Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 5056284, 10 pages
https://doi.org/10.1155/2018/5056284

Research Article

Random Forests-Based Operational Status Perception Model in Extra-Long Highway Tunnels with Longitudinal Ventilation: A Case Study in China

Chao Qian,[1] **Jianxun Chen**,[2] **Yanbin Luo**,[2] **and Shuguang Li**[1]

[1]*School of Electronic & Control Engineering, Chang'an University, Xi'an 710064, China*
[2]*School of Highway, Chang'an University, Xi'an 710064, China*

Correspondence should be addressed to Jianxun Chen; chenjx1969@chd.edu.cn

Received 1 November 2017; Accepted 10 June 2018; Published 5 July 2018

Academic Editor: Juan C. Cano

An increasing number of extra-long highway tunnels have been built and put into operation around the world, but the quantified segmentation criteria for evaluating the in-tunnel operational status have not yet been enacted up till the present moment. Meanwhile, ventilation facilities could not satisfy the dynamic requirements of fresh air demand under fast spatial-temporal variation of traffic conditions and operating environment. In this study, the operational data collected from an extra-long highway tunnel were deeply analyzed using big data technology. By combining traffic flow and environmental monitoring data, a data-driven perception model based on the Random Forests was structured. The prediction results show that the proposed model provides better performances as compared to contrast models, indicating it had better ability to adapt to the dynamic changes of in-tunnel operational status while realizing accurate prediction. The designed intelligent control strategies of ventilation facilities and traffic operation applying for different operational status would provide a theoretical basis and data support for lifting the level of intelligent control as well as promoting energy saving and consumption reducing in extra-long highway tunnels.

1. Introduction

By the end of 2016, 815 extra-long highway tunnels with a total length of 3622.7 km were built in China [1]. Owing to the influence of traffic volume and fleet composition, vehicle emissions accumulate sequentially. These emissions are difficult to disperse, especially in the case of extra-long highway tunnels with high traffic loads and frequent traffic congestions. Tunnel ventilation has become the primary problem during operation periods.

For road tunnels, there exist several, very different approaches to ventilation concepts [2]. They have common objectives, opposite in nature: (a) the pollution levels within admissible margins and (b) the energy consumption for ventilation facilities to fulfill objective (a) should be minimal. Under some circumstances, it is difficult to meet both objectives concurrently by using simple ventilation control algorithms [3]. Thus many advanced control methodologies have been proposed in recent decades.

Appropriate and accurate ventilation control systems can not only decrease energy consumption and save operation cost but also provide drivers with a comfortable and safe driving environment. Standard linear feed-backward control was applied in early ventilation automatic control schemes such as PI or PID. However, these conventional control schemes reach their limits of applicability as soon as nonlinear effects become increasingly dominant. Funabashi et al. (1991) and Koyama et al. (1993), respectively, proposed ventilation control systems for longitudinal ventilation road tunnels with nonlinear programming and fuzzy control applications [4, 5]. Chen et al. (1998) designed a fuzzy logic control model for prediction of pollutant concentrations and adjustment of jet fans [6]. Chu et al. (2008) demonstrated a genetic algorithm in combination with fuzzy control to maintain an adequate level of the pollutants and minimize power consumption [7]. Bogdan et al. (2008) developed a model predictive and fuzzy control algorithm for a longitudinal ventilation system [8]. The predictive controller

estimates fresh air requirements (depending on traffic and weather conditions) and calculates the number of necessary jet fans, while the fuzzy controller compares measured and admissible levels of pollutants and adjusts a predicted number of jet fans to keep the pollutant levels within predefined boundaries. Euler-Rolle et al. (2017) applied a model based nonlinear dynamic feedforward control in the longitudinal tunnel ventilation to enhance standard feedback control and improve the closed-loop behavior [9]. However, all these contributions focused rather on the specific pollutants control than on the overall control and dynamic characteristics of in-tunnel operational status. Unchanging ventilation mode and unreasonable control strategy lead to enormous energy consumption and economic loss [10].

The in-tunnel operational status can be considered as a result generated by the combined action of four transportation elements, including the driver, vehicle, road, and environment. Li et al. (2015) focused on the diffusion properties of CO, NO, and $PM_{2.5}$ influenced by in-tunnel traffic force [11]. Yamada et al. (2016) and Martin et al. (2016) concentrated on the impact of in-tunnel tunnel environment (e.g., NO_2 level and particle number concentrations) on the driver and the passenger [12, 13]. Up till the present moment, the quantified segmentation criteria for evaluating the operational status in extra-long highway tunnels have not been enacted. Meanwhile, the analysis and mining of the in-tunnel operational status by deeply combining the real-time traffic flow and environmental information have also seldom been studied.

The decision tree is a classical classification algorithm, which is essentially a data recursive partitioning process based on a series of rules. Since the single decision tree has some drawbacks, such as low precision and overfitting, the ensemble learning method, which summates simple machine learning algorithms to produce better predictive performance than could be achieved by the most sophisticated solutions, has become popular in research in the field of machine learning. Practitioners created various solutions to improve a decision tree by replicating it many times and averaging results. For classification task, the ensemble can be used as a voting system, choosing the most frequent response class as an output for all its replications.

Aiming at finding the best way to replicate the trees in an ensemble, Breiman (1996) tested the effects of bootstrap sampling (sampling with replacement), which not only leaves out some noise but also creates more variation in the ensembles, improving the results. This technique is called "**b**ootstrap **agg**regat**ing**" and use the acronym **bagging** [14]. Noticing that results of an ensemble of trees improved when the trees differ significantly from each other, Breiman (2001) proposed a new ensemble model, Random Forests (RF), which add a layer of randomness to bagging [15, 16]. Random Forests change how the classification or regression trees are constructed by constructing each tree using a different bootstrap sample of the data, which turns out to perform very well compared with many other classifiers, including discriminant analysis, support vector machines, and neural networks,and is robust against overfitting [17].

