural network to infer the transportation carbon emission in the future period of time.

### 5.1 3-layer perceptron neural network (3-layerPNN)

The 3-layerPNN predicts the transportation carbon emission, utilizing the extracted features and the labeled data. The extracted features consist of $F_T$, $F_Mo$, $F_P$, $F_RN$ and $F_Me$. The labeled data is calculated using the “top-down” approach, which is formulated in Equation 1. The main training and learning algorithm for our 3-layerPNN is given in Algorithm 1. The graphical structure of our 3-layerPNN is shown in Figure 9.

$$f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}, \quad (13)$$

In this Figure, our 3-layer perceptron neural network consists of three layers, that is the input layer, the hidden layer and the output layer. The hidden layer contains $n$ nodes. The output layer contains $k$ nodes. We use the solid circles to denote the inputs of the network and utilize hollow circles to denote the multiple-inputs neurons. The solid circles labeled “+1” are used to represent the bias units and correspond to the intercept term $b_i$. Our neural network has parameters $(W, b) = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$, where $W^{(l)}_{i,j}$ is the weight associated with the connection between unit $j$ in layer $l$ and unit $i$ in layer $l + 1$. $b^{(l)}_i$ denotes the bias associated with unit $i$ in the layer $l + 1$. $a^{(l)}_i$ denotes the output value of unit $i$ in the lay $l$. The transfer function $f$ of each neural in the hidden layer is the activation function $tanh$, as designed in Equation 13.
Algorithm 1 The training and learning algorithm

Input: A set of features \( F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \), the labeled data \( D_t \) for each grids, a threshold \( \alpha \) controlling the number of epochs to run the optimizer.

Output: A set of learning parameters \( \theta \) of the neural networks.

1: \( 3\text{-layer} PNN \leftarrow \text{construct a three layers perceptron} \)
2: \( \theta \leftarrow \text{initialize the parameters} \theta \)
3: \( D \leftarrow \{ F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \} \)
4: \( \text{epoch} \leftarrow 0 \)
5: while \( \text{epoch} < \alpha \) do
6: \( \text{epoch} \leftarrow \text{epoch} + 1 \)
7: while True do
8: \( \text{miniBatch} \leftarrow \text{read one unit from} D \text{ and} D_t \)
9: if \( \text{miniBatch} \text{ is empty} \) then
10: \( \text{break} \)
11: end if
12: \( \text{put miniBatch into the Stochastic Gradient Descent algorithm} \)
13: \( \text{minimize the loss function of} \ 3\text{-layer} PNN \)
14: \( \text{for each sample} \)
15: \( \text{update the parameters} \theta \text{ of} 3\text{-layer} PNN \)
16: end while

The output layer uses the logistic regression to analyze and predict the transportation carbon emission utilizing the input data of this layer, as formulated in Equation 15. The prediction \( y_{pred} \) is the class whose probability is maximal, as formulated in Equation 15.

\[
P(Y = i|x, W^{(2)}, b^{(2)}) = \text{softmax}(W^{(2)}x + b^{(2)}) = \frac{e^{W^{(2)}x + b^{(2)}}}{\sum e^{W^{(2)}x + b^{(2)}}},
\]

\[
y_{pred} = \text{argmax}_i P(Y = i|x, W^{(2)}, b^{(2)}),
\]

where \( Y \) denotes a stochastic variable. \( x \) denotes the input vector. \( W^{(2)} \) is the weight matrix between units in the hidden layer and units in the output layer. \( b^{(2)} \) is the bias vector associated with units in the output layer.

Formally, the computation that this 3-layer perceptron neural network represents is given by the equations as follows:

\[
a^{(2)} = f(W^{(1)}x + b^{(1)}),
\]

\[
a^{(3)} = \text{softmax}(W^{(2)}a^{(2)} + b^{(2)}),
\]

where \( a^{(2)}, a^{(3)} \) respectively denotes the output vector of the hidden layer and the output layer. \( b^{(1)} \) respectively denotes the bias vector associated with units in the hidden layer. \( W^{(1)} \) is the weight matrix between units in the input layer and units in the hidden layer.

