Journal of Cleaner Production xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Identification of high impact factors of air quality on a national scale using big data and machine learning techniques

Jun Ma [a, b], Yuexiong Ding [b], Jack C.P. Cheng [a], Feifeng Jiang [c], Yi Tan [d], Vincent J.L. Gan [a], Zhiwei Wan [e, *]

[a] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[b] Department of Research and Development, Big Bay Innovation Research and Development Limited, Hong Kong, China
[c] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong, China
[d] College of Civil Engineering, Shenzhen University, Shenzhen, China
[e] School of Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Article history:
Received 17 May 2019
Received in revised form
30 August 2019
Accepted 18 October 2019
Available online xxx

Handling editor: Tomas B. Ramos

Keywords:
Air quality index
Big data
GIS
National scale
Variable importance
XGBoost

ABSTRACT

To effectively control and prevent air pollution, it is necessary to study the influential factors of air quality. A number of previous studies have explored the relationships between air pollution and related factors. However, the methods currently used either cannot well address the multicollinearity problem or fail to explain the importance of the influential factors. Moreover, most of the existing literature limited their studied area in a city or a small region and studied factors in one aspect. There is a lack of studies that analyze the influential factors from the perspective of a country or take into consideration multiple variables. To fill the research gap, this paper proposes a multivariate analysis in the national scale to investigate the most important factors of air quality. In order to study as much influential factors as possible, 171 features ranging from environmental, demographical, economic, meteorological, and energy, were collected and analyzed. To tackle such a "big data" problem, a non-linear machine learning algorithm namely Extreme Gradient Boosting (XGBoost) is utilized to model the relationship and measure the variable importance. Geographical Information System (GIS) is employed to preprocess the diversified variables and visualize the results. Performance of XGBoost is compared with other models and its parameters are tuned using Bayesian Optimization. Experimental results of a case study in the U.S. show that our methodology framework can effectively uncover the important factors of air quality. Six kinds of factors are found to have the largest impact on air quality. Practical suggestions are also proposed from the six aspects to control and prevent air pollution.

1. Introduction

1.1. Background and literature review

In recent years, air pollution has become one of the major environmental issues due to urbanization and industrialization. According to the World Health Organization (WHO), 91% of the world's population lives in places where air quality exceeds the WHO guideline limits (out of 10 people worldwid, 2018). Serious air pollution has caused over one-third of deaths from stroke, lung cancer, and chronic respiratory disease, and one-quarter of deaths from ischemic heart disease (WHO | Air pollution and health: Summary and WHO, 2018). In regard to the huge damages of air pollution to human health, governments and research institutions have paid much attention to prevent and control air pollution. Governments have introduced relevant policies to control the emission of air pollutants from the sources. Scholars have been studying air quality from various aspects, such as upgrading the air quality prediction system (Li et al., 2017; Ma et al., 2019), analyzing the effectiveness of pollution prevention measurements (Wang et al., 2009; Beckett et al., 2000), and exploring the relationships between air pollution and related factors (Selden and Song, 1994; Haagenson, 1967).

Studying the influential factors of air pollution can help uncover the cause-effects behind, and therefore provide more specific suggestions to governments on policymaking. To numerically

* Corresponding author.
E-mail address: wanzhiwei0213@outlook.com (Z. Wan).

https://doi.org/10.1016/j.jclepro.2019.118955
0959-6526/

analyze the relationships, scholars have tried different kinds of methods. Among the existing methods, statistical models have been employed by most literature. For example, Vardoulakis and Kassomenos (2008) investigated the factors affecting PM10 levels in two European cities using principal component and regression analysis. Ngo et al. (2018) studied the impact of Chinese New Year and sandstorms on air quality in the U.S. using t-tests and ordinary least squares (OLS). Yang and Tang (2018) estimated a difference-in-differences (DID) model to assess the relationship between the fares of public transit and the air quality in Beijing. Monteiro et al. (2018) used cross-correlation to discuss the impact of the economic crisis on air quality in Portugal. Hao and Liu (2016) utilized spatial econometric models to investigate the influential factors of urban PM2.5 concentration in China. However, most of the statistical methods are based on the linear assumption, which is opposite to the non-linear characteristics of the real world. Therefore, their performance is limited and the results might be biased.

Recently, machine learning methods have been an alternative approach to address the non-linearity problem and improve model performance. For example, Pearce et al. (2011) investigated the influence of synoptic-scale meteorology on air quality using self-organizing maps (SOMs) and generalized additive modeling (GAM). Reich et al. (1999) identified the relationship between meteorological characteristics and air pollution sources using artificial neural networks (ANNs). Ali et al. (2014) employed support vector machine (SVM) for spatio-temporal air pollution analysis and examination of possible causes. Orun et al. (2018) adopted the Bayesian Networks to unveil hidden links of traffic-related air pollution (TRAP). However, although the aforementioned machine learning methods have generated satisfying modeling results on non-linear problems, they failed to non-linearly evaluate the variable importance. They are often referred as “black box” methods, and cannot help turn the numerical relationships into actionable suggestions (Casalicchio et al., 2018; Jun and Cheng, 2017). Therefore, a methodology framework that has the strong modeling power of machine learning algorithms, and is also capable of calculating the variable importance like statistical methods is required.

Furthermore, there are some common limitations in previous studies. Firstly, most existing literature limited their studied areas in a city or a small region. There is a lack of studies that analyze the influential factors of air pollution from the perspective of a country. This is not beneficial for the national government to propose nationwide countermeasures. Secondly, most of the previous studies only analyzed one kind of driving factors of air pollution, such as meteorology and transportation. However, reasons behind the air quality could be diversified. It is necessary to take into consideration various kinds of possible factors and conduct a comprehensive multivariate analysis. Thirdly, Geographic Information System (GIS), which is a system designed to capture, store, manipulate analyze, manage and present spatial or geographic information (Geographic information sy, 2018), has been widely used in assessing air pollution exposure and dispersion (Khan et al., 2018). However, most of the studies only utilized GIS to present or visualize the data. The ability of GIS in managing multi-source and multi-dimension data has been rarely adopted.

1.2. Research objective

To fill the research gap and overcome the limitations, this paper proposes a methodology framework that takes the advantages of non-linear machine learning methods and GIS to study the leading causes of air pollution in the U.S. The objectives of this study lie in two aspects:

- Explore and propose a methodology framework that could better support the analysis of the spatial influential factors on city/county/regional level factors on a national scale on the perspective of big data analysis.
- Identify the most influential factors among all the collected 171 counties level factors on the air quality of different counties in the U.S, and analyze the possible cause-effect behind.

The proposed method implemented Extreme Gradient Boosting Decision Tree (XGBoost) to model the non-linear relationships and calculating the feature importance. GIS was utilized in data fusion and post analysis. The second objective is in fact a verification of the proposed framework. 171 features ranging from environmental, demographical, economic, meteorological, and energy, were collected and analyzed. Based on the importance rank of the XGBoost algorithm, we were able to identify the most important factors on AQI (Air Quality Index) on a national scale. Actionable suggestions could be then given on policy making and city management.

2. Methodology

The methodology is proposed based on typical machine learning procedures. Machine learning (ML) is the scientific study of algorithms and statistical models that used to perform a specific task without using explicit instructions, but relying on the pattern and inference on the data sets instead (Ma and Cheng, 2016a). The inference in supervised machine learning is usually called the label. It is guiding the machine learning algorithms to learn the relationships between the features and the labels. The algorithms will build a mathematical model based on sample data, known as the training data, and then implemented the trained model on testing data to make predictions or decisions (Ma and Cheng, 2016b).

