Prediction of Transportation Carbon Emission using Spatio-temporal Datasets and Multilayer Perceptron Neural Network

G. Uma Mahesh
Dept of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, A.P, India

M.Thrilok Reddy
Dept of Information Technology
Sree Vidyanikethan Engineering College
Tirupati, A.P, India

Abstract: At present greenhouse effect is the major problem in urban cities. This effect is caused due to the release of greenhouse gases like carbon dioxide, water vapour, methane etc. Transportation carbon emission is major source of greenhouse gases. These gases harm the human health and climate in the environment. It is very essential to know the information about transportation carbon emission in real time. Old way is to get the information of transmission carbon emission by calculating the combustion of fossil fuel. The proposed method finds the prediction accuracy of carbon emission from taxicabs information in the whole city using spatio-temporal datasets observed in the city, that is taxi GPS data, transportation carbon emission data, road networks, points of interests (POIs) and meteorological data. Proposed system uses a Multi-layer perceptron neural network (MultilayerPNN) method to learn the characteristics of collected data and to know the transportation carbon emission. It evaluates method with extensive experiments based on five real data sources.

Keywords: Transportation carbon emission, urban big data, multilayer perceptron neural network.

I. INTRODUCTION

Transportation carbon emission is the main source of greenhouse gases (GHG).Transportation sector contributed about 11% to the total annual anthropogenic GHG emissions increase. It is essential for policy makers to obtain the real time and fine grained information about carbon emission.

Even in the same city, transportation carbon emission differs in the different places and multiple factor, such as road traffic, human mobility and structure of road networks. 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. Taking advantage of the excellent feature learning

ability of multi-layer network architectures, this model proposes a multi-layer perceptron neural network which learns the model parameters leveraging diverse feature datasets. 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. At first 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, will train our model to learn the neural network parameters and use the trained model to predict the future transportation carbon emission. A Multi-layer perceptron is proposed to infer the real-time carbon emission in each region based on heterogeneous spatio-temporal datasets. The main contributions of our project include:

- 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 expansion of worldwide temperatures causes our planet's environmental change in the ways, which will have critical consequences for human well being and condition.
It is important to acquire the ongoing and fine-grained data about carbon emission as per nearby conditions. Unfortunately even in the same city, transportation carbon emission differs in different places and depends on multiple factors, such as road traffic, human mobility and structure of road network.
We propose a Multi-layer perceptron neural network (multi-layer PNN) model to infer the real-time and fine-grained transportation carbon emission in each region, based on heterogeneous spatio-temporal data sources in the city, such as meteorology, road traffic, human mobility, structure of road networks and POIs. The environment scientists and governmental agencies have proposed some greenhouse gas emission calculation methods and models, such as top-down and bottom-up models.

II. 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. There are two major approaches to estimate the GHGs emissions, that are "top-down" and "bottom-up" models proposed by the IPCC[4]. 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. 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[2].

2.2 Urban Computing

Devising a multilayer perceptron neural network model to predict the transportation carbon emission, leveraging multiple urban data sources. However we utilize the all feature datasets to train a multi-layerP NN model. Extending the application of multi-layer neural networks to the field of carbon emission analytics. Yang [5] proposed a Wifi-based, real-time monitoring of a carbon monoxide system for application in the construction industry.

2.3 Traffic optimization

With the increasing deployment of sensors on the roads and vehicles, it will help us get more urban traffic data. concentrating 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.

III. OVERVIEW

3.1 Preliminary

Definition 1: Transportation carbon emission.
It is a measure of the total amount of carbon dioxide emissions emitted through the combustion of fossil fuels.
The IPCC guideline for national greenhouse gas gives two calculation methods.

They are: 1) Top-down approach
2) Bottom-up approach

Top-down approach is formulated in Equation 1 as follows:

$$W = \sum_{i \in N} \sum_{j \in N} K_{i,j}.n_{i,j}.s_{i,j}.e_{i,j} \quad (1)$$

Where W is the total amount of transportation carbon emission.
K_{ij} 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 type of fuel.
n is number of vehicles.
s represents the transport mileage of vehicles.
e stands for the average intensity of consumption per unit mileage.

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

Definition 3: POI.
POI is a specific point location in the physical world, consisting of a name, category, longitude, latitude. Indicate the land use and the function of the region.

Definition 4: Road network.
Have a strong correlation with traffic flows. A good complementary of traffic modeling.
Total length of highway f_h,
Total length of other road segments f_r.

