Non-Linear Models of Light-Duty Vehicle Emission Factors Considering Pavement Roughness

Qing Li, Fengxiang Qiao* and Lei Yu

1Innovative Transportation Research Institute, Texas Southern University, USA

Abstract

Emission factors are very important measures for developing an emission inventory, making decisions, designing control strategies, mitigating climate change, and even improving public health, in terms of respiratory system diseases. The emission factors could be either measured from field tests or estimated by an emission model. Existing models seldom consider the impacts of some special factors such as pavement roughness. As the impacts of the pavement roughness on emissions are very complicated, a linear model or physical model may not depict the mappings from affecting factors to resulted emission factors. In this paper, two non-linear models, including K-Nearest Neighbor (K-NN) and Neural Network (NN) were built to estimate vehicle emission factors using roughness involved input data. A best fitted model was identified to illustrate the emission pattern along a wide range of pavement roughness. Multiple field tests were conducted in five regions of the State of Texas, United States, with a total of 1,609 km test length. One dedicated test vehicle was employed throughout the test. Pavement roughness was tested using a smartphone based application. All tested data were separated into four groups, each representing a different range of roughness, while the modeling was conducted within each group. The predictive performance of each model was evaluated by (1) correlation coefficient; (2) relative errors; and (3) two tailed unequal variance t-test. Results suggest that, K-NN can be better than NN to model the emission factors for the Texas highway system, and driving on a smoother and rougher pavement result in higher vehicle emissions.

Keywords: Modeling; Emission factors; Pavement roughness; K-Nearest neighbor; Neural network

Introduction

The U.S. Environmental Protection Agency (EPA) recommends building the emission inventory as parts of a State Implementation Plan (SIP) [1]. Emissions factors are important measures in the development of national, regional, state, and local emissions inventories for decision-making and control strategies. Users of emission factors include agencies in federal, state, and tribal levels, as well as consultants and industries [2,3]. Proper estimation of emission factors (EF) could also help in developing countermeasures in not only the environmental protections and congestion mitigation [4], but also the public health improvements [5,6]. The emission factors are also used to report to national greenhouse gas inventories under the United Nations Framework Convention on Climate Change (UNFCCC) [7].

EF could be measured directly from on-road measurement equipment and in-lab testing devices [8-13] or estimated using a suitable model such as the EPA models MOBILE 6.2 [14] and MOVES [3]. Field and in-lab tests are limited to the availabilities of equipment and testing environment/scenarios, while the model estimation may not consider all real conditions and might induce more or less errors [15]. Many studies found that vehicle emissions are very subject to many factors, such as driving behaviors [16], vehicle information [17], pavement materials [18], route’s slope conditions [19], traffic control system [20], and traffic situations, such as the situation at a work zone and a signalized intersection [21-23]. However, most of the emission models seldom incorporate the impacts of pavement roughness [24-26] into the independent variables. It is hypothesized that the vehicle emissions are nonlinearly correlated to pavement roughness.

The objective of this paper is to identify a best fitted nonlinear model to illustrate the impacts of pavement roughness on vehicle emission pattern, based on real world measurements by a test vehicle in five regions of the State of Texas.
### Table 1: Classification of pavement roughness based on Texas emission measurement (source: Li et al., 2016d).

| Range       | Center | Avg. | Std  | Evaluation |
|-------------|--------|------|------|------------|
| (0.00-1.99) | 1.36   | 0.051| 0.055| High       |
| (1.99-3.21) | 2.54   | 0.032| 0.017| Low        |
| (3.21-6.00) | 4.07   | 0.030| 0.016| Low        |
| > 6.00      | 7.07   | 0.039| 0.014| High       |

Table 1 shows higher vehicle emission factors were observed on the smoother and rougher pavement denoted by category A and D. This implies that pavement roughness is also one of determinants in air emissions. The roughness involvement may improve the accuracy of the vehicle estimation.