In this study, the machine learning methodology is utilized to model the relationships between various influential features and the AQI in the US counties (the labels). Considering the “no explicit instructions” idea in ML, the proposed methodology is expected to be applicable on other large regions, such as city-level or national-level spatial analysis. It consists of three parts, as shown in Fig. 1. The first part is data collection and preprocessing. Raw data are collected and preprocessed, and features are engineered and fused through Geographical Information System (GIS). Based on the preprocessed data, an Extreme Gradient Boosting (XGBoost) model is constructed and its parameters are fine-tuned using Bayesian Optimization (BO). Modeling performance of XGBoost is evaluated by comparing with other models. Feature importance then is measured. Based on the calculated feature rank, the most influential features are uncovered. Cause-effects behind these factors are analyzed afterwards.

2.1. Data fusion in GIS

After the raw data collection, there is an important step called data fusion in this kind of analysis. This is because when analyzing the influential features on air quality under a big data concept, it is necessary to collect and analyze data from various kinds of sources, such as temperature, city population/income, and power plant stations. To combine these kinds of data and form a unified dataset is not an easy task, since the database key for different data sets are different (Geographic information sy, 2018). For example, as shown in Fig. 2, we summarized the commonly seen data types in spatial analysis into three different categories, including the typical data frame, the point data, and the image data. We also picked an example in each category. They are city level personal income (PI),

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

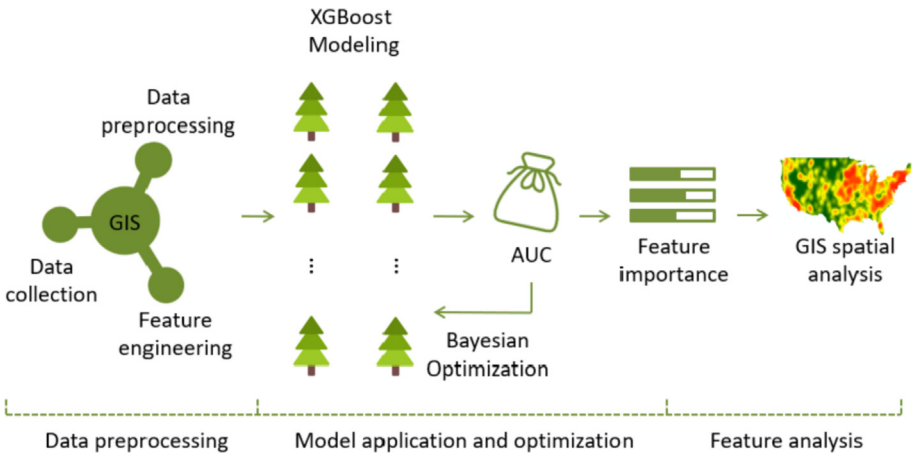

Fig. 1. The proposed methodology framework. It first integrates the labels and features in GIS, and then models the relationships using XGBoost while the algorithm parameters are tuned using Bayesian Optimization. The calculated important factors were later analyzed in GIS.

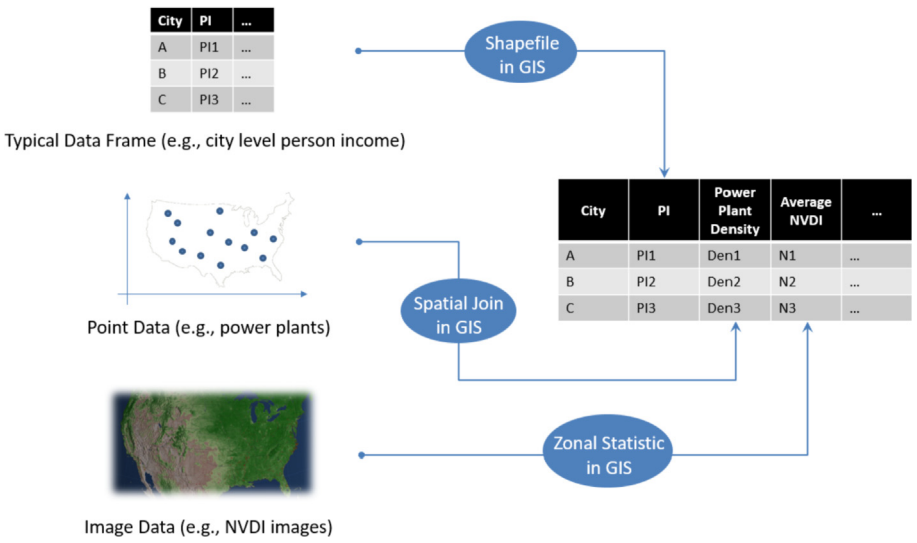

Fig. 2. Data fusion of three typical data types in GIS.

location of the power plants, and the Normalized Difference Vegetation Index (NDVI) images. Traditionally, it is very difficult to combine these three kinds of data into the features for different cities. You may have to develop a complicated calculation strategy to achieve this and it still can be inaccurate. However, by using the shapefiles and the geospatial tools in GIS, things can be very straightforward. As shown in Fig. 2, the city level PI data can be connected to the shapefile using the city names. The number of the power plants in each city can be calculated using the spatial join tool in GIS. The average NDVI value that reflect the vegetation level of a city can be calculated using the zonal statistic tool.

Besides the useful data fusion functionality, GIS also helps a lot in analyzing the spatial correlations for different features in the post-learning stage. More details on this will be introduced in the case study.

2.2. Extreme Gradient Boosting Decision Tree

After preprocessing, it can be seen from Fig. 1 that Extreme Gradient Boosting (XGBoost) is the key component in the framework to model the non-linear relationship between air pollution and related factors. In fact, its ability in classification and regression has been widely used and has achieved great performance in previous literature. For example, Zheng et al. (2017) adopted the XGBoost algorithm for feature importance evaluation for short-term load forecasting. Torlay et al. (2017) applied XGBoost to identify atypical language patterns and differentiate patients with epilepsy from healthy subjects. Zhang and Zhan (2017) utilized XGBoost for rock facies classification. In this study, the algorithm is also implemented to learn the relationships between the inputs and outputs, which are the county level factors and the high/low AQI respectively. Besides, this is one of the pioneering studies in air quality research that have utilized this algorithm in identifying the variable importance and analyzing the cause-effect.

Extreme Gradient Boosting Decision Tree (XGBoost) is an ensemble technique developed based on the Gradient Boosting proposed by Friedman (2002) but with some improvements and better performance (Chen and Guestrin, 2016). It builds a set of classification and regression trees (CARTs) as base learners in a parallel way and gives the result by summing up the score of each CART. The model therefore can be written as Equation (1).

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

$$\hat{\mathbf{Y}} = \sum_{m=1}^{M} f_m(\mathbf{X}) \tag{1}$$

where $\mathbf{X}$ and $\hat{\mathbf{Y}}$ are the inputs and outputs of the model, M represents the number of CARTs and f_m denotes each independent CART tree with leaf scores.

As mentioned above, compared with the original Gradient Boosting (GBDT) designed by Friedman (2002), some improvements are applied in XGBoost. One of them is the introduction of the regularized objective to the loss function (Nobre and Neves, 2019). Calculation of the regularized objective $\mathscr{L}^m$ to optimize for the mth iteration and the regularization term Ω is shown respectively in Equation (2) and Equation (3).

$$\mathscr{L}^m = \sum_{i=1}^{n} l(y_i, \hat{y}_i^m) + \sum_{j=1}^{m} \Omega\left(f_j\right) \tag{2}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{k=1}^{T} w_k^2 \tag{3}$$

Where n represents the number of samples, $\hat{y}_i^m$ represents the prediction of the sample i at iteration m, $l()$ represents the training loss function. T represents the number of nodes and w represents the weight of each node. γ and λ are two constants used to control the regularization degree.