The main goal of this work was to fuse the in-tunnel traffic flow data (such as fleet segmentation and traffic volume) and ambient air data (such as the concentrations of toxic gas and particular matter and air velocity) based on big data technology and to build a Random Forests-based perception model realizing accurate prediction of the in-tunnel operational status.

2. Material and Methods

2.1. Operational Monitoring Data. The Xi'an–Hanzhong Expressway (Xihan Expressway) is one of the most critical sections of the G5 Beijing–Kunming Expressway (a part of the China National Expressway Network, commonly known as the Jingkun Expressway), which connects north and southwest China, in Shaanxi province. A critical controlling project in the Xihan Expressway, the Qin Mountains tunnel group (Figure 1), comprises three extra-long highway tunnels, No. 1 tunnel, No. 2 tunnel, and No. 3 tunnel, passing through the Qin Mountains. The mountains are the most important geographical entities that divide northern and southern China.

No. 1 tunnel is a twin-bore tunnel with unidirectional traffic in each bore. The tunnel comprises southbound (SB) and northbound (NB) tunnels, with each direction having two lanes for motor vehicles. Figure 2 depicts the overall structure of the SB tunnel. In total, 11 lay-bys (emergency parking bays), numbered from ESA-1 to ESA-11, have been built along the length of the tunnel. The ventilation mode is longitudinal and is powered by 30 jet fans; an inclined shaft is reserved, and the air supply and exhaust system with additional axial fans had not yet been equipped. Since it was constructed and opened to traffic in 2007, the traffic has consistently increased. Among all vehicles, heavy-good vehicles (HGV) have shown the most notable increase.

In this study, four lay-bys (ESA-1, ESA-4, ESA-8, and ESA-11) were selected as the monitoring or data-collection sites in the driving direction. A real-time monitoring experiment for the operational environment was performed from Nov 27, 2016, to Dec 3, 2016. Through the experiment, raw monitoring data of the operational status were obtained. The details of data and monitoring instruments are listed in Table 1.

Raw monitoring data were preprocessed at statistical or resampling intervals of 15 min. That is, traffic flow data were converted to cumulative values every 15 min. The other monitoring data were calculated as average values for each 15 min interval. Finally, the statistical dataset of the operational environment was obtained.

The proportions of passenger cars (PC), light-duty vehicles (LDV), and HGV were 29.46%, 3.21%, and 67.32%, respectively. LDV had the lowest proportion, which smoothly changed; hence, its impact on the in-tunnel operational status could be ignored. The Pearson correlation coefficients indicated that PC had weak positive correlations with CO and NO_2. Consequently, the impact of PC on the in-tunnel operational status could also be ignored. Finally, only HGV was retained in the traffic flow data. Profiles of pollutant concentration in the driving direction exhibited a triangular distribution characteristic, increasing consistently from the tunnel entrance to the tunnel exit; this characteristic is consistent with the conventional wisdom of longitudinal ventilation systems. In conclusion, five types of data collected

Table 1: Monitoring instruments and data.

Data category	Data name	Instrument and model	Unit	Statistical interval
Ambient air	CO	ThermoFisher Model 48i CO Analyzer	ppm	15-min average value
	NO_2	ThermoFisher Model 42i NO_x Analyzer	ppm	
	Air velocity	AZ Instrument 9871	m/s	
	$PM_{2.5}$	ThermoFisher Model 5030i Particulate Monitor	mg/m^3	
Traffic flow	PC	Laser vehicle detector	vehicle/15 min	15-min accumulative value
	LDV		vehicle/15 min	
	HGV		vehicle/15 min	

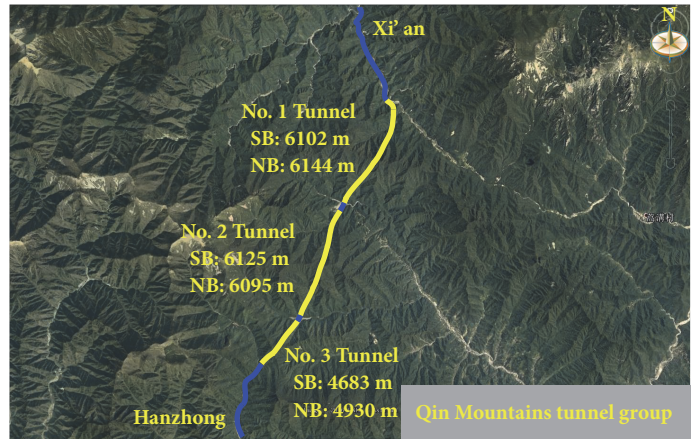

Figure 1: Qin Mountains tunnel group.

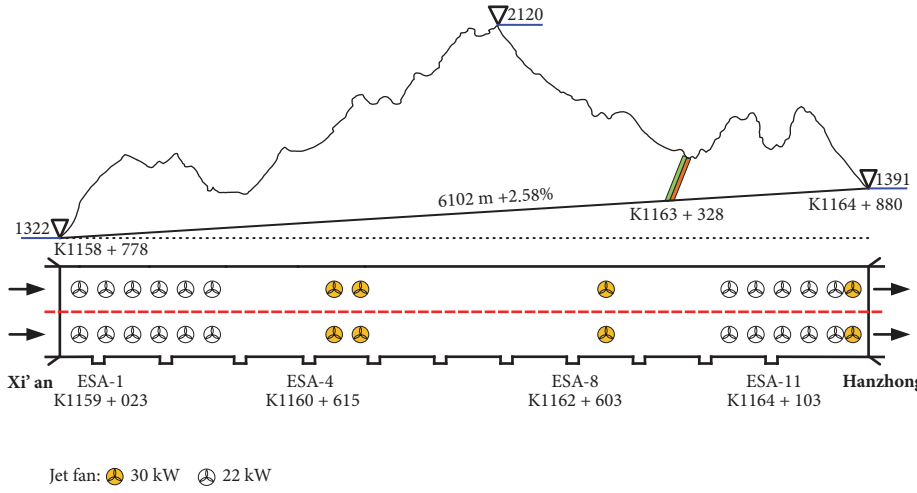

Figure 2: Overall structure of Qin Mountains No. 1 tunnel in the southbound direction.

from ESA-11, the monitoring site with the highest degree of pollution, were selected as the sample dataset; these data were the CO, NO_2, air velocity, $PM_{2.5}$, and HGV data.