The nonlinear mapping in Figure 1 should be implemented through a nonlinear model, the output of which could be emission factors of major air emissions, such as: carbon dioxide (CO₂) in g/mi, carbon monoxide (CO) in mg/mi, hydrocarbon (HC) in mg/mi, and nitrogen oxides (NOₓ) in mg/mi. In order to have a uniform comparison of multiple air emissions (m), this study adopted the normalized emission index proposed by Li et al. The Normalized Emission Factor (N) is calculated by using equation (1).

\[
N_{ij} = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{x_{ij} - \text{Min}_{j}(x_{ij})}{\text{Max}_{j}(x_{ij}) - \text{Min}_{j}(x_{ij})} \right)
\]

where:

- \(N_{ij}\) = The i-th normalized emission factor in the j-th air emission,
- \(x_{ij}\) = The i-th emission factor of the j-th air emission (g/mi or mg/mi),
- \(m\) = The number of studied air emissions, here is 4 for CO₂, CO, HC, and NOₓ,
- \(\text{Min}_{j}\) = The minimum emission factor of the j-th air emission (g/mi or mg/mi), and
- \(\text{Max}_{j}\) = The maximum emission factor of the j-th air emission (g/mi or mg/mi).

Any nonlinear models could be candidates for the required nonlinear mapping. In this paper, two typical nonlinear models were employed to estimate emission factors based on the data collected from field: (1) the K-nearest neighbors (K-NN) model, and (2) the Neural Network (NN) model. Both are machine learning based multidimensional in their respective featured spaces. The two models are also memory-based. They start with training observations, and assume that the response class of nearly observations is likely to be similar.

#### The K-NN model

The K-nearest neighbors (K-NN) algorithm is based on an assumption that class probabilities are locally approximately constant. However, for most neighborhoods, it is not constant. To bring out a feasible constant class probability, distance metric needs to be improved. There are many types of distance metrics, such as Mahalanobis distance, city block metric, Minkowski metric, cosine distance, and so on [28]. Euclidean distance is a commonly used, expressed by equation (2).

\[
d^2 = (x - y)' (x - y)
\]

where:

- \(d^2\) = Euclidean distance squared between vector \(x\) and \(y\),
- \(x = \{x_1, x_2, \ldots, x_m\}\), or \(x\),
- \(y = \{y_1, y_2, \ldots, y_m\}\), or \(y\).

Meanwhile, the number of nearest neighbors called \(k\) is essential for deliver a precise estimated result. A smaller \(k\) may result in higher variance, whereas larger \(k\) may lead to higher bias. The selection of \(k\) quite depends on the nature data. Therefore, cross-validation is often adopted to seek for a proper nearest neighbor size with the lowest Error Log (el) described by equation (3).

\[
el = \log 10\left(\sum_{j=1}^{k} \sqrt{(y_j - \hat{y}_j)^2}\right)
\]

where:

- \(k\) = The best number of neighbors,
- \(y_j\) = The measured output at the j-th nearest neighbor, and
- \(\hat{y}_j\) = The estimated output at the j-th nearest neighbor.

The estimated output \(\hat{y}_j\) is an average of \(k\) weighted nearest neighbors, described by equation (4).

\[
y_j = \frac{\sum_{i=1}^{k} w_{ij} y_{ij}}{k}
\]

where:

- \(w_{ij}\) = The weight of the j-th input nearest neighbor at the \(i\)-th measured output neighbor,
- \(y_{ij}\) = The measured output of the \(i\)-th input neighbor at the \(j\)-th measured output neighbor,
- \(k\) = The best number of neighbors.

As the K-NN model excuses based on its training dataset, any noise or irrelevant features become sensitive for the model results. Meanwhile, more frequent classes may dominate the modeled result.

#### The neural network model

A neural network (NN) is flexible for linear and nonlinear. For nonlinear relationship between dependent and independent variables, the neural network could provide precise estimated results. Commonly two models were executed within the network, including Multiplayer perception (MLP) and Radial Basis Function (RBF), while the RBF provides a linear combination of radial basis functions of the inputs and neuron parameters, MLP used a supervised learning technique called backpropagation for training, which allows predicting more complex relationships. MLP was chosen in this study. MLP maps sets of input data onto a set of appropriate output, which is consisted of multiple layers of nodes. Each layer is fully connected to the next one. A typical neural network structure is presented in Figure 2. [29]

In Figure 2, there are \(p\) dependent variables, which are input to one or more hidden layers with \(q\) neurons. In each neuron, there is a linear activation function, which maps the weighted input variables to the output of each neuron. The main activation function is a hyperbolic tangent function in equation (5), which can be replaced by a sigmoid function in equation (6).
\[ y(x_i) = \tanh(x_i) \]  
\[ y(X) = (1-\exp(-Xi))^{-1} \]

where:  
\[ y = \text{The output variable(s)}, \]
\[ x = \text{The input variables}. \]

The machine learning based model results could be improved by increasing training and scoring times, which will be reflected by its structure as well. For a MLP, there would be up to two hidden layers with multiple neurons. Cross-check would be required to obtain an optimal structure, including the number of layer and neurons.