Another improvement of XGBoost over GBDT is the utilization of additive learning strategy (Zhang et al., 2019). Instead of applying stochastic gradient descent method to complement the corresponding optimization procedure, XGBoost adds the best tree model $f_m(x_i)$ into the current classification model to give prediction result for the mth iteration. It can be expressed as Equation (4).

$$\hat{y}_i^m = \hat{y}_i^{m-1} + f_m(x_i) \tag{4}$$

By applying Equation (4) to the regularized objective, Equation (2) can be further formulated as follow:

$$\mathscr{L}^m = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{m-1} + f_m(x_i)\right) + \Omega(f_m) + \text{constant} \tag{5}$$

Furthermore, XGBoost adopts the Taylor Expansion second order to the objective function, which means Equation (5) can be further expanded to Equation (6).

$$\mathscr{L}^m = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{m-1}\right) + g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) \right] + \Omega(f_m) \tag{6}$$

where $g_i = \partial_{\hat{y}^{m-1}} l(y_i, \hat{y}^{m-1})$ and $h_i = \partial^2_{\hat{y}^{m-1}} l(y_i, \hat{y}^{m-1})$ are the first and second order derivatives on the loss function respectively.

2.3. Feature importance

In addition to the ability in modeling non-linear classification and regression problems, XGBoost is also capable of ranking the feature importance by averaging the feature importance in each tree. The importance of a features in a single XGBoost tree is calculated by the amount of information gain after splitting the tree using the feature. Its calculation is shown in the following equation.

$$\mathrm{IG}(\mathrm{T},\ \mathrm{F}) = \mathrm{H}(\mathrm{T}) - \mathrm{H}(\mathrm{T}|\mathrm{F}) = -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{F} p(F) * \sum_{i=1}^{J} -p(i|F)\log_2 p(i|F) \tag{7}$$

where H(T) and H (T|F) are the entropy of the parent node and the child notes of the split based on the F feature, p_i is the fraction of each labeled samples in one node.

To sum, this paper combines the non-linear machine learning method XGBoost and the Geographic Information System (GIS) technique to model the relationship between air quality and possible influential factors and figure out the most important factors. Taking the advantages of XGBoost and GIS, the proposed methodology framework can effectively model the non-linear relationship, calculate the feature importance, reduce the complexity of data preprocessing as well as generate persuasive results and analysis. To test the effectiveness of the methodology, a case study is performed in the following of the paper.

3. Case study

3.1. Data preparation

Due to data availability, the U.S. is selected as the study area for the case study. Four years (2012–2016) of air quality data in the U.S. counties are collected from the United States Environmental Protection Agency (EPA) as the study target. Since this paper aims at exploring and analyzing the influential factors of air quality, different aspects of possible influential factor data need to be collected. With the help of GIS, we gathered features from six aspects, including meteorology, energy, economics, demography, transportation and environment. The main data sources include United States Environmental Protection Agency (US EPA), United States Census Bureau (US Census), and United States Energy Information Administration (US EIA). Feature counts and examples of the prepared six categories are presented in Table 1. In total, we obtained 171 features.

3.2. Data preprocessing

Before applying the XGBoost on the collected dataset, the samples and features need to be pre-processed to better fit the model. The procedures in this experiment mainly include labeling, missing data imputation, and feature selection.

3.2.1. Labeling

4-year average AQI of each county is the label of the machine learning model in this study. Note that before calculating the average AQI for labeling, we filter out the days that have some extreme values, because they may be outliers caused by unusually events such as random wild fires, recording mistakes, etc. These could negatively affect the analysis, and therefore should be excluded. In this study, it is achieved by using the 3-sigma rule in statistics. That is when calculating the average AQI in each county, we excluded the days that have an AQI larger than $\mu_i + 3\sigma_i$, or lower than $\mu_i - 3\sigma_i$. Here μ_i and σ_i are the mean and standard deviation of the raw AQI in a county. Some of the counties did not have relevant records in the EPA data set, so they were excluded from this study. This results in 1014 counties in this experiment.

In addition, according to the international AQI standard (Air Quality Index (I) B, 2019), when AQI is smaller than or equal to

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

Table 1
Categories, examples and feature counts of the prepared 171 features.

Category	Feature Counts	Examples
Meteorology	15	Average Wind Speed; Average Temperature; Highest Daily Average Temperature; Precipitation; Snow Days;
Energy	9	Heating Degree Days; Cooling Degree Days; Number of Power Plants;
Economics	48	Personal Income; Farm Earnings; Construction Earnings; Forestry Earnings;
Demography	61	Population; Number of Workers 16 years and over; Number of Adults 16 years and over with Bachelor degree or higher
Transportation	27	Number of People with 3 or more vehicles available; Percentage of People taking public transportation to work;
Environment	11	Water Area; Percentage of Water Areas; NDVI; Elevation; Land Area; Total Area;
Total	171	

50, air quality is considered satisfactory and air pollution poses little or no risk; when AQI is larger than 50, air quality becomes worse and the potential of air pollution to affect public health becomes higher. Thus, to further benefit the modeling, label of the study is binarized into positive case and negative case. Counties of which the average AQI is smaller than or equal to 50 are marked as positive cases, while those larger than 50 are marked as negative cases. In total, there are 937 positive cases and 77 negative cases. The distribution of the counties in the US is presented in Fig. 3. Green represents the positive cases while red represents the negative cases. Grey means that the AQI data of the county is unavailable.

3.2.2. Missing value and feature normalization

Among the prepared 171 features, some of them contain different level of missing values. These should be handled before model implementation. For the features that have a missing rate higher than 50%, they have already been excluded out from the selected 171. While for those with less than 50% missing rates, the blanks are filled using the average value of the corresponding state.

Besides, to mitigate the influence from high dimension, and speed up the model training, z-score transformation is conducted to normalize the data. Calculation of z-score transformation is shown in Equation (8).

$$X_{transform} = \frac{X - \mu}{\sigma} \tag{8}$$

where μ represents the mean value, and σ represents the standard deviation.

3.2.3. Feature selection

Next, feature selection is performed. In the collected dataset, not all the features are numerically influential factors of air quality. Redundant features will contribute little or even nothing to the feature analysis but increase the complexity of the modeling. Therefore, they need to be excluded.

Commonly seen feature selection methods include filter methods, wrapper methods, and embedded methods (Ma and Cheng, 2016c). Filter methods will rank the features based on their relationship with the target label, and then select the top ones. Wrapper methods is more time consuming compared with filter methods. It will depend the model performance to decide which subset of features to keep. Embedded methods are those incorporated into the machine learning algorithms already.

After referring some literatures, this study implemented several representative methods in each category. The eventual model performance of each method is shown in Table 2. For filter methods, we compared correlation-based feature selection (CFS) and Principle Component Analysis based (PCA) feature selection (PFA). CFS

Table 2
Performance comparison of different feature selection methods with parameters tuned. The models are all trained and tested using the optimized XGBoost algorithm. Here 10-CV means 10-fold cross validation. AUC is a [0,1] range model measurement, and it will be introduced in later sections. The higher number means better models.

Category	Methods	Number of Selected Features	10-CV AUC
Filter	CFS	87	0.846
	PCA	58	0.834
Wrapper	RFE	121	0.875
Embedded	Elastic Net	75	0.841
N/A	N/A	171	0.853

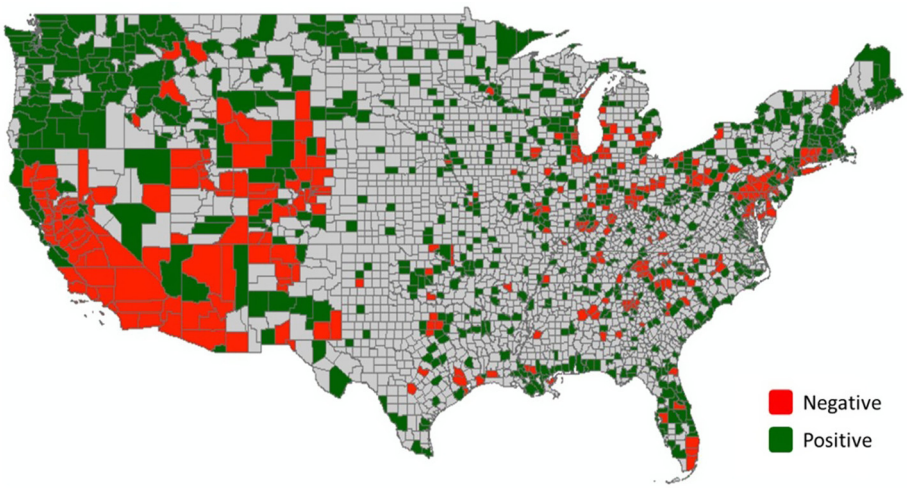

Fig. 3. The distribution of counties with positive and negative AQI values. Grey means the data is not available.