2.2. Clustering Method. A five-dimensional space was obtained from the sample dataset. Clustering analysis for the operational status is the task of grouping the sample dataset in such a way that status data in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other clusters. In centroid-based clustering, the task can be summarized as finding the cluster centers and assigning the sample data to the nearest cluster center such that the squared distances from the cluster are minimized and thus obtaining a classification method for multiclass operational statuses.

Fuzzy C-Means (FCM) clustering is a fuzzy clustering algorithm based on an objective function; this algorithm was developed by Dunn [18] and improved by Bezdek [19]. Given

its advantages in big data applications, FCM clustering was chosen in this study. Consider that the ith sample data $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ denote a five-dimensional monitoring result, namely, the values of CO, NO_2, air velocity, $PM_{2.5}$, and HGV. The sample dataset containing N measured values is denoted by X. Then X can be expressed by a $N \times 5$ matrix, as shown in the following:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & x_{N4} & x_{N5} \end{bmatrix} \quad (1)$$

The FCM aims to minimize the following objective function:

$$\sum_{v=1}^{k} \sum_{i=1}^{N} u_{iv}^2 \left\| x_i - m_v \right\|^2 = \sum_{v=1}^{k} \sum_{i=1}^{N} u_{iv}^2 \sum_{f=1}^{5} \left(x_{if} - m_{vf} \right)^2 \quad (2)$$

where k is a preset number of operational status, i.e., cluster numbers; v is the sequence number of a cluster; m_v is the center of the cluster v; u_{iv}^2 stands for the unknown membership of sample x_i in cluster v with a membership exponent 2 to determine the level of cluster fuzziness; $\|x_i - m_v\|^2$ denotes the squared Euclidean distance between x_i and m_v; f is the sequence number of five-dimensional space; and m_{v1}, m_{v2}, m_{v3}, m_{v4}, and m_{v5} represent the values of cluster center m_v corresponding to CO, NO_2, air velocity, $PM_{2.5}$, and HGV, respectively. Cluster center m_{vf} can be calculated by the following equation:

$$m_{vf} = \frac{\sum_{i=1}^{N} u_{iv}^2 x_{if}}{\sum_{i=1}^{N} u_{iv}^2} \quad (3)$$

Kaufman and Rousseeuw (2008) proposed a new fuzzy clustering algorithm FANNY based on FCM [20]. The FANNY algorithm has some definite advantages over FCM: lower sensitivity to outliers or otherwise erroneous data and better recognition of nonspherical clusters. In the FANNY algorithm, the following equation is derived from (2):

$$\min \sum_{v=1}^{k} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} u_{iv}^2 u_{jv}^2 d\left(x_i, x_j\right)}{2 \sum_{j=1}^{N} u_{jv}^2} \quad (4)$$

where $d(x_i, x_j)$ represents the given distances (or dissimilarities) between samples x_i and x_j; Euclidean distance is in common use. Each pair is encountered twice because $d(x_j, x_i)$ also occurs, and the factor 2 in the denominator compensates for this duplicity. The membership function is subject to the following constraints:

$$\begin{aligned} u_{iv} &\geq 0, \quad i = 1, \cdots, N; \\ \sum_{v=1}^{k} u_{iv} &= 1, \quad i = 1, \cdots, N. \end{aligned} \quad (5)$$

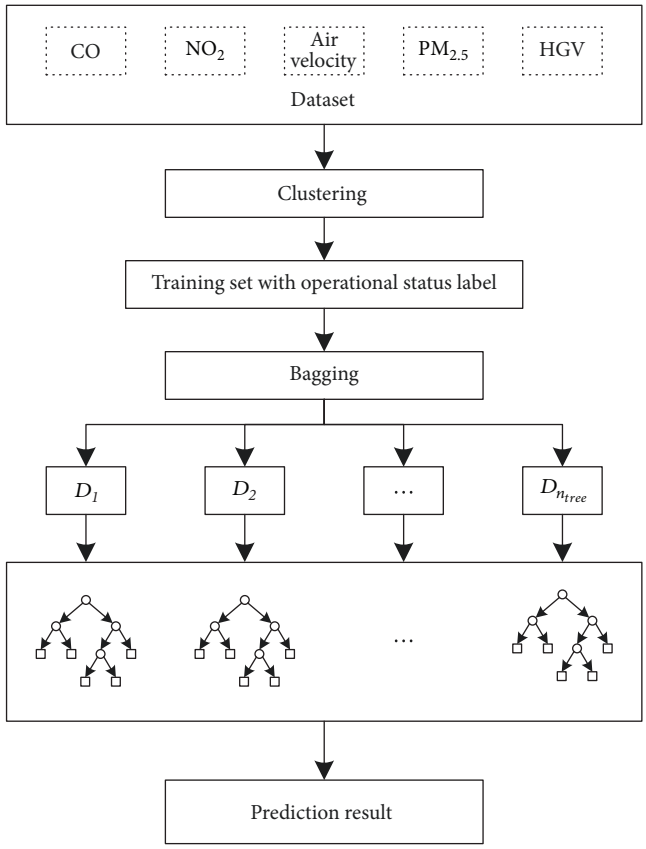

Figure 3: Diagram of Random Forests combining with clustering analysis.

The optimization problem is solved as shown in (4) to calculate and obtain the membership coefficients of all samples in every cluster $u_{iv} (1 \leq i \leq N, 1 \leq v \leq k)$ and each cluster center m_v. Thus, each sample is assigned to the cluster in which it has the largest membership, and the fuzzy clustering is completed.