Definition 5: 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}, $w_{12, \dots,}$ $w_{1R.}$
b is a bias.
The f represents the activation function.
a represents the output of the neuron.
The neuron can be formulated as the equation 2.

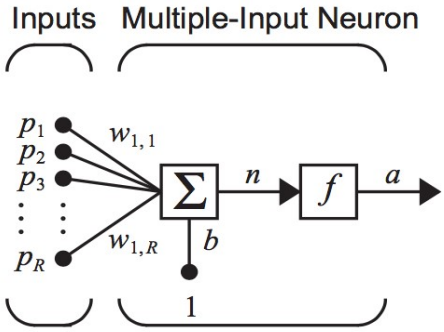

Fig. 1. Multiple-Input Neuron

$$a = f(n) = f(\sum_{i=0}^{R}(P_i.W_{1i}) + b),\ R \in N \qquad (2)$$

3. 2 Framework

To infer the future data of transportation carbon emission all through the city multilayer perceptron neural system is trained. Basically gather and filter the important source information as indicated by the earlier learning. Preprocess the raw data and concentrate the highlights that are firmly identified with the transportation carbon emission.

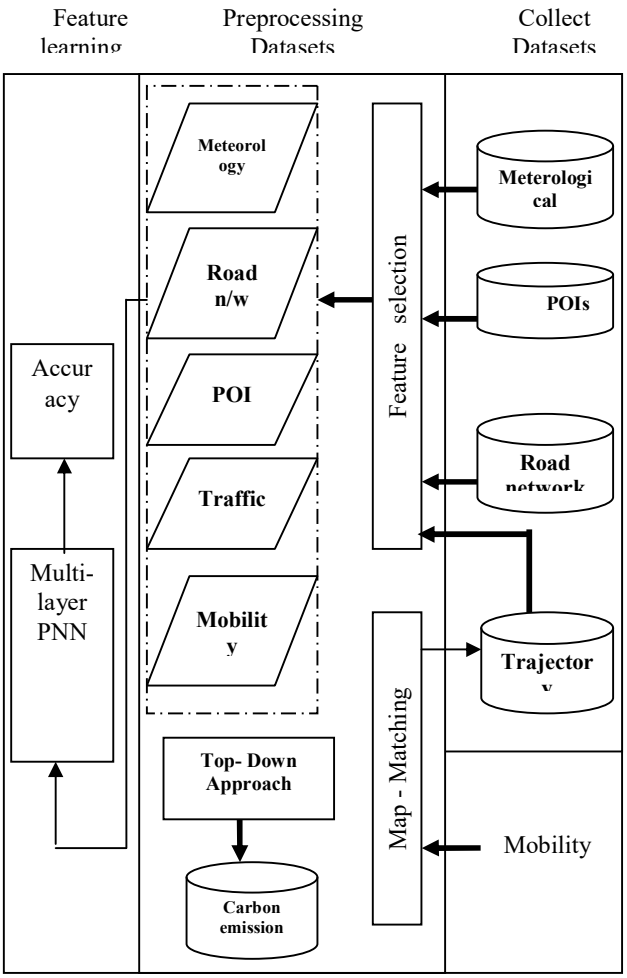

Fig. 2. Framework for the proposed system

Collecting data sources:
Four real world datasets are collected, that are taxi trajectories, road networks, POIs and meteorological data.
Preprocessing datasets:
Every direction is extracted from the raw direction datasets and mapped onto the road network to enhance the quality of the information. At that point we use the "top-down" way to to calculate the amount of transportation carbon emission of every grid utilizing the cleaned direction information. Likewise, we get the transportation carbon emission dataset.
Then extract the diverse spatio temporal features that are closely related to transportation carbon emission.
Feature learning:
Feed the extracted highlights into our multi-layer PNN model to prepare the neural network structure and optimize the learning parameters. The transportation carbon emission information is utilized as the labeled data for the supervised learning.
Inference:
Based on trained network structure and the learning parameters, use the logistic regression method to infer the future transportation carbon emission.
Problem Statement:
Given a collection of grids $G = \{g_1, g_2, .., g_n\}$, where gi.W is the carbon emission of the grid gi, 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 gi.W0 in the future.