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

is developed based on a concept that "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other" (Hall and Smith, 1997; Cheng and Ma, 2015a). PFA tried to find uncorrelated representatives of the features sets (Lu et al., 2007). For wrapper methods, we picked the typical Recursive Feature Elimination (RFE) procedure. It recursive eliminates the lowest rank features based on model performance. For embedded methods, we compared the feature sets selected by Elastic Net. It uses L1 and L2 penalty to shrink the coefficients of the regression (Ma and Cheng, 2016c). Note that the parameters for different methods are all tuned for optimal performance of the relevant models.

It can be seen from Table 2 that the wrapper method produced the best model performance. Therefore, it is selected as the feature selection method in this study. Note that although the feature sets generated using the wrapper method may contain features that have high correlations in-between, the multicollinearity problem can be mitigated by the bootstrap process in XGBoost (Ma and Cheng, 2016a). Therefore, the feature importance calculated by XGBoost is more stable and closer to the real fact than the weights in traditional linear regressions.

As a result, the preprocessing eventually prepared a data set with 1014 samples and 121 potential influential features.

4. Modeling and results analysis

4.1. Model measurement and cross-validation

It is noticed that for the designed classification problem, the samples are imbalanced with 937 positive cases and 77 negative cases. Therefore, the traditional prediction accuracy cannot provide an objective measurement for the algorithm performance. To solve this problem, this study implements the Area Under the receiver operating characteristics Curve (AUC, or AUROC) as the measurement. As shown in Fig. 4, the receiver operating characteristics curve (ROC) is drawn by the true positive rate (TPR) and the false positive rate (FPR) (Ma and Cheng, 2017). Since the modeling outputs of the raw XGBoost model is numerical values, its prediction on positive or negative depends on whether the calculated value is smaller or larger than the threshold. So, when the threshold rises from 0 to 1, each threshold will have a different pair or TPR and FPR. Connecting all the pairs produces the ROC curve, and the area under the ROC curve becomes the AUC value. Thanks to the search-over process, AUC can be more objective on imbalanced data sets than commonly used measures like F1 score (Ma and Cheng, 2017).

In addition, since 1014 samples is not a very large number, to prevent the results from the overfitting problem, this study implemented 10-fold cross validation to test the algorithms and parameters. Unlike the typical training and testing partition, 10-fold cross validation will separate the dataset into ten parts without replacement. Then the algorithm will be trained and tested ten rounds. Each round will use nine of ten parts for training and the remaining one part for testing. Ten rounds make all the ten parts be tested once. This could significantly reduce the chance for overfitting (Cheng and Ma, 2015b). Averaging the performance of the ten rounds gives the final evaluation of the algorithm.

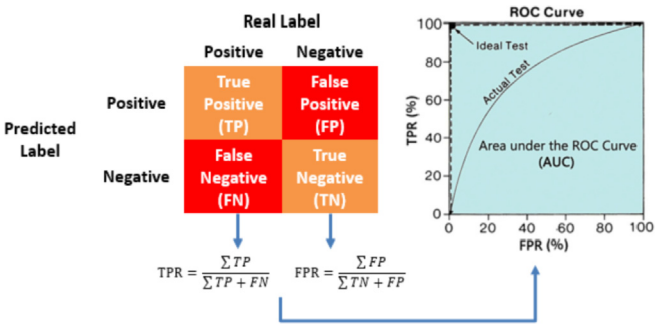

Fig. 4. Calculation of the AUC measurement.

4.2. Parameter tuning using Bayesian Optimization

Before inputting the cases into XGBoost and modeling the relationship between AQI and possible influential factors, the parameters of XGBoost need to be optimized to further improve the modeling performance. In XGBoost, there are six important parameters that could affect the model performance:

- N_T: number of boosted trees. This parameter represents the number of iterations XGBoost train on the data. Too small will lead to under fitting, while too large will cause overfitting.
- lr: learning rate. This parameter adjusts the size of the learning steps. Too small will lead to local optimum and slow calculation, while too large may miss the optimal value and not converge.
- sub_sample and colsample_bytree: these two parameters represent the proportion of training sample and variables used in training. They are like the bootstrap process in Random Forest, and could help control the randomness and mitigate multicollinearity.
- max_depth: the maximum depth a tree can grow. A larger value may help fit the data better, but also increase the risk for overfitting.
- min_child_weight: the minimum sum of instance weight (hessian) needed in a tree child. This value reflects when to stop split once your sample size in a node goes below a given threshold. It is used to prevent overfitting.

Tuning these six parameters all together will make the search space too large and not controllable. To address this problem, an efficient optimization procedure should be proposed. The objective of this step is to figure out the most optimal set of parameters for the problem in this paper. The most typical methods include grid search (GS) and random search (RS). These two methods are the most traditional and easy to implement, but still can cost much computation and usually cannot obtain an optimal result (Simon, 2013). To improve these, researchers later proposed population based methods such as genetic algorithms (GA), particle swarm optimization (PSO), and evolutionary strategies (ES) (Feurer et al., 2019). These methods usually consume less time than GS and RS in obtaining an optimal result. However, their performance are not state-of-art (Simon, 2013; Feurer et al., 2019).

In the era of deep learning and modern machine learning, scholars in computer science have discovered a higher performance framework for parameter optimization, and it is namely Bayesian Optimization (BO). It has been reported to achieve state-of-art in many research domains including image recognition, speech recognition, and neural language modeling (Feurer et al., 2019). However, no study ever before has explored the combination of XGBoost and BO in identifying the influential factors on air quality on a national scale.

The method has two key components. One is the probabilistic surrogate model, and the other is the acquisition function. The surrogate model is used to fit the observations to the target function in each iteration. A commonly seen surrogate model is the Gaussian process $G(\mu(\lambda), \sigma^2(\lambda))$. This method assumes the model

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

predictions at parameter set λ follow the normal distribution. Its mean $\mu(\lambda)$ and variance $\sigma^2(\lambda)$ are calculated using Equations (9) and (10).

$$\mu(\lambda) = \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{Y} \tag{9}$$

$$\sigma^2(\lambda) = k(\lambda, \lambda') - \boldsymbol{k}_*^T \boldsymbol{K}^{-1} \boldsymbol{k}_* \tag{10}$$

where $\boldsymbol{k}_*$ is the vector of covariances between λ and all the previous observations. $\boldsymbol{K}$ is the covariance matrix of all previously evaluated configurations and $\boldsymbol{Y}$ is the observed function values. The quality of the Gaussian process depends solely on the covariance function. A common choice is the Mátern 5/2 kernel, with its hyperparameters integrated out by Markov Chain Monte Carlo (Feurer et al., 2019).

The second component is the acquisition function. It uses the predictive distribution of the probabilistic model, determines the utility of different candidate points, trading off exploration and exploitation (Feurer et al., 2019). A common choice is called the expected improvement (EI) (Feurer et al., 2019), which is shown in Equation (11).

$$E(I(\lambda)) = (y_{min} - \mu(\lambda)) \cdot \Phi\left(\frac{y_{min} - \mu(\lambda)}{\sigma^2(\lambda)}\right) + \sigma^2(\lambda) \cdot \phi\left(\frac{y_{min} - \mu(\lambda)}{\sigma^2(\lambda)}\right) \tag{11}$$

where Φ and ϕ are the standard normal density and standard normal distribution function.