### Performance measures

The predictive performance of each modeling stage was evaluated by (1) the correlation coefficient between the observed values and the modeled values, (2) the relative errors, which were calculated from the difference between the observed and modeled emission factors divided by the observed values, and (3) the two tailed unequal variance t-test. The null hypothesis is that the observed and modeled population means are the same but the two population variances may differ. The null hypothesis will be accepted if the p value is greater than 95%. Compared with paired t-test, the unequal variance t-test is able to quantify how far apart the two means of the two population are [30].

### Emission tests and data collection

On-broad vehicle emission tests were conducted along Texas highways in various regions, including El Paso, Austin, San Antonio, College Station, Houston, and other Southeast regions, from November 2014 to June 2015 during sunny days. The tested routes include high-speed freeways, rural highways, arterial roads, and local street, covering a wide range of speed limit roadway facilities with a wide range of pavement roughness.

A Portable Emission Measurement System (PEMS) was equipped inside a dedicated test vehicle to provides second-by-second emission rate. The test vehicle is 10 years old with 10,000 starting mileage. A Global Positioning System (GPS) was paced on the top of the test vehicle to collected real-time test vehicle’ geo-location information, including latitude, longitude, and altitude. Meanwhile, the engine’ dynamic operation information, such as IAT, MAP, speed, acceleration and rpm were also recorded through a set of sensor arrays that were also connected to and synchronized with the PEMS.

Meanwhile, a smart phone installed with roughness measurement application (app) was mounted to the front of the windshield inside the vehicle by a phone car rack. Before each test, a simple calibration procedure was conducted. Concretely, the phone position was adjusted as straight (vertically or horizontally) as possible in order to set the phone’s three dimensions (x, y and z) as close to zero as possible, which serves as a reference point for the roughness model in the app. The correlation of the calculated IRI towards laser beam measured IRI is 80% above [31]. The app provides real-time calculated International Roughness Index (IRI) for every 20-meter distance.

A total of 1,609 km (1,000 miles) highway routes were tested and about 210,800 emission rates for CO₂, CO, HC, and NOₓ were recorded. To synchronize the IRI data, the collected emission rate data were calculated and interpolated into emission factors for every 20-meter distance. It turns out that 19,099 valid data pairs (20 meters each) were prepared. Seventy percent of the data pairs were used to train the models, while the rest were evenly separated for testing and validation.

### Results and Discussion

#### Models structure identification

A total of 19,019 data pairs were prepared. Based on the four categories of pavement roughness in Table 1, the data pairs were divided into four datasets. Most of test pavement roughness fell into category A, 14,078. 3,585 data pairs were classified into category B. Only few data pairs met the category C and D with 1,304 and 52, respectively. These datasets were further randomly divided into three groups for training (70%), testing (15%), and validation (15%) in the modeling process.

**K-NN model**: Cross-validation was conducted to identify the number of k, the optimal number of nearest neighbors. Table 2 presents a list of k with the highest correlation coefficient R values in a validation stage for the four categories, which ranges from 3 to 5.

In the most cases, 4 nearest neighbors were chosen. Meanwhile, with the k-values in Table 2, the validated emission factors were highly correlated to the observed values. Besides, a two tailed t-test was used to examine the variance of the observed and modeled emission factors. The null hypothesis is that two samples are equal variance. When p is greater than 0.05, the null hypothesis is accepted. All p-values in Table 2 are greater than 0.05, which means the variance of the observed and models emission factors is equal.