IV. FEATURE EXTRACTION

4.1 Traffic Features: F_T

Traffic features are important characteristics in inferring the transportation carbon emission. Usually, the vehicles have different emission rates in different operating conditions like cold start, hot start and hot stabilized. Apart from this, vehicle speed and travel length are also important features in transportation carbon emission. Mainly has two features for each area, the number of vehicles and the travel length . These features are extracted from GPS trajectories generated by vehicles traversing the area at a particular time.

Number of vehicles: NumVeh.

The number of vehicles entering into an area is calculated by retrieving the vehicles travel trajectories. It is believed that the vehicle has to access an area if some points of trajectories fall into affecting region of an area. The same vehicle may visit a grid several times in an hour, but the number of vehicles is increased by one.

Travel length: Len

To calculate total travel length of all trajectories , we sum the total distance between two consecutive points as shown in equation 3.

$$Len = \sum Dist(p_i, p_i+1),\ p_i.l, p_i+1.l \in g.R \qquad (3)$$

These features are extracted from GPS trajectory dataset over taxicabs. Usually, the more emission sources always produce more emissions. Just we know that the vehicle will consume more fuel when the vehicle starts. So the transportation carbon emission is different and more than the usual times.

4.2 Mobility Features

Urban transportation is to meet the fundamental necessities
of human. Uncovering the qualities of human mobility in various areas is important to the research on urban mining. Based on characteristics of human mobility in a specific region, we can assess the notoriety of this region. It adds to the deduction of transportation carbon discharge. In our examination, we select two human mobility related highlights that is number of people arriving (numArr) and leaving (numLea) a framework's influencing district g.R in the pass hour. We remove the two features from the dataset produced by vehicles navigating the grid in the previous hour. Given directions in the dataset, we retrieve the pickup points(p_{up}) and the relating drop-off focuses (p_{off}) falling in the grid. Getting these two features of every grid, we just
Need to traverse all directions once.
numArr and numLea are calculated using the equation 4

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

$$numLea = numLea + 1,\ p_{up}.l \in g_i.R,\ p_{off}.l\ \ g_j.R \qquad \text{--- (4)}$$

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 $_R$ 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.

4.3 POI FEATURES:

POI is a specific point location in the physical world, consisting of a name, category, longitude, latitude. The classification and density of POIs tend to imply that the region's popularity and regional functions. The density of
POIs in each area is very different. The POI dataset can be divided into 12 classes, namely $\{C_1, C_2, \cdots, C_{12}\}$.

Table 1: The category of POIs.

C1: Vehicle Services(sales, repairs)	C7: Hotels and Residences
C2: Food and Beverages	C8: Scenic Spots
C3: Shopping Services	C9: Culture and Education
C4: Life Services	C10: Infrastructure services
C5: Sports and Leisure	C11: Financial Services
C6: Health Care	C12: Companies

4.4 Road-network Features:

Road networks are very important for road traffic in each area and they also have great influence on the road traffic. It include 2 Features for each grid g, that is the total length of highways fhighway and the other roads froad. The transportation carbon emission in each grid is strongly correlated with froad. The longer froad, the

more transportation carbon emission. On the surface, there is a small relationship between transportation carbon emission and $f_{highway}$.

$$f_{highway} = \sum len(r_{highway}),\ r_{highway} \in g.R, \qquad (5)$$
$$f_{road} = \sum len(r_{other}), r_{other} \in g.R\ ,$$

Where $len(r_{highway})$ and $len(r_{other})$ are the length of the highway r highway and the other road r other respectively.

4.5 Meteorological Features:

The purpose is to identify what useful information from the meteorology can contribute to the analysis of transportation carbon emission. The two features for each grid are weather f w and temperature f_t.
F_w= weather, weather ∈ M
F_t = temp, temp ∈ Z
where the weather is the weather in the grid g. The M represents the set of weather types. The temp is the temperature in the grid g.

V. LEARNING AND INFERENCE

In this, the extracted features are combined to train our multilayer perceptron neural network and optimize the learning parameters. The main purpose is to utilize the combination of multiple temporal and spatial features to improve the prediction. Here this model proposes a multi-layer perceptron neural network to infer the transportation carbon emission in the future period of time.