After an initial exploration of the parameters, this study pre-set the range for the mentioned six parameters as $N_T = [0, 1000]$, $l_r = [0, 0.3]$, sub_sample $= [0, 1]$, colsample_bytree $= [0, 1]$, max_depth $= [0, 10]$, and min_child_weight $= [0, 10]$. Then the parameters set λ was optimized using the proposed BO method. Eventually, it is discovered that when $N_T = 200$, $l_r = 0.1$, sub_sample $= 1.0$, colsample_bytree $= 0.8$, max_depth $= 4$, and min_child_weight $= 5$, the XGBoost provides an optimal performance with AUC $= 0.875$.

4.3. Model comparisons

Moreover, a model comparison experiment is conducted to evaluate the performance of XGBoost. Since no other study ever before have studied the relationships between city level factors and the air quality in a national scale, so this study compared the modeling performance between the proposed method and some other latest machine learning methods in air quality research. These include linear based methods (Multiple Linear Regression-MLR, Logistic Regression-LR), typical machine learning methods (Decision Tree-CART, K Nearest Neighbors-kNN, Support Vector Machine-SVM, Artificial Neural Network-ANN), bagging and boosting methods (Gradient Boosted Decision Trees-GBDT, Random Forest-RF, Bagging and Boosting LR/SVM), and deep learning methods (Deep Neural Network-DNN). Details are presented in Table 3.

To make the comparison more objective, 10-fold cross-validation is also applied to the other comparison models, and each algorithm is fine-tuned using the similar BO procedure. It can be seen from Table 3 that (1) linear based methods and typical machine learning methods produced the lowest AUC. This is reasonable since for modeling the non-linear real-world problems, those methods are outdated. (2) Bagging and Boosting methods significantly improved the performance of the weak learners, especially for CART and LR. While for SVM, its improvement is quite rare. (3) Deep learning has gained much attention on various topics in these years. However, in our experiment, its performance did not stand out, and was even worse than GBDT and RF. Note that in this study, we tested the DNN performance using 3 to 10 layers and 4 to 128 neurons per layer, and use the optimal one for comparison. Its lower AUC compared with the top methods may because, in nowadays, most deep learning models are developed for specific problems, for example, CNN for visual and image problem, LSTM for time series and language problems. But for more complicated multi-source and multi-dimensional data features like this study, simply making the network deeper may not easily provide an excellent result. Therefore, it is worth of exploration a well fit deep network structure for this kind of problems. But this is out of the scope of this paper, and can be addressed in future research. (4) XGBoost outperforms all the other algorithms in our 10-fold cross validation test. This proves that it is a reasonable choice in our problem.

4.4. Results analysis

After the parameters of XGBoost are optimized, XGBoost is used to model the relationship between AQI and potential influential factors and measure the feature importance. Based on the modeling results, the top 10 features which have the highest importance are shown in Table 4. The most important feature is PI (personal income), followed by PCP, PPD, CTV, etc. To understand the relationship between AQI and these 10 most important features, detailed analysis are presented in the following sections.

4.4.1. Personal income

As Table 4 indicates, total personal income (PI) has the largest impact on AQI. Note that, in economic data sets, PI refers to the sum of all the incomes received by all the individuals and households in a county. It is an important variable to calculate Gross Domestic Product (GDP). An observation of Fig. 3 suggests that most of the developed areas with high personal income are negative cases, such as Los Angeles and Washington. It seems that PI has a negative correlation with air quality. To further investigate the relationship between PI and AQI, the average total PI of positive counties and negative counties are calculated. Results are presented in Fig. 5. It shows that the average PI of negative counties is almost four times as much as that of positive counties. This indicates that higher PI will lead to larger AQI and worse air quality.

PI is one of the indicators of the economic development level (Indicators, 2003). Higher PI means higher development level to some extent. Thus, based on Fig. 5, it can be concluded that the development level has a high negative correlation with air quality. This is because that as the economy develops, more people move into the county and more factories are built, which requires more energy consumptions and causes more emissions of air pollutants (Gurjar et al., 2008). At the same time, during the urbanization, artificial buildings and facilities occupy more and more areas where used to be green plants, lakes, and wetlands. This weakens the self-adjustment of nature in absorbing and degrading air pollutants (Beckett et al., 2000). Therefore, air quality is deteriorated.

4.4.2. Precipitation and water area

Average daily precipitation (PCP) ranks the second most important influential factors of AQI in Table 4. Previous studies have also proved that air quality is closely related to the precipitation (Ravindra et al., 2003; Pillai et al., 2001; Davies, 1967). To uncover the relationship between PCP and AQI, this paper divides the studied counties into counties with high precipitation (PCP_H) and counties with low precipitation (PCP_L) based on the average PCP of the studied 1014 counties. The distribution of counties with PCP_H and PCP_L is presented in Fig. 6 with the help of GIS. Red means counties with high precipitation and green means counties

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

Table 3
Comparison of the proposed XGBoost method with other machine learning algorithms.

Algorithms	Description	10-fold CV AUC	Improvements using XGBoost
MLR	Linear Based	0.764	14.53%
LR	Methods	0.793	10.34%
CART	Typical Machine Learning	0.674	29.82%
kNN	Methods	0.781	12.04%
SVM		0.855	2.34%
ANN		0.846	3.43%
BB LR	Bagging and Boosting (BB)	0.853	2.58%
BB SVM	Methods	0.858	1.98%
RF		0.861	1.63%
GBDT		0.863	1.39%
DNN	Deep Learning	0.859	1.86%
XGBoost	XGBoost	0.875	N/A

Table 4
Top 10 important features based on XGBoost outputs.

Rank	Feature Abbreviation	Description	Feature Importance
1	PI	Total Personal Income	0.0853
2	PCP	Average Daily Precipitation	0.0612
3	PPD	Density of Power Plants	0.0581
4	CTV	Number of Population Per Square Meter Taking Car, Truck or Van to Work	0.0569
5	2_Vs	Number of Population Per Square Meter With 2 Vehicles available	0.0496
6	DTR	Average Daily Diurnal Temperature Range	0.0421
7	HDD	Heating Degree Days	0.0403
8	CDD	Cooling Degree Days	0.0374
9	Constr	Yearly Earnings in Construction Industry Per Square Meters	0.0358
10	Water	Proportion of Water Areas	0.0351

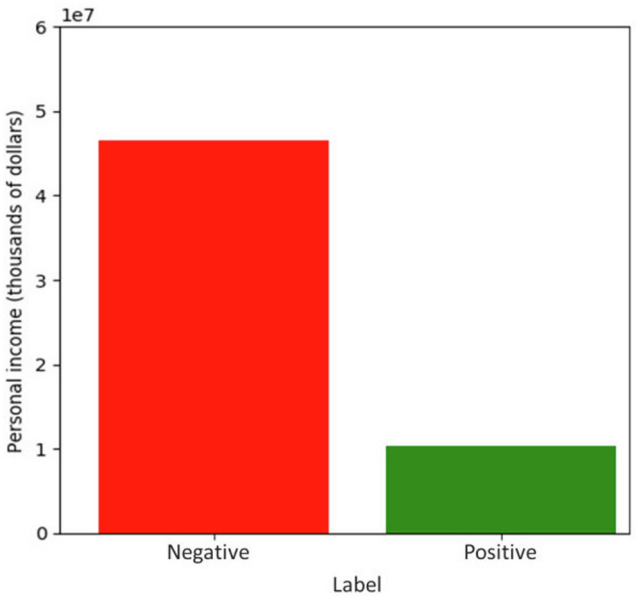

Fig. 5. The average total PI of all the positive and negative counties in the U.S.

with low precipitation. Compared Fig. 6 with Fig. 3, it can be observed that most of the counties with PCP_H are positive counties. Besides, the average PCP of positive counties and negative counties are calculated in Fig. 7. It can be seen that the average precipitation of negative counties is only around two-thirds of that of positive counties. Hence, it can be inferred that precipitation has a positive correlation with air quality.