2.3. Perception Model

Definition 1 (the perception of in-tunnel operational status). Given a training set $T = \{(x_1, y_1), \cdots, (x_N, y_N)\} \in (X^5 \times Y)^N$, $x_i \in X^5$ is the ith sample in the training set and it includes the values of CO, NO_2, air velocity, $PM_{2.5}$, and HGV; $y_i \in Y = \{c_1, c_2, c_3, c_4\}$ corresponds to one of the four operational statuses of the ith sample—lightly polluted, moderately polluted, heavily polluted, and severely polluted; and $i = 1, \cdots, N$ is the serial number of the training set. According to algorithmic modeling [21], the target is to find a function $f(x) : X^5 \longrightarrow Y$—an algorithm that operates on X^5 to predict the responses of in-tunnel operational status Y.

The ensemble of the Random Forests combining with clustering analysis is shown in Figure 3, in which clustering results of operational status are taken as inputs of Random Forests-based perception model. For perception model, first of all, bootstrap samples of size n_{tree} with replacement from the training set are taken, and a new series of training subsets are formed by the bagging technique. Then, randomly

Table 2: Confusion matrix for in-tunnel operational status.

Actual status	Perception status				Total
	Lightly polluted (c_1)	*Moderately polluted* (c_2)	*Heavily polluted* (c_3)	*Severely polluted* (c_4)	
Lightly polluted (c_1)	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$N_{1,.}$
Moderately polluted (c_2)	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$N_{2,.}$
Heavily polluted (c_3)	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	$n_{3,4}$	$N_{3,.}$
Severely polluted (c_4)	$n_{4,1}$	$n_{4,2}$	$n_{4,3}$	$n_{4,4}$	$N_{4,.}$
Total	$N_{.,1}$	$N_{.,2}$	$N_{.,3}$	$N_{.,4}$	N

select partial features in training subset for finding the best split variable whenever splitting the sample in a tree and create a complete tree using the bootstrapped examples. Next, compute the performance of each tree using examples that were not chosen in the bootstrap phase (out-of-bag data). Finally, calculate a vote on new cases when completing all the trees in the ensemble. Declare for each of them the winning class as a prediction.

2.4. Modeling Approach. There are the following two crucial parameters in Random Forests modeling, namely, n_{tree} and m_{try}:

(1) n_{tree}—the number of trees to grow;

(2) m_{try}—the number of variables randomly sampled as candidates at each split.

Herein, n_{tree} determines the overall scale of the whole Random Forests, and m_{try} defines the structure of a single decision tree. In other words, n_{tree} and m_{try} determine the construction of the Random Forests at macroscopic and microcosmic levels, respectively.

In R, the *randomForest* package provides an interface to the Breiman and Cutler's Fortran programs of Random Forests, and *randomForest()* function implements the algorithm for classification and regression [22]. The function prototype is as follows:

randomForest (*formula, data, mtry, ntree, na.action*)

in which *formula* describes the model to be fitted; *data* is a data frame containing the variables in the model; *mtry* is the number of variables randomly sampled; *ntree* is the number of decision trees; *na.action* specifies the action to be taken if NAs are found.

Since the bootstrap performs sampling with replacement from the training set, its probability to be chosen as the out-of-bag (OOB) sample is $(1 - 1/N)^N$. For large N, the number of OOB samples is expected to be a fraction $e^{-1} \approx 0.368$ of the training set. It means each decision tree is grown by using approximately $1 - e^{-1} \approx 63.2\%$ of the training samples, leaving $e^{-1} \approx 36.8\%$ as the OOB samples. Since the OOB part of the data has not been used in tree construction, it can be used to estimate the ensemble prediction performance in the following way.

Let D_b^{OOB} be the OOB part of the data for the bth tree. Then use the bth tree to predict D_b^{OOB}. Since each training sample x_i is in an OOB sample set, on the average approximately $e^{-1} \approx 36.8\%$ of the time the ensemble prediction $\hat{Y}^{OOB}(x_i)$ can be calculated by aggregating only its OOB predictions. Calculate an estimate of the error rate (ER) for classification by

$$ER \approx ER^{OOB} = n^{-1} \sum_{i=1}^{n} I\left(\hat{Y}^{OOB}(x_i) \neq Y_i\right) \tag{6}$$

where $I(\cdot)$ is the indicator function.

2.5. Evaluation Metric. A status set $Y = \{c_1, c_2, c_3, c_4\}$ is used to denote the four classes of the in-tunnel operational statuses—lightly polluted, moderately polluted, heavily polluted, and severely polluted. Then the confusion matrix (as shown in Table 2) is chosen to describe the classification performance.

In Table 2, $n_{i,j}$ denotes the number of actual statuses identified as c_j by the classification model. The confusion matrix reflects the distribution of status set Y, among which the jth column reflects the precision of c_j and ith row reflects the recall (also known as sensitivity) of c_i. Thus, for the particular operational status, e.g., c_j, the precision (P_{c_j}) and recall (R_{c_j}) are calculated by the following.

$$P_{c_j} = \frac{n_{j,j}}{N_{.,j}} \times 100\% \tag{7}$$

$$R_{c_j} = \frac{n_{j,j}}{N_{j,.}} \times 100\% \tag{8}$$

Besides, the other evaluation metric is the harmonic average of the precision and recall and is called F-measure (F). It is calculated as follows:

$$F_{c_j} = \frac{2P_{c_j}R_{c_j}}{P_{c_j} + R_{c_j}} \times 100\% \tag{9}$$

3. Results

3.1. Classification of Operational Status. Determining the optimal number of clusters is a fundamental issue in clustering analysis. In this study, this value was estimated by the optimum average silhouette width [23]. Suppose a data set is partitioned into k clusters, the silhouette width of sample x_i is then defined as

$$S_k(x_i) = \frac{B_k(x_i) - A_k(x_i)}{\max\left(A_k(x_i), B_k(x_i)\right)} \tag{10}$$