**Neural network**: A cross-check was conducted in the validation stage to seek for an optimal structure, with which the modeled emission factors may be highly correlated to the observed ones. There are two steps in this check. The first step was to identify the number of neurons at the first hidden layer. The second was to identify the number of...
hidden layer and the neural number in the second layer (if possible). Figure 3 shows the check results. In Figure 3, a blue line tells that R values increase with the increase in neurons in one hidden layer. When 10 neurons were used, the R value would be over 0.97. Thus, 10 neurons were identified for the first layer. For the possible second hidden layer, a cross-check was continued. A red line in Figure 3 demonstrates that adding second layer does not improve the R values. By the contrary, it dropped when the neuron further increases to 7 and 8. In response to this, 1 hidden layer with 10 neurons was confirmed for the structure of the NN model.

Model testing and validation

Two models were executed three stages, including training, testing, and validation, based on the identified structure. Correlation coefficients were adopted to evaluate the level of curve fitting at the three stages. An overview of the correlation coefficients performed by the two models is listed on Table 3.

Table 3 illustrates that the R-values by the two models are mostly higher than 0.50, which indicates a good fit with the observed values. More specifically, the R values of CO2 are overall higher than other air emissions in the three modeling stages. Except the R value of 0.77 and 0.64 by NN in validation for category A and D, the R values are higher than 0.89. This implies that the two models can estimate CO2 emission more accurate than other air emissions. This could be attributed to their different emission patterns. Compared with CO2 emission pattern, other air emission patterns are more complex. The CO2 emission is proportional to the demand of power need for motion, which can be quantified for the four categories with different level of pavement roughness, whereas there may be insufficient data points in category D to provide a generalized picture of the CO emission pattern. Hence, the modeled CO emission factor in category D may not be reliable. Moreover, the distribution of CO emission factors in category A (Figure 4a) are apparently more dispersive than in category B and C (Figures 4b and 4c), up to 250 mg/mi. The emission pattern is the inverse of the emission pattern for HC. The unburnt fuel can be easily escaped from an exhaust pipe as HC at higher ambient temperature.

Few R values marked in red are observed in the CO and NOx and Normalized Index (N). In particular, the R values of CO presents lower correlative to the observed values. It is more likely that the emission pattern of CO is different from other studied air emissions here, which would be explored in next sub-section. Besides, the most lower R values in red were performed by NN model. Thus, in terms of curve fitting for these datasets, K-NN can estimate vehicle emissions more accurate than NN model.

Fitted regression line: To obtain an insight into the emission pattern, several typical fitted regression lines are plotted in Figure 4. Figure 4 shows that there are obviously more data points in category A, whereas there may be insufficient data points in category D to provide a generalized picture of the CO emission pattern. Hence, the modeled CO emission factor in category D may not be reliable. Moreover, the distribution of CO emission factors in category A (Figure 4a) are apparently more dispersive than in category B and C (Figures 4b and 4c), up to 250 mg/mi. The on contrary, most CO emission factors are within 40 and 30 mg/mi in category B and C.

Similarly, the NOx emission factors in category A (Figure 4e) distribute more dispersive than in category B (Figure 4f), and the emission level is clearly higher as well. Figures 4g and 4h provided a similar view in Figures 4a and 4b.

Emission factor: In this study, specific emission factors were quantified for the four categories with different level of pavement roughness, based on on-board emission tests. Based on the observed results, two models were developed. The comparison of the modeled values and the observed values is illustrated in Figure 5.

Figure 5 shows that the relationship between pavement roughness and emission factors is not linear, which is consistent with the previous study by Li et al. [6]. The smoother or the rough pavement may induce higher vehicle emissions. Besides, it seems that the developed k-NN slightly under estimate the emission factor across the four air emissions, particularly the estimation in category D. This could be explained by insufficient data points for such rough pavement during the on-road tests. Likewise, NN model also did not deliver a better estimate results for category D. Further, NN model slightly overestimates CO2 and NOx emissions, and underestimates HC emissions.