Another water-related influential factor in Table 4 is the proportion of the water area. Similarly, this study calculates the average proportion of the water area in positive counties and negative counties. Fig. 8 shows the results. It can be observed that the average water area of negative counties is only half of that of positive counties. This suggests that water area is positively related with air quality.

Based on Figs. 7 and 8, it can be seen that precipitation and water area both have a positive relationship with air quality. Higher precipitation and larger water area will lead to better air quality. This is because that on the one hand, rain can remove particulate matter (PM), SO_2, NO_2 as well as other pollutants in the air (Jacob and Winner, 2009). Therefore, air is cleaned and air quality is improved. On the other hand, due to the differences in surface characteristics, air temperature and atmospheric pressure above water area and land area are different (Hewson and Olsson, 1967). This induces the breeze circulation between water area and land area, which will accelerate the air movement. Consequently, dispersion of air pollution is speed up and air quality is improved. Furthermore, compared to land area, air moisture around water area is higher, which has the same function of rainfall of removing air pollutants.

Thus, to achieve better air quality, government can increase artificial precipitation and install urban spraying system to increase the moisture in the air. Also, more environmental protection policies should be introduced to protect natural lake, wetland and vegetation plant. Financial expenditure on urban greening should be increased.

4.4.3. Power plant density

The third most important factor uncovered by XGBoost is power plant density (PPD). Note that this paper only considers the influence of thermal power plants on air quality since hydroelectric power stations, wind farms, solar power stations and others that use clean energies do not or pose little impact on environment. Also, report of Energy Information Administration (EIA) showed that until 2016, around 30% of power plants in America are still thermal power plants (U.S., 2018). To study the impact of PPD on

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

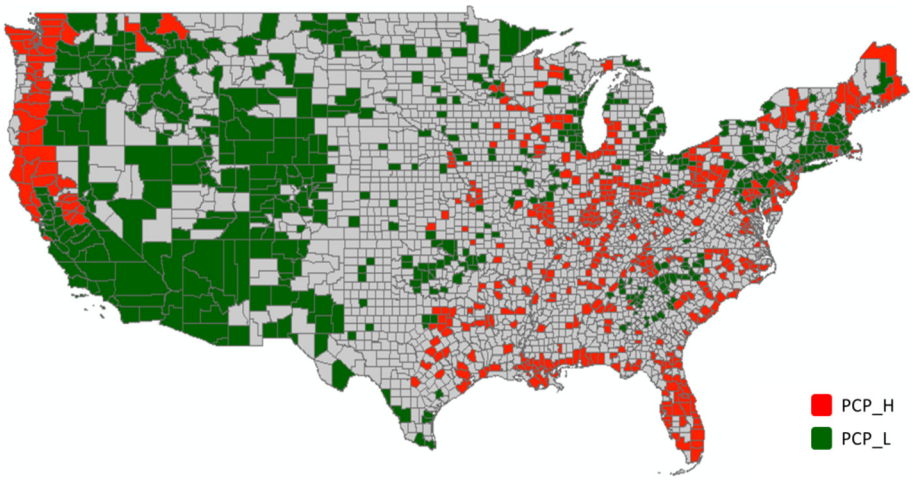

Fig. 6. The distribution of counties with PCP_H and PCP_L.

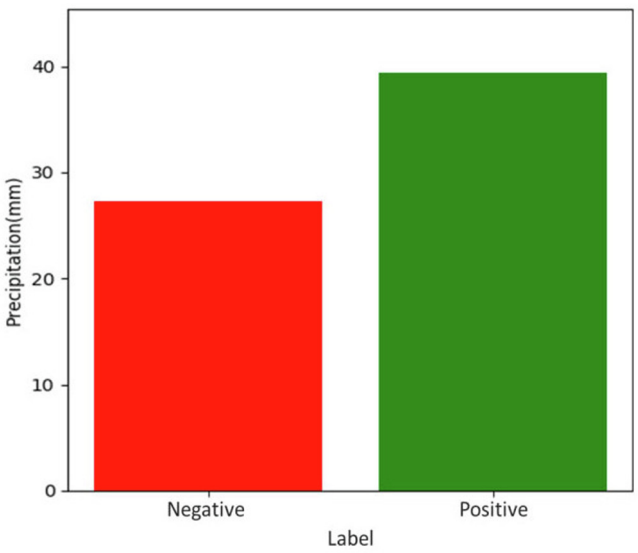

Fig. 7. The relationship between annual precipitation and AQI.

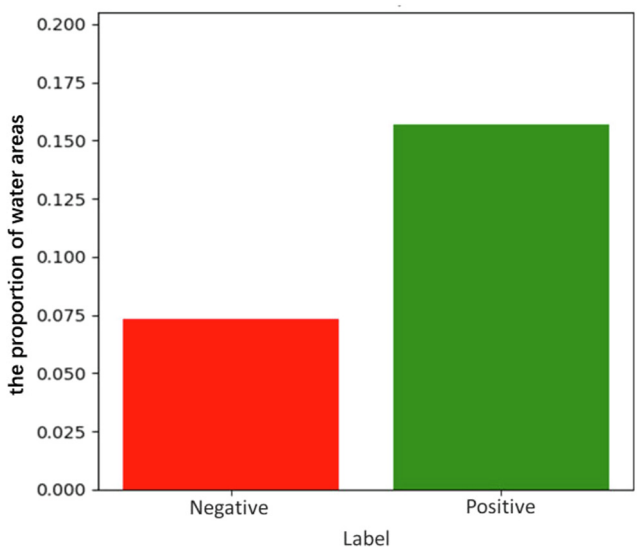

Fig. 8. The relationship between the proportion of water area and AQI.

AQI, the distribution of the density of thermal power plants in America is presented in Fig. 9 using GIS. Red means higher density of thermal power plants while green means lower density. Compared Fig. 9 with Fig. 3, it can be observed that most of the areas with higher density of thermal power plants are negative counties, such as the western coastal area. This means that higher density of thermal power plants will cause larger AQI and worsen the air quality.

The reason behind is that during the power generation process of thermal power plants, many air pollutants, such as NO_x, CO_x, SO_x, PM, etc., are exhausted. This, of course, will deteriorate air quality. In fact, the density of power plants to some degree reflects the demand for energy in an area. Higher density means larger demand. The seventh influential factor heating degree days (HDD), and the eighth factor cooling degree days (CDD) in Table 4 also demonstrate the demand for energy. Therefore, it can be concluded that the demand for energy is highly correlated with air quality. Larger energy demand commands more power plants and therefore, leads to more air pollution. Also, larger energy demand normally means higher personal income and higher economic development level, which further supports the positive correlation between personal income and air quality in Section 4.4.1.

Therefore, to improve air quality, it is important to manage thermal power plants. In the short term, government should introduce more policies to help rectify and improve high-pollution power plants, such as installing more tail gas treatment devices and smoke filter devices. In the long term, government should attach more importance to the development and research of clean energies, such as wind power, hydropower and nuclear power, and increase the number of low-pollution power plants to replace high-pollution plants.