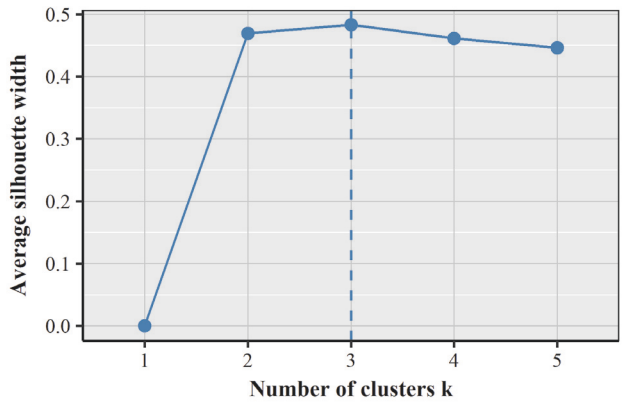

FIGURE 4: Optimal number of clusters; higher average silhouette widths are preferred.

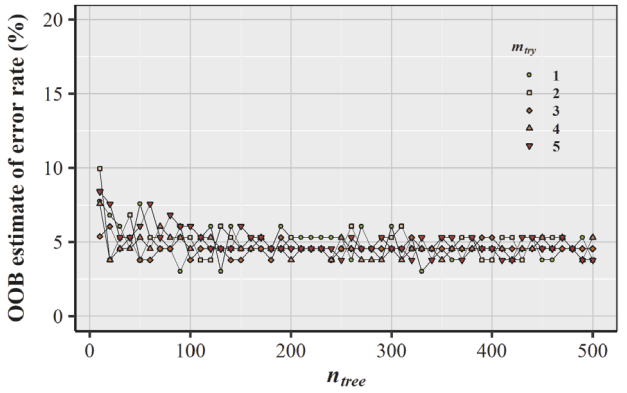

FIGURE 5: Influence of the combined parameters (n_{tree} and m_{try}) on OOB error.

where $A(x_i)$ is the average dissimilarity between x_i and all other samples in the cluster to which x_i belongs. Similarly, $B(x_i)$ is the minimum average dissimilarity between x_i and all other clusters to which x_i does not belong. The average silhouette method computes the average silhouette width (S_k) of all N samples for different values of k:

$$S_k = \frac{1}{N}\sum_{i=1}^{N} S_k(x_i), \quad k = 1, 2, 3 \cdots \tag{11}$$

The optimal number of clusters k is the one that maximizes the average silhouette width over a range of possible values for k.

The average silhouette widths with $k = 1, 2, 3, 4, 5$ for this study are shown in Figure 4. The silhouette plot shows that the k value of 3 corresponded to the maximum width, so the optimal number of in-tunnel operational status is 3 for the actual monitoring dataset. One of the four in-tunnel operational statuses did not appear in the experiment, so the next step is to verify which status was missing.

By applying FANNY algorithm to the preprocess data, three cluster centers are obtained using the following equation.

$$m_{vf} = \begin{bmatrix} 0.82 & 0.081 & 3.2 & 0.28 & 0 \\ 5 & 1.1 & 5.9 & 0.46 & 56 \\ 14 & 1.8 & 6.7 & 0.72 & 115 \end{bmatrix} \tag{12}$$

In (12), the five elements in each row represent the values of cluster centers in the following order: CO (ppm), NO_2 (ppm), air velocity (m/s), $PM_{2.5}$ (mg/m^3), and HGV (veh/15 min). The number of HGV is 0 in the first cluster, and the NO_2 concentrations in the second and third cluster exceed the PIARC standard (1 ppm) [24]. Thus, the three rows represent the cluster centers of lightly polluted, heavily polluted, and severely polluted statuses. The moderately polluted status did not appear when the traffic volume increased slightly.

3.2. Modeling of Status Perception

3.2.1. Optimal Combined Parameter. Before tuning the parameters in the Random Forests, the in-tunnel operational dataset is divided into a training set and a test set in the ratio 7:3. The former was used for parameter tuning and variable importance calculation. The latter was used for model evaluation. The minimum OOB ER principle is considered to be the reference to optimize the combination of parameters n_{tree} and m_{try}. R implementation code performed on a desktop PC running Windows 10, with a 3.6 GHz Intel i7 quad-core CPU and 16 GB RAM is shown as in Algorithm 1.

Assuming $n_{tree} = 10, 20, \cdots, 500$ and $m_{try} = 1, 2, 3, 4, 5$, 250 combinations of n_{tree} and m_{try} are run iteratively, and the relation between the combined parameters and OOB estimate of ER is obtained as shown in Figure 5.

Figure 5 shows that OOB ER was largely influenced by parameter n_{tree}; the error decreased with increasing n_{tree}, making the perception results more accurate. However, the time consumed for each iteration remained on the order of milliseconds, and, hence, the calculation time could be ignored. When $n_{tree} > 200$, OOB ERs tended to converge. The parameter m_{try} had less impact on the OOB ER. The results further validated that the Random Forests would be less likely to overfit, and the classification error would converge with an increasing number of decision trees. Consider the classification accuracy; an optimal combined parameter was identified as $n_{tree} = 500$ and $m_{try} = 1$, corresponding to 4.55% as an unbiased estimation of the OOB ER. The R code of status perception modeling is shown below.

```
Status.rf <- randomForest(Status ~ ., data = trainingset, mtry = 1, ntree = 500, na.action = na.omit)
print(Status.rf)
```