5.1 Multi-layer perceptron neural network
The Multi-layer PNN 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.

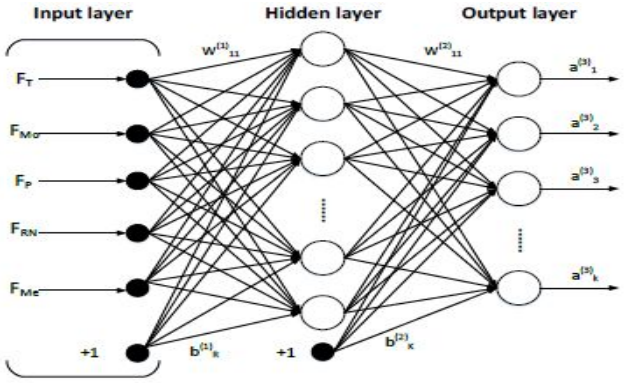

Fig. 3. Structure of Multi-layer PNN

In the Figure3, our multi-layer perceptron neural network consists of six layers that is the input layer, 4 hidden layers
and the output layer. Each hidden layer contains n nodes. The output layer contains k nodes. The neural network has
the following parameters if it has one hidden layer

$$(W, b) = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}).$$

where $W^{(1)}$ is the weight matrix between units in the input layer and the units in the hidden layer.$W^{(2)}$ is the weight matrix between units in the hidden layer and the units in the output layer.$b^{(1)}$ is the bias vector associated with units in the hidden layer.
$b^{(2)}$ is the bias vector associated with units in the output layer.

Training and learning algorithm

Input: A set of features (F_T, F_{Mo}, F_P, F_{RN}, F_{Me}), the labeled data Dl for each grids.

Output: Accuracy of multi-layer perceptron

1: 3-layerPNN ← construct a three layers perceptron
2: D ← {F_T , F_{Mo}, F_P , F_{RN}, F_{Me}}
3:Divide the dataset into test and train part
4: epoch ← 0
5:training_epoch←100
6:learning_rate←0.3
7:Number of neurons for each hidden layer are defined
8:weights:h1,h2,h3,h4 and biases:b1,b2,b3,b4 are defined
9: while epoch < training_epoch do
10: epoch ← epoch + 1
11: while True do
12: miniBatch ← read one unit from D and Dl
13: if miniBatch is empty then
14: break
15: end if
16: put miniBatch into the Stochastic Gradient Descent algorithm using learning rate.
17: minimize the loss function of 3-layer PNN for each sample
18: update the parameters θ of 3-layer PNN
19: end while
20: end while
21: calculate accuracy using test dataset.
22:print final accuracy using argmax()

VI. EXPERIMENTS

6.1 Datasets

In this model, five types of real world datasets are used to validate the proposed system. Detailed description of these datasets is described in Table2.
The five real world datasets are:

Taxi trajectories: The GPS trajectory dataset is generated by over taxi cabs. This dataset is used to calculate the F_T and
F_{Mo}. According to the annual report of the traffic development, 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. Using map matching, numbers of taxicabs in each region is counted easily and detect the condition of taxicabs on each road in the past hour. This dataset is used to represent traffic patterns. In zheng's[6] paper also same taxicab dataset is used to represent the traffic patterns.

Transportation carbon emission: Usually, there are two approaches to calculate the transportation carbon emission. They are "top-down" and "bottom-up" approaches defined by IPCC[4]. The data required for 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. The carbon emission coefficient in different countries may be different.

Mobility: Each record of mobility has number of people arriving and number of people leaving.

POI's: Each record of the POI dataset contains the POI's ID, name, type, latitude and longitude, etc. The POIs are divided into multiple parts. The number of the POIs in each part is shown in Figure 4.

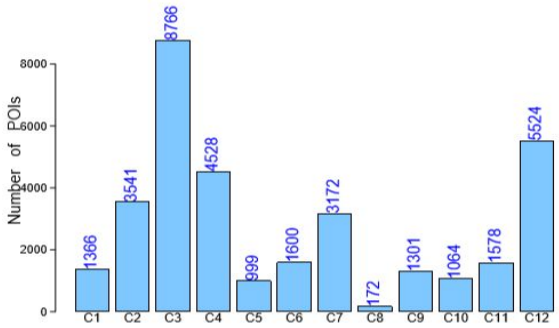

Fig 4: Number of POI's in different categories

Meteorological data: We collect the meteorological data from a public website every day. Each record of this dataset consists of two messages, that is temperature and weather.