4.4.4. Commuting mode and vehicle ownership

A transportation commuting related factor ranked the fourth. CTV refers to the number of populations per square meter taking car, truck or van to work. This factor can be break down into four categories, including car_truck or van_drove alone (CTV_A), car_truck or van_carpooled (CTV_C), taxicab_motorcycle or other means (TM), car_truck or van (CTV), and public transportation (PT). To explore the impact of commuting mode on AQI, this paper calculates the average number of people taking different commuting modes in positive counties and negative counties, of which the

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

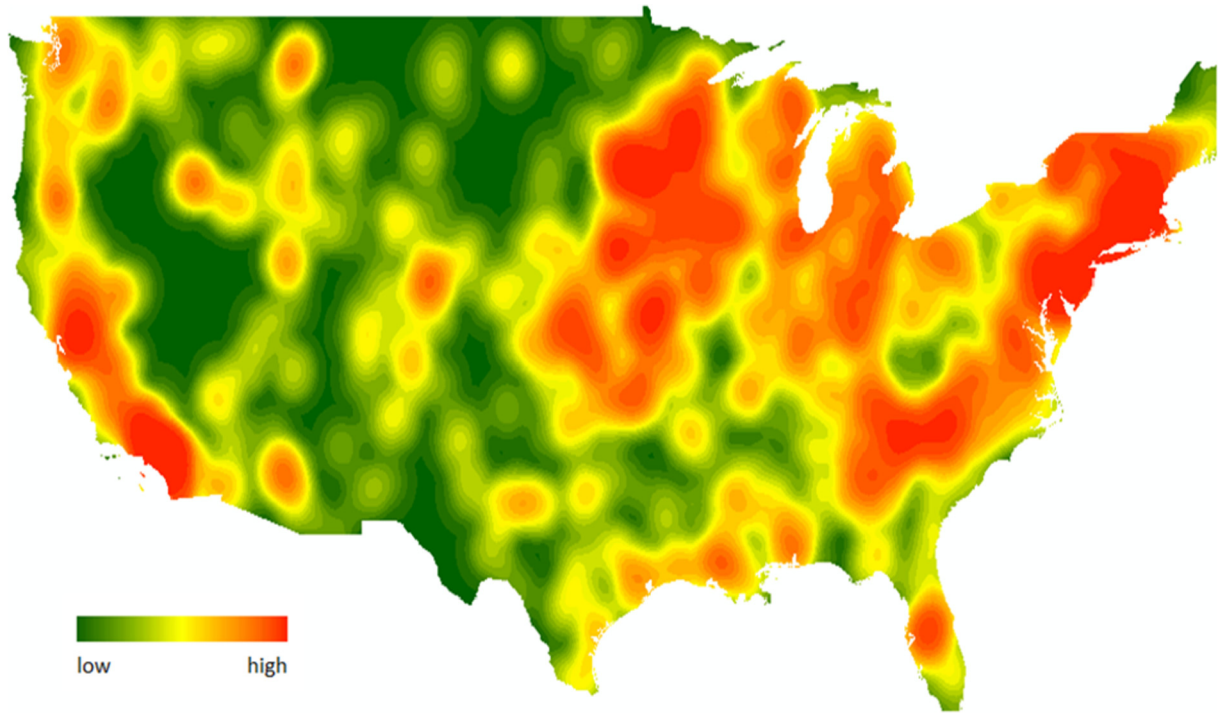

Fig. 9. The distribution of the density of thermal power plants in America.

results are presented in Fig. 10. It can be observed that the number of people using the commuting modes of CTV and CTV_A are much larger in negative counties than positive counties. CTV and CTV_A both reflect the vehicle ownership. Larger CTV means larger vehicle ownership. This leads to potentially high traffic volume/density. As new energy vehicles have not been popularized, most of the vehicles on the road use fuel as driving energy, which will generate a large amount of air pollutants. Consequently, larger traffic volume leads to more serious air pollution (Gurjar et al., 2008; Borrego et al., 2000; Subramani, 2012).

Another vehicle related factor uncovered in Table 4 is the number of population per square kilometer with two vehicles available (2_Vs). Different average values in positive and negative counties are calculated in Fig. 11. It can be seen that the number in negative counties is all higher than that in positive counties. Particularly, the 2_Vs and 3_Vs in negative counties are almost as twice as that in positive counties. This relationship is also resulted from the potentially higher traffic volumn/density.

As the development of transportation must keep pace with the development of economy, limiting the traffic flow or even forbidding the personal vehicles might be unreasonable. Therefore, to mitigate the impact of traffic emissions on air quality, research and

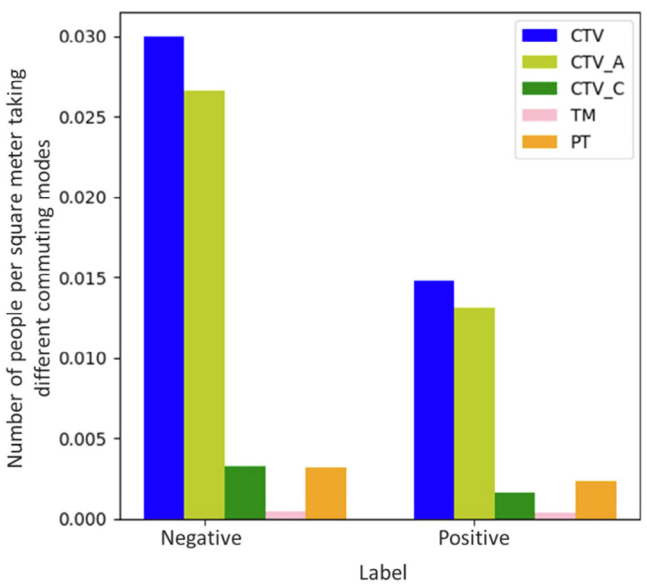

Fig. 10. The relationship between commuting modes and AQI.

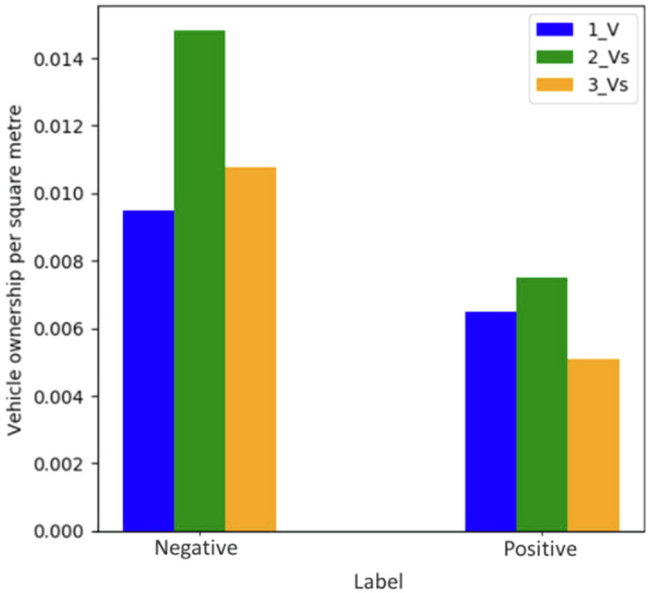

Fig. 11. The relationship between vehicle ownership and AQI.

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

development of new energy vehicles should be strengthened, new policies should be introduced to encourage people to use new energy vehicles. Furthermore, establishing a new energy public transportation system that is accessible in all directions could be effective in solving the problem. It can attract more people to use public transportation, and therefore, reduce the personal driving rate and gradually reduce the number of vehicle ownership.

4.4.5. Diurnal temperature range

Diurnal temperature range (DTR) is also an important influential factor of AQI as Table 4 indicates. To find out the relationship between DTR and AQI, this paper divides the collected DTR range into seven intervals and calculates the average AQI within each interval. Results are shown in Fig. 12. The horizontal axis represents the DTR values and the vertical axis describes the AQI values. It can be seen from Fig. 12 that as DTR increases, AQI gradually decreases. These two variables presented a negative correlation.