3.2.2. Importance Measurement of Variables. Another important feature of the Random Forests is the measurement of variable importance, which allows ranking variables regarding the importance and optimizing the variable subset, thus avoiding the problems created by dimensionality and reducing the computational complexity. There are two indexes to measure variable importance: mean decrease accuracy (MDA) and mean decrease Gini-index (MDG). The former is defined as the average decrease between the percentage of votes for the correct class in the untouched OOB data and

```
# division of training set and test set
set.seed(100)
ind <- sample(2, nrow(Mydataset), replace = TRUE, prob = c(0.7,0.3))
trainingset <- Mydataset[ind==1, ] # training set accounts for 70%
testset     <- Mydataset[ind==2, ] # test set accounts for 30%
# combined parameters in Random Forests
library(randomForest)
library(caret)
M <- ncol(trainingset)
ntree <- 10*c(1:50)
result <- data.frame()
# model training
set.seed(100)
for(m in 1:(M-1)){
  for(n in ntree){
    fit.rf <- randomForest(Status ~., data = trainingset, mtry = m, ntree = n, na.action = na.omit)
    OOB.ER <- 1-confusionMatrix(as.table(fit.rf$confusion[,c(-4)]))$overall["Accuracy"]
    result <- rbind(result, data.frame(m, n, OOB.ER))
  }
}
print(result)
```

Algorithm 1

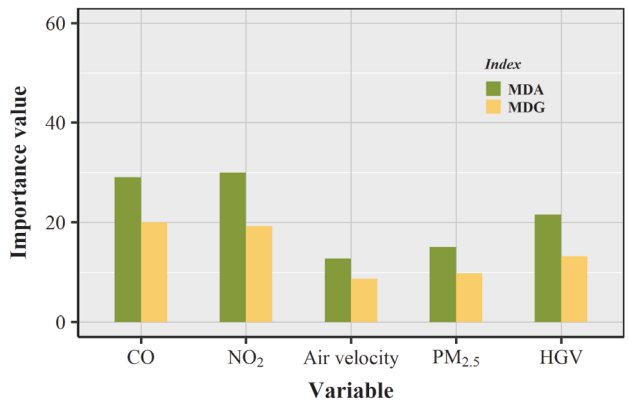

Figure 6: Comparison of importance of different variables.

the percentage of votes for the correct class in the variable permuted OOB data averaged over all trees. The latter is defined as the total decrease in the Gini-index from splitting on the variable averaged over all trees [22, 25]. The bigger the MDA and MDG, the more important is the variable. The optimal combined parameter in the training set was applied, and the importance indexes of each variable were calculated, as shown in Figure 6.

As seen from Figure 6, during the dynamic evolution process of in-tunnel operational status, the importance order of variables from largest to smallest is as follows: NO_2, CO, HGV, $PM_{2.5}$, and air velocity. NO_2 and CO, the two main types of gaseous pollutants, are the primary factors that affect the changes in the in-tunnel operational status.

3.3. Perception Results of Operational Status. In this study, the Naïve Bayes, Support Vector Machine (SVM), and Random Forests-based perception model were applied to predict operational status in the test set. Evaluation metrics for these three models are listed in Table 3.

Naïve Bayes classifier assumes that the value of a particular feature is independent of the value of any other feature, which is always invalid in practice. The naive design and apparently oversimplified assumption affect classification performance of Naïve Bayes. SVM can efficiently perform a nonlinear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. However, a significant practical question, the selection of the kernel function parameters, is still not entirely solved. Table 3 indicates that the precision, recall, and F-measure for the Random Forests-based model were better than those for the Naïve Bayes or SVM model. For further calculation, the average precision, recall, and F-measure in the Naïve Bayes model were 94.85%, 86.79%, and 90.19%, respectively. In the SVM-based model, the average precision, recall, and F-measure were 96.20%, 90.83%, and 93.24%, respectively. In contrast, the average precision, recall, and F-measure in the Random Forests-based model were 98.83%, 95.52%, and 97.07%, respectively. The results validated that the Random Forests-based perception model offers the best performance among the three models, indicating its better adaptability to the dynamic changes of the operational status in extra-long highway tunnels.

4. Discussion

4.1. Optimal Number of Clusters. Determining the number of clusters in a dataset, a quantity often labeled k, is a frequent problem in data clustering and is a distinct issue from the process of actually solving the clustering problem. The correct choice of k is often ambiguous, with interpretations depending on the shape and the scale of distribution of points in a dataset and the desired clustering resolution of the user.

Table 3: Comparison of evaluation metric for different perception models.

Operational status	Naïve Bayes			SVM			Random Forests		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Lightly polluted	100	80	88.89	100	85	91.89	100	90.00	94.74
Heavily polluted	88.71	98.21	93.22	92.44	98.21	95.24	96.49	100	98.21
Severely polluted	95.83	82.14	88.64	96.15	89.29	92.59	100	96.55	98.25

In this study, the silhouette method was chosen for assessing the natural number of operational statuses. Frankly, the determination of $k = 3$ was of a little subjectivity; $k = 2$ or $k = 3$, after all, was only slightly smaller than it. In consequence, long-term accumulation of in-tunnel operational monitoring data is crucial for the rational classification of operational status.

4.2. Management and Control Strategy for Ventilation and Traffic Flow. The perception model can be used to determine the real-time in-tunnel operational status by using a combination of pollutant concentration and traffic volume monitoring results. In the SB direction of Qin Mountains No. 1 tunnel, the percentages of the operational environment with heavily polluted and severely polluted statuses were 59% and 31%, respectively. The lightly polluted status contributed less than 10% of the operational environment. The box-plots presenting the distributions of CO, NO_2, air velocity, $PM_{2.5}$, and HGV under different statuses are shown in Figures 7(a)–7(e).

Although the moderately polluted status did not appear, the fluctuation range of CO (Figure 7(a)), NO_2 (Figure 7(b)), and $PM_{2.5}$ (Figure 7(d)) exhibited a tendency to intensify with the deterioration of the in-tunnel operational status, which was basically the same as the tendency for HGV. Figure 7(c) shows that there was a minimal number of HGV in regions with the lightly polluted status. Natural ventilation mode was used during that period, and, hence, the air velocities underwent a rather significant fluctuation influenced by the movement of vehicles (piston effect). For heavily and severely polluted statuses, all jet fans were turned on, and the air velocities were relatively stable (Figure 7(e)). Even so, the concentration of NO_2 still exceeded the PIARC standard.