To train our neural network, we use the negative log-likelihood as the loss. We use the stochastic gradient descent (SGD) to minimize the loss function to optimal the parameters, then update all parameters of the neural networks. The likelihood \( \mathcal{L} \) and the loss \( \ell \) are respectively defined as follow:

\[
\mathcal{L}(\theta, D) = \sum_{i=0}^{D} \log(P(Y = y^{(i)}|x^{(i)}, \theta)),
\]

\[
\ell(\theta, D) = -\mathcal{L}(\theta, D),
\]

where \( \theta \) is the set of parameters of the neural networks, \( \theta = \{W^{(2)}, b^{(2)}, W^{(1)}, b^{(1)}\} \). \( D \) is the training data set.

6 Experiments

6.1 Datasets

In the experiments, we validate our prediction system using the five kinds of real world datasets in Zhuhai. We show these datasets in Table 2. Details are as follows.

**Taxi trajectories:** The GPS trajectory dataset is generated by over 3,000 taxicabs in Zhuhai city from August 1st to October 14th, 2015. We use this dataset to calculate the \( F_T \) and \( F_{Mo} \). Every taxicab generates a GPS record every 10 seconds and lasts 24 hours a day. All the taxicabs can produce about 38 million records and generate about 0.116 million kilometers in one day. According to the annual report of the traffic development in Zhuhai in 2014, the residents travel about 4.5 million times a day, and the taxicabs and small passenger cars account for about 31.8%. Each GPS record contains vehicle ID, record time, latitude, longitude, speed, direction and passenger status, etc. Utilizing map matching, we can easily count the number of taxicabs in each region and detect the condition of taxicabs on each road in the past hour. In our study, this dataset is big enough to represent the traffic patterns there. In Zheng’s paper [5], he also uses the taxicab dataset to represent the traffic patterns.

**Transportation carbon emission:** There are two widely accepted approaches to calculate the transportation carbon emission, that is “top-down” and “bottom-up” approaches defined by IPCC(2006) [1]. The data required by the “bottom-up” approach is very difficult to obtain, and there is no complete data provided to this approach. However, the “top-down” approach can effectively avoid the difficulty of data acquisition. Accordingly, we use the “top-down” approach to calculate the transportation carbon emission in our experiments. The carbon emission coefficient of the taxicab is 0.5 kg CO₂/km, that is defined by the ministry of science and technology of China according to China’s national conditions [21]. Additionally, the carbon emission coefficient in different countries may exist differences.

**POIs:** We extract the POI dataset of Zhuhai from the Auto Navi MAP. Each record of the POI dataset contains the POI’s ID, name, type, latitude and longitude, etc. The POIs are divided into multiple parts, as described in Table 1. The number of the POIs in each part is shown in Figure 10. The geographical distribution of the POIs is described in Figure 1(b).

**Road-networks:** The road-network data of Zhuhai is also extracted from the Auto Navi MAP. Each record of this dataset contains the road-network’s ID, road class, length and the location of the starting point, mid point and end point of the road, etc.

**Meteorological data:** We collect the meteorological data from a public website every day from August 1st to October 31st, 2015 in Zhuhai. Each record of this dataset consists of two messages, that is temperature and weather.
6.2 Comparison Methods

According to the characteristics of our datasets, we compare our method with three typical machine learning algorithms (Gaussian Naive Bayes, Linear Regression, Logistic Regression) and two deep learning methods (Stacked Denoising Autoencoder, Deep Belief Networks).