![Figure 3: Cross-validation result of correlation coefficient for CO2](image)

**Table 3:** Correlation coefficients in three modeling stages.

| category | Stage | CO2 | CO | HC | NOx | Normalized Index |
|----------|-------|-----|----|----|-----|------------------|
|          |       | K-NN | NN | K-NN | NN | K-NN | NN | K-NN | NN | K-NN | NN |
| A        | Training | 0.96 | 0.98 | 0.84 | 0.73 | 0.77 | 0.58 | 0.78 | 0.74 | 0.84 | 0.71 |
|          | Testing | 0.96 | 0.97 | 0.87 | 0.41 | 0.85 | 0.72 | 0.77 | 0.42 | 0.88 | 0.36 |
|          | Validation | 0.98 | 0.77 | 0.85 | 0.04 | 0.80 | 0.68 | 0.78 | 0.36 | 0.86 | 0.08 |
| B        | Training | 0.96 | 0.98 | 0.85 | 0.85 | 0.78 | 0.79 | 0.87 | 0.82 | 0.86 | 0.60 |
|          | Testing | 0.97 | 0.89 | 0.19 | 0.94 | 0.52 | 0.74 | 0.92 | 0.29 | 0.26 | 0.93 |
|          | Validation | 0.97 | 0.89 | 0.66 | 0.33 | 0.83 | 0.80 | 0.89 | 0.43 | 0.71 | 0.27 |
| C        | Training | 0.96 | 0.99 | 0.59 | 0.82 | 0.54 | 0.84 | 0.86 | 0.79 | 0.66 | 0.92 |
|          | Testing | 0.99 | 0.91 | 0.69 | 0.35 | 0.98 | 0.60 | 0.95 | 0.43 | 0.81 | 0.51 |
|          | Validation | 0.93 | 0.98 | 0.89 | 0.31 | 0.89 | 0.80 | 0.91 | 0.65 | 0.94 | 0.56 |
| D        | Training | 0.92 | 0.89 | 0.78 | 0.69 | 0.87 | 0.97 | 0.83 | 0.85 | 0.89 | 0.89 |
|          | Testing | 0.97 | 0.96 | 0.98 | -0.43 | 0.99 | 0.34 | 0.99 | 0.83 | 0.99 | 0.97 |
|          | Validation | 0.99 | 0.64 | 1.00 | 0.12 | 0.96 | 0.92 | 1.00 | 0.72 | 0.99 | 0.98 |
Figure 4: Fitted regression lines of CO in the four categories and NOx in category A and B.
Conclusion

In this research, the K-NN and NN models are employed to model the emission factors based on the 1,609 km on-board emission tests in the state of Texas in United States. The modeling was conducted separately based on the range of pavement roughness (categories A, B, C, and D). The input variables include vehicle operational and engine information. Results show that, the K-NN model poses more accurate on the estimate of emission factor than the NN model in the four categories of pavement roughness. Meanwhile, the nonlinear relationship between vehicle emissions and pavement roughness is further validated. Driving on a smoother and rougher pavement result in higher vehicle emissions, which is consistent with the previous study by Li et al. [27].

Acknowledgements

Supports for this research in part by the U.S. National Tier 1 University Transportation Center (UTC) TranLive #DTRT12GUTC17/KLK900-SB-003, and the U.S. National Science Foundation (NSF) CREST #1137732 are gratefully acknowledged.

References

1. Kuykendal W (2014) Emissions inventory guidance for implementation of ozone and particulate matter national ambient air quality standards (NAAQS) and regional haze regulations. DIANE Publishing, U.S. Environmental Protection Agency (US EPA).
2. US Environmental Protection Agency (US EPA) (2016a) Technology transfer network clearinghouse for inventories and emissions factors.
3. US Environmental Protection Agency (US EPA) (2016b) Moves (Motor Vehicle Emission Simulator) 2014.
4. Qiao F, Yu L, Li L (2007) Estimating impact of nonrecurring congestion on vehicle emissions. In: Transportation Research Board 86th Annual Meeting (No. 07-0427).
5. Caiazzo F, Ashok A, Waitz IA, Yim SH, Barrett SR (2013) Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005. Atmospheric Environment. 79: 198-208.
6. Li Q, Qiao F, Yu L (2016) Vehicle emission implications of drivers' smart advisory system for traffic operations in work zones. Journal of the Air and Waste Management Association 66(5): 446-455.
7. Intergovernmental Panel on Climate Change (IPCC) (2006) IPCC guidelines for national greenhouse gas inventories.
8. Qiao F, Yu L, Vojtisek-Lom M. (2005) On-road vehicle emission and activity data collection and evaluation in Houston, Texas. Transportation Research Record: Journal of the Transportation Research Board (1941): 60-71.
9. Frey HC, Kim K (2005) Operational evaluation of emissions and fuel use of B20 versus diesel fueled dump trucks.
10. Sjödin A, Jerksjö M (2008) Evaluation of European road transport emission models against on-road emission data as measured by optical remote sensing.
11. Franco V, Kousoulidou M, Muntean M, Ntziachristos L, Hausberger S, et al. (2013) Road vehicle emission factors development: A review. Atmospheric Environment. 70:84-97.
12. Kousoulidou M, Fontaras G, Ntziachristos L, Bonnel P, Samaras Z, et al. (2013) Use of portable emissions measurement system (PEMS) for the development and validation of passenger car emission factors. Atmospheric Environment. 64:329-38.
13. Li Q, Qiao F, Yu L (2016) Calibrating emission factors for highways considering pavement roughness information. Proceedings of the 2016 Air Quality Measurement Methods and Technology.15-7.
14. EPA U. (2003) User’s guide to Mobile 6. 1 and Mobile 6. 2. Environmental Protection Agency.
15. Qiao F, Li Q, Yu L (2016) Neural network modeling of in-vehicle noises with different roadway roughness. Bridging the East and West. 27:174.

16. Munni J, Qiao F, Li Q, Yu L (2015) Driving behavior and emission analysis at yellow interval with advanced warning message under foggy weather condition: A simulator test. In: Proceedings of the 56th Annual Transportation Research Forum in Atlanta, Georgia.

17. Li Q, Qiao F, Yu L (2016) Estimating vehicle idle emissions based on on-board diagnostic II data. Proceedings of the 2016 Air Quality Measurement Methods and Technology. 15-7.

18. Li Q, Qiao F, Yu L (2015a) Implications of wireless communication system for traffic operations on vehicle emissions. Proceedings of the 107th Air & Waste Management Association (AWMA), June 22-25, 2015. Raleigh, North Carolina, USA.

19. Mansour HH, Aidin M (2012) Developing pollutant emission models for diesel public vehicles in Tehran. International Conference on Ecological, Environmental and Biological Sciences (ICEEBS’2012) Jan. 7-8, 2012 Dubai.

20. Li Q, Qiao F, Yu L (2014) Impacts of vehicles to infrastructure communication technologies on vehicle emissions. Proceedings of the International Conference on Environmental Science and Technology, American Academy of Sciences. 1: 9-13.

21. Li Q, Qiao F, Wang X, Yu L (2013) Impacts of P2V wireless communication on safety and environment in work zones through driving simulator tests (paper # 26-179). Proceedings of the 26th Annual Conference of the International Chinese Transportation Professionals Association (ICTPA). Tampa. 24-26.

22. Qiao F, Li Q, Yu L (2014) Testing impacts of work zone X2V communication system on safety and air quality in driving simulator. Publication in the proceedings of the 21st ITS World Congress. Detroit. 7-11.

23. Li Q, Qiao F (2014) How drivers’ smart advisory system improves driving performance? A simulator imitation of wireless warning on traffic signal under sun glare. LAMBERT Academic Publishing.

24. Chatti K, Zaabar I (2012) Estimating the effects of pavement condition on vehicle operating costs. Transportation Research Board.

25. Li Q (2016) Impact of freeway weaving segment design on environment and public health. Ph.D. Dissertation. Environmental Toxicology. Texas Southern University.

26. Li Q, Qiao F, Yu L (2015) Will vehicle and roadside communications reduce emitted air pollution? International Journal of Science and Technology. 1:17-23.

27. Li Q, Qiao F, Yu L (2016) Clustering pavement roughness based on the impacts on vehicle emissions and public health. J Ergonomics 6:1

28. Corry L (1997) Hermann Minkowski and the postulate of relativity. Archive for History of Exact Sciences 51: 273-314

29. Qiao F, Yang H, Lam W (2001) Intelligent simulation and prediction of traffic flow dispersion. International Journal of Transportation Research.

30. GraphPad (2010) The unequal variance (Welch) t-test

31. Roadroid (2013) Road conditioning monitoring using smart phones.