The classification of in-tunnel operational statuses provides a scientific way to develop strategies for intelligent ventilation and traffic management and control. For lightly polluted status, consider switching off the fans and depending only on natural ventilation. For moderately polluted status, consider switching on the fans with a variable-frequency drive (VFD) to save energy consumption. For heavily polluted status, consider operating the jet fans at the fully open position and activating the axial fans in the inclined shaft in a timely manner. For severely polluted status, the in-tunnel air quality is terrible and the tunnel is filled with smog and smoke, threatening driving safety; therefore, all jet fans and axial fans should be fully operated. If the tunnel is operated under the severely polluted status for extended periods of time, temporary traffic control measures should be executed to ensure driving safety [26], for instance, limiting HGV powered by diesel engines passing through the tunnel or diverting them upstream of the tunnel.

4.3. Impact on Ecology and Environmental Management. The ecology and environmental impact of transportation are significant because transportation is a major consumer of energy and burns most of the world's petroleum. According to the annual report of Chinese Ministry of Environmental Protection, more than 246 million vehicles emitted 45.47 million tons of pollutants in China in 2014 [27]. Vehicle emissions have become one of the principal sources of air pollution and a significant cause of dust–haze and photochemical smog. Reducing transportation emissions will produce considerable positive effects on Earth's air quality, acid rain, smog, and climate change. Although stricter vehicle emission standards have been implemented, a vast number of old vehicles are still rolling down the road, exceeding the emission limit by several times. Consequently, effective measures should be made to accelerate the elimination of aging automobiles or retrofit them with approved pollutant control devices.

5. Conclusions

In this study, the operational monitoring data in an extra-long highway tunnel were analyzed in detail using big data technology. By combining monitoring results of CO, NO_2, air velocity, $PM_{2.5}$, and HGV, a data-driven model for in-tunnel operational status perception was structured. The major conclusions are as follows.

By applying the FANNY algorithm, the optimal number of clusters for obtaining the in-tunnel operational status was determined following the principle of maximum average silhouette width. Owing to the restriction of the total experimental duration, the clustering results did not contain all four operational statuses. Unfortunately, the moderately polluted status was not observed. The next step is to perform long-term monitoring of the in-tunnel operational environment and obtain massive data, thus realizing more scientific and reasonable classification of in-tunnel operational statuses.

A Random Forests-based perception model was built for determining in-tunnel operational status. Taking the perception accuracy into consideration primarily, an optimal combined parameter of the Random Forests was identified. Prediction results indicated that the proposed model was better than the contrast models and had the better adaptability to dynamic changes of operational status in extra-long highway tunnels, thus realizing accurate predictions.