1) Gaussian Naive Bayes (GaussianNB): In many practical applications, the naive Bayes classifiers can be trained very efficiently in a supervised learning setting [22]. We compare with the Gaussian Naive Bayes, which is shown in Equation 19.

\[ p(x = v | c) = \frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(x - \mu_c)^2}{2\sigma_c^2}}, \] (19)

where \( x \) denotes a continuous attribute of the training data. \( v \) is some observation value. \( c \) denotes one class of the \( C \).

2) Linear Regression (LinearR): A linear regression model assumes that the relationship between the dependent variable \( y_i \) and the \( p \)-dimensional vector of regressors \( x_i \) is linear [23]. Given a dataset \( \{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n \), the linear regression prediction model can be formulated as follows:

\[ y_i = \beta_1x_{i1} + \cdots + \beta_px_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n \] (20)

where \( y_i \) denotes the regressand, \( x_{i1}, x_{i2}, \cdots, x_{ip} \) are the input variables. \( \beta_1, \beta_2, \cdots, \beta_p \) are the regression coefficients.

3) Logistic Regression (LogisticR): The Logistic regression utilize a logistic function to measure the relationship between the categorical dependent variable and one or more independent variables [24]. The logistic regression prediction model can be written as the Equation 14. In this Equation, the function \( \text{softmax} \) is a standard logistic function.

4) Stacked Denoising Autoencoder (SDA): The SDA is a deep neural network stacked by the denoising autoencoders (DA) [25]. Usually, the performance of the deep learning models is better when learning high level features. In this experiment, we construct a deep learning neural network with five layers, including one input layer, three hidden layers and one output layer. These hidden layers are built by the DA. Each hidden layer consists of 100 neurons. The output of each hidden layer is the input to the next layer. The output layer is a logistic regression layer. The fine-tuning loop of the SDA is very similar to that in our 3-layer PNN.

5) Deep Belief Networks (DBN): The DBN is a deep neural network stacked by the restricted Boltzmann machines (RBM). The DBN is a graphical model which learns to extract a deep hierarchical representation of the training data [26]. In this experiment, we construct a DBN with four layers, including one input layer, two hidden layers and one output layer. we use the RBM to build the hidden layers. Each hidden layer consists of 100 neurons. The output layer is also the logistic regression layer.

6.3 Results

Evaluation on Features: We first compare and analyze the prediction performance of the single feature dataset and the fusion of multiple feature datasets. We use these different feature datasets to train the 3-layer PNN model, respectively. Table 3 shows the training results of five representative combinations of the feature datasets (\( F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \)). As described in the first three rows in the table, the prediction performance of the fusion of \( F_T \) and \( F_{Mo} \) is better than the performance of \( F_{Mo} \) and worse than the performance of \( F_T \). The prediction accuracy of the fusion of \( F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \) is best. The prediction accuracy of the model is 90.86% utilizing the fusion of \( F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \). The results indicate that the performance of the fusion of multiple feature datasets is not always superior to the performance of the single feature dataset. However, when combined with \( F_T, F_{Mo}, F_P, F_{RN}, F_{Me} \), the accuracy rate of prediction becomes better. It proves that the feature datasets we selected are very efficient. In addition, these results also suggest that feature selection is very importance for neural network training and learning.

### Table 3: The results related to features

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy of the prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_T )</td>
<td>88.87 %</td>
</tr>
<tr>
<td>( F_{Mo} )</td>
<td>77.45 %</td>
</tr>
<tr>
<td>( F_T + F_{Mo} )</td>
<td>88.68 %</td>
</tr>
<tr>
<td>( F_P + F_{RN} + F_{Me} )</td>
<td>66.02 %</td>
</tr>
<tr>
<td>( F_T + F_{Mo} + F_P + F_{RN} + F_{Me} )</td>
<td>90.86 %</td>
</tr>
</tbody>
</table>
Overall Results: Leveraging the extracted five kinds of feature datasets ($FT, F_{Mo}, F_P, F_{RN}, F_{Me}$), we train the proposed 3-layer PNN and the other five aforementioned methods. The performance of these algorithms is shown in Figure 11. The predication accuracy of the 3-layer PNN is 90.86%. It is clear that the 3-layer PNN is more accurate than other algorithms. The results demonstrate that our method has the advantage to infer the transportation carbon emission over the other five methods. Additionally, the predication accuracy of SDA and DBN is 32.348% and 32.306% respectively. The deep learning algorithms utilize several processing layers to model high-level abstractions in data, which may lead to the distortion of training data, especially for the case of fewer feature datasets. Moreover, compared to the two deep learning methods with the other methods in our experiments, it is also shown that the predication performance of the neural network with more processing layers is not always better in the transportation carbon emission areas, and the deep learning algorithm is difficult to obtain the efficient high-level abstractions based on the manually extracted feature datasets.

![Fig. 11. Overall results of different methods](image)

7 Conclusion

In this paper, from the perspective of big data, we predict the future transportation carbon emission based on five real world datasets (taxi trajectories, road networks, POIs, meteorological data and carbon emission data) observed in Zhuhai city. We identify five kinds of feature datasets ($FT, F_{Mo}, F_P, F_{RN}, F_{Me}$) based on the relationship between the real world datasets and transportation carbon emission. Additionally, using the “top-down” method, we calculate the amount of transportation carbon emission as the labeled data. Accordingly, we train the proposed 3-layer PNN model to infer the transportation carbon emission in the future, utilizing the extracted features and the labeled data. We evaluate our method on the basis of the spatio-temporal data obtained in the city. The results show the predication accuracy of our method being above 90.86%, which is better than the accuracies of the other methods. Moreover, the results also demonstrate that our method is superior to the well-known machine learning algorithms (Gaussian Naive Bayes, Linear Regression, Logistic Regression) and deep learning algorithms (Stacked Denoising Autoencoder, Deep Belief Networks).

The key experiences we learn from the research lie in three aspects. Primarily, feature selection and fusion are very important for neural network training and learning. Secondly, when the extracted features are relatively few, the performance of deep learning methods may be not ideal. For instance, the predication accuracies of the SDA and DBN algorithms are worse than the other machine learning algorithms in our experiments. Finally, the selection of the learning algorithms should be based on the unique characteristics of dataset.

In the future, we plan to extend this work in several directions. First of all, we will apply our method to more cities, in order to better characterize transportation carbon emission. Then, to reduce carbon emissions, we will consider the urban planning and vehicle scheduling to assist the decision making. Given the different amount of vehicle exhaust emissions in all regions, the management departments should optimize the traffic in different ways to cut the regional emissions. Subsequently, we would further study the data mining algorithms and theory to design more excellent algorithms for the heterogeneous urban data.

Acknowledgments

The work is supported by NSFC (No.61472149), the Fundamental Research Funds for the Central Universities (2015QN67), the National 863 Hi-Tech Research and Development Program under grant (2015AA01A203), and JSPS KAKENHI Grant Number JP16K00117, JP15K15976, KDDI Foundation.

References


Hai Jin is a Cheung Kung Scholars Chair Professor of computer science and engineering at Huazhong University of Science and Technology (HUST) in China. Jin received his PhD in computer engineering from HUST in 1994. In 1996, he was awarded a German Academic Exchange Service fellowship to visit the Technical University of Chemnitz in Germany. Jin worked at The University of Hong Kong between 1998 and 2000, and as a visiting scholar at the University of Southern California between 1999 and 2000. He was awarded Excellent Youth Award from the National Science Foundation of China in 2001. Jin is the chief scientist of ChinaGrid, the largest grid computing project in China, and the chief scientists of National 973 Basic Research Program Project of Virtualization Technology of Computing System, and Cloud Security.

Jin is a senior member of the IEEE and a member of the ACM. He has co-authored 22 books and published over 700 research papers. His research interests include computer architecture, virtualization technology, cluster computing and cloud computing, peer-to-peer computing, network storage, and network security.