DTR represents the variation between the highest temperature and the lowest temperature that occurs during the same day (Diurnal temperature varia, 2018). Changes in DTR can be influenced by multiple possible causes, such as season, urban heat, land use change, geography, etc. In particular, local effects such as urban growth, irrigation and variations in local land use have greater impacts on DTR (Diurnal Temperature Range, 2018). This makes this parameter an important measurement for urban heat island (UHI) effect. The cause effect between DTR and AQI may because of two reasons that relate to UHI. First is from lower DTR to higher AQI. Because lower DTR reflects more significant UHI (Diurnal Temperature Range, 2018), it usually happens at developed urban areas. Those areas consume more energy and have more complicated transportation, both of which worsen the air quality. Second is from higher AQI to lower DTR. Places with worse air quality means denser concentration of PM2.5, PM10, SO_2, O_3, etc. Those pollutants themselves act like a blanket covering the cities, and aggravating the UHI effect. In the daytime, it absorbs more heat than clean air, and in the night, it significantly blocks the heat loss from the urban surface. This phenomenon leads to a lower DTR.

Hence, to improve air quality, efforts should be made to mitigate the UHI effect. It can be accomplished through the changes of land surfaces, such as increasing greening areas, using higher-albedo surface materials of buildings as well as developing green roofs (Oberndorfer et al., 2007). On the other hand, height of buildings should be controlled to accelerate the air movement between urban areas and rural areas.

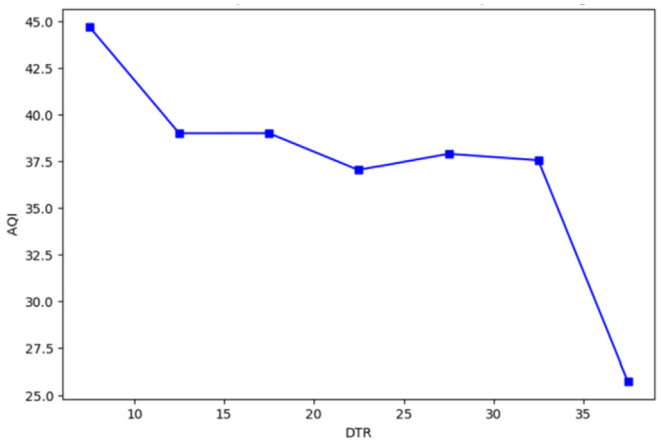

Fig. 12. The relationship between DTR and AQI.

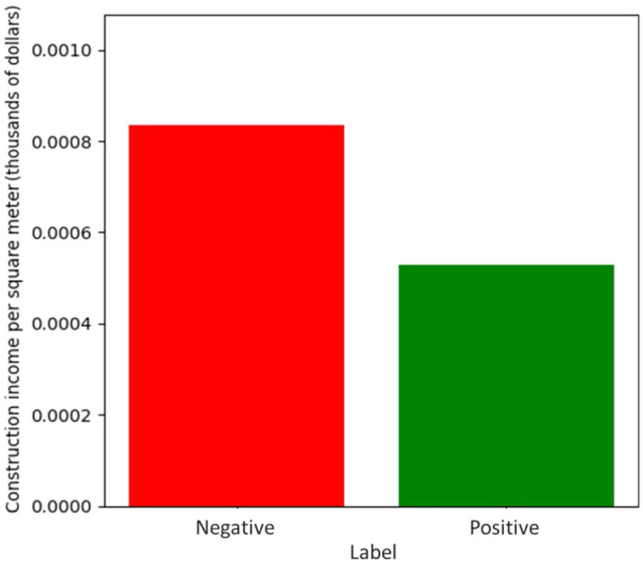

Fig. 13. The relationship between construction income and AQI.

4.4.6. Construction income

Construction income (Constr) is another important influential factor of AQI. Similarly, to figure out the relationship between Constr and AQI, this paper calculates and compares the average construction income of negative counties and positive counties. Fig. 13 presents the results. It can be seen that the construction income of negative counties is much higher than that of positive counties. This means that higher construction income will result in larger AQI and more serious air pollution.

Higher construction income means more construction activities. Construction activities, however, will cause a lot of problems, such as wholesale destruction of vegetation, bare sand land, fly dust and nonstandard transportation of construction waste (Muleski et al., 1995). These all increase the particulate emissions and result in higher PM concentrations. Also, construction industry to some extent reflects the development level of a region. More construction activities mean higher development level, and consequently, more city infrastructures, such as road. Development of transportation infrastructure furthermore means more vehicles and more vehicle emissions. Moreover, some improper construction activities which ignore the disposal of dust will cause severer air pollution.

Therefore, to improve air quality, particulate emissions from construction activities need to be reduced. Many measures can be conducted to effectively reduce particulate emissions. For example, harden the bare land in the construction sites, increase the amount and frequency of water spraying in the construction sites, strictly supervise the transportation of cement, sand and other materials, avoid earthwork excavation, house blasting and other activities that will raise dust in windy days, etc. (Muleski et al., 1995).

5. Conclusion

To conclude, this paper proposes a non-linear framework based on Extreme Gradient Boosting (XGBoost) and Geographical Information System (GIS) to analyze the influential factors of air quality. Abilities of XGBoost in modeling non-linear relationship and measuring feature importance and GIS in managing multiple variables and visualizing results are utilized in this paper. The methodology framework also compares the classification performance of XGBoost with other machine learning models to show the

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

reasonability of choice. To validate the effectiveness of the proposed framework, a case study is conducted in America. Experimental results show that:

- Compared to other machine learning models, XGBoost exhibits better classification performance with its AUC value around 0.871.
- The proposed methodology framework shows great performance in performing multivariate analysis on a national scale.
- Six kinds of features are found to have the largest impacts on the air quality in America among 171 considered factors. They are personal income, precipitation and water area, power plant density, commuting mode and vehicle ownership, diurnal temperature range (DTR) as well as the earnings in the construction industry.
- Personal income, power plant density, commuting mode and vehicle ownership, and construction income all have a negative impact on air quality. Precipitation and water area, and DTR, on the other hand, are positively associated with air quality.
- According to the influential factors, actionable suggestions are provided for the improvement of air quality.

Still, there are limitations in this study. Due to the data availability, the collected AQI dataset only contains the information of 1014 counties, which is less than half of the total number of counties in America. Therefore, many other counties are not studied in this paper. Also, the scope of the discovered results are only applicable to the counties in the U.S. Whether city level or state level has the same influential factors requires further studies to verify. Furthermore, the scope of the proposed method is expected to be applicable in similar spatial analysis in other domains, such as county/city level transportation/traffic accident analysis, county/city level house price analysis, but whether it is as useful as in this study also requires further studies to verify.

References

Air quality Index (AQI) basics. n.d. https://airnow.gov/index.cfm?action=aqibasics.aqi. accessed January 3, 2019.

Ali, S., Tirumala, S.S., Sarrafzadeh, A., 2014. SVM aggregation modelling for spatio-temporal air pollution analysis. In: 17th IEEE Int. Multi Top. Conf., vol. 2014, pp. 249–254. https://doi.org/10.1109/INMIC.2014.7097346.

Beckett, K.P., Freer-Smith, P., Taylor, G., 2000. EFFECTIVE TREE SPECIES FOR LOCAL AIR- QUALITY MANAGEMENT, p. 8.

Borrego, C., Tchepel, O., Barros, N., Miranda, A.I., 2000. Impact of road traffic emissions on air quality of the Lisbon region. Atmos. Environ. 34, 4683–4690. https://doi.org/10.1016/S1352-2310(00)00301-0.

Casalicchio, G., Molnar, C., Bischl, B., 2018. Visualizing the Feature Importance for Black Box Models. ArXiv 180406620 Cs Stat.. http://arxiv.org/abs/1804.06620. accessed August 23, 2018.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 16, 785–794. https://doi.org/10.1145/2939672.2939785.

Cheng, J.C.P., Ma, L.J., 2015. A data-driven study of important climate factors on the achievement of LEED-EB credits. Build. Environ. 90, 232–244. https://doi.org/10.1016/j.buildenv.2014.11.029.

Cheng, J.C.P., Ma, L.J., 2015. A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. Build. Environ. 93