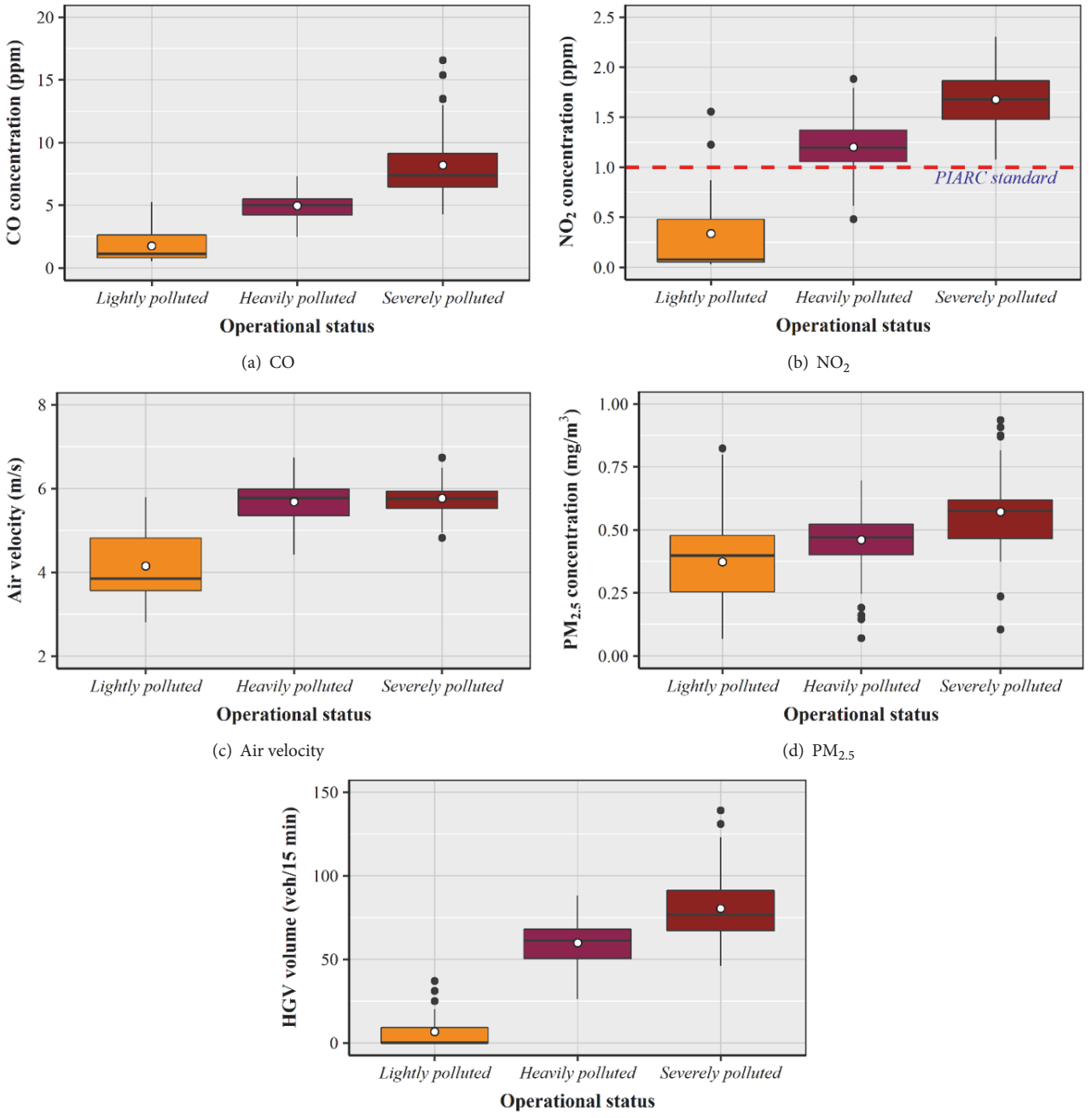

Figure 7: Distributions of particular variables under different statuses.

The distribution of individual variable under different operational statuses were analyzed. The management and control strategies for ventilation and traffic flow under lightly polluted, heavily polluted, and severely polluted statuses were discussed. These strategies could help improve the operation and management level of extra-long highway tunnels and provide a scientific method to realize energy saving and emission reduction.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant no. 51678063 and

Fundamental Research Funds for the Central Universities of Ministry of Education of China under Grants nos. 310832161006 and 310821173102. The authors thank the administration center of Qin Mountains tunnel group of Shaanxi provincial expressway construction group corporation Xihan branch for assistance with the experiments. They also thank graduate students Penglei Sun, Yuhui Zhai, Wei Li, Tongzhan Liu, and Xiang Ji from Chang'an University who participated in the experiments.

References

[1] Ministry of Transport of China, *Statistical Bulletin: Transportation Industry in 2016*, Beijing, China, 2017.

[2] PIARC, "Road tunnels: operational strategies for emergency ventilation," La Défense, 2011.

[3] S. Bogdan and B. Birgmajer, "Model Predictive Fuzzy Control of Longitudinal Ventilation System in a Road Tunnel," *Automatika – Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 47, pp. 39–48, 2006.

[4] M. Funabashi, I. Aoki, M. Yahiro, and H. Inoue, "A fuzzy model based control scheme and its application to a road tunnel ventilation system," in *Proceedings IECON '91 International Conference on Industrial Electrlnics, Control and Instrumentation*, Kobe, Japan, 1991.

[5] K. Toshihiro, Y. Tatsuro, W. Takahiro, S. Masanori, M. Miyako, and E. Hisashi, "Road Tunnel Ventilation Control Based on Nonlinear Programming and Fuzzy Control," *IEEJ Transactions on Industry Applications*, vol. 113, no. 2, pp. 160–168, 1993.

[6] P.-H. Chen, J.-H. Lai, and C.-T. Lin, "Application of fuzzy control to a road tunnel ventilation system," *Fuzzy Sets and Systems*, vol. 100, no. 1-3, pp. 9–28, 1998.

[7] B. Chu, D. Kim, D. Hong et al., "GA-based fuzzy controller design for tunnel ventilation systems," *Automation in Construction*, vol. 17, no. 2, pp. 130–136, 2008.

[8] S. Bogdan, B. Birgmajer, and Z. Kovačić, "Model predictive and fuzzy control of a road tunnel ventilation system," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 5, pp. 574–592, 2008.

[9] N. Euler-Rolle, M. Fuhrmann, M. Reinwald, and S. Jakubek, "Longitudinal tunnel ventilation control. Part 1: Modelling and dynamic feedforward control," *Control Engineering Practice*, vol. 63, pp. 91–103, 2017.

[10] C. Guo, M. Wang, L. Yang, Z. Sun, Y. Zhang, and J. Xu, "A review of energy consumption and saving in extra-long tunnel operation ventilation in China," *Renewable & Sustainable Energy Reviews*, vol. 53, pp. 1558–1569, 2016.

[11] Q. Li, C. Chen, Y. Deng et al., "Influence of traffic force on pollutant dispersion of CO, NO and particle matter (PM2.5) measured in an urban tunnel in Changsha, China," *Tunnelling and Underground Space Technology*, vol. 49, pp. 400–407, 2015.

[12] H. Yamada, R. Hayashi, and K. Tonokura, "Simultaneous measurements of on-road/in-vehicle nanoparticles and NOx while driving: Actual situations, passenger exposure and secondary formations," *Science of the Total Environment*, vol. 563-564, pp. 944–955, 2016.

[13] A. N. Martin, P. G. Boulter, D. Roddis et al., "In-vehicle nitrogen dioxide concentrations in road tunnels," *Atmospheric Environment*, vol. 144, pp. 234–248, 2016.

[14] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] L. Breiman and A. Cutler, "Random Forests," 2001, https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[17] A. Liaw and M. Wiener, "Classification and regression by randomforest," *The R Journal*, vol. 2, no. 3, pp. 18–22, 2002.

[18] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[19] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Springer, New York, NY, USA, 1981.

[20] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 2008.

[21] L. Breiman, "Statistical modeling: the two cultures," *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, vol. 16, no. 3, pp. 199–231, 2001.

[22] A. Liaw and M. Wiener, "Breiman and Cutler's Random Forests for Classification and Regression," 2015, https://CRAN.R-project.org/package=randomForest.

[23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[24] PIARC, "Road tunnels: vehicle emissions and air demand for ventilation," La Défense, 2012.

[25] L. Breiman, "Manual On Setting Up, Using, And Understanding Random Forests V3.1," 2002, https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.

[26] PIARC, "Integrated approach to road tunnel safety," La Défense, 2007.

[27] Ministry of Environmental Protection of China, *China Vehicle Emission Control Annual Report*, Beijing, China, 2015, Ministry of Environmental Protection of China, China Vehicle Emission Control Annual Report.