Table 2: Details of the datasets

Dataset	**Features**
Taxi trajectory	Vehicle id
	Record time
	Latitude
	Longitude
Mobility	Number of people arriving
	Number of people leaving
Roads	Id
	Travel length
POI's	Id
	POI name
	Latitude
	Longitude
Meteorology	Temperature
	Weather

6.2 Comparison Methods

According to characteristics of our datasets, this method is compared with two typical machine learning algorithms (Gaussian Naive Bayes, Logistic Regression)

1) Gaussian Naive Bayes (GaussianNB):

In many practical applications, the naive Bayes classifiers[7] can be trained very efficiently in a supervised learning setting .We compare with the Gauss Naive Bayes, which is shown in equation6.

$$p(x = v|c) = 1/\sqrt{2\pi\sigma^2} * e^{-(v-\mu c)2/\,2\sigma 2\,c} \quad \text{--(6)}$$

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

2) Logistic Regression (LogisticR):

The Logistic regression [3] utilize a logistic function to measure the relationship between the categorical dependent variable and one or more independent variables. The logistic regression prediction model can be written as the below Equation. In this Equation7, the function softmax is a standard logistic function.

$$P(Y = i|x, W^{(2)}, b^{(2)}) = \text{softmax}_i(W^{(2)}x + b^{(2)}) = e^{W(2)ix+b(2)\,i} / \sum e^{W(2)jx+b(2)j} \quad \text{--(7)}$$

6.3 Results

Evaluation on Features:

In this model, first compare and analyze the prediction performance of the single feature dataset and the fusion of multiple feature datasets. These different feature datasets are used to train the multi-layer PNN model. 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 performance of the fusion of F_T, F_{Mo}, F_P, F_{RN}, F_{Me} is best. The predication accuracy of the model is 75% utilizing the fusion of F_T, F_{Mo}, F_P, F_{RN} and 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

Features	Accuracy of the prediction
F_T	70%
F_{Mo}	70%
F_T+F_{Mo}	55%
F_P+F_{RN}+ F_{Me}	50%
F_T+ F_{Mo}+ F_P+F_{RN}+ F_{Me}	75%

Overall Results:
Using the five kinds of feature datasets (F_T, F_{Mo}, F_P, F_{RN}, F_{Me}), we train the proposed multi-layer PNN and compare with the other two mentioned methods. The performance of these algorithms is shown in Figure5. The prediction accuracy of the multi-layer PNN is 75%.
Using WEKA, the prediction accuracy for Naïve Bayes and Logistic regression is only 65%. So it is clear that this model is more accurate than other algorithms. The results demonstrate that this method has the advantage to infer the transportation carbon emission over the other two methods.

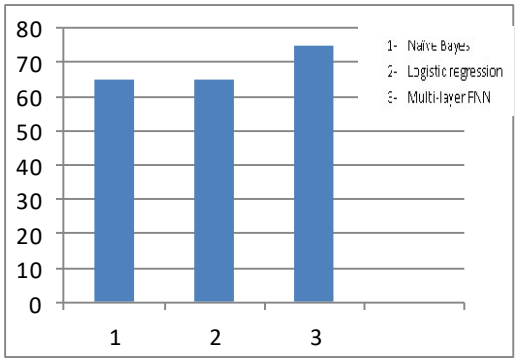

Fig 5: Overall results of different methods

VII. CONCLUSION

Using the five real world datasets, future transportation carbon emission is predicted. Calculated the amount of transportation carbon emission based on relation between five 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. We propose a multi-layerPNN model to find the accuracy for transportation carbon emission. The results demonstrate that multi-layer PNN method is better to the machine learning algorithms (Gaussian Naïve Bayes, Logistic Regression).

REFERENCES

[1] Lu, Kaoru Ota, Mianxiong Dong, CheYu, and Hai Jin,"Predicting Transportation carbon emission with urban big data," vol. 2, issue. 4.

[2] U. DOT, "Transportations role in reducing us greenhouse gas emissions," US Department of Transportation, Washington, DC, 2010.

[3] D.W. Hosmer Jr and S. Lemeshow, Applied logistic regression. John Wiley & Sons, 2004.

[4] I. P. on Climate Change, 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Intergovernmental Panel on Climate Change, 2006.

[5] J. Yang, J. Zhou, Z. Lv, W. Wei, and H. Song, "A real-time monitoring system of industry carbon monoxide based on wireless sensor networks," Sensors, vol. 15, no. 11, pp. 29 535–29 546, 2015.

[6] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: when urban air quality inference meets big data," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013, pp. 1436–1444.

[7] H. Zhang, "The optimality of naive Bayes," AA, vol. 1, no. 2, p. 3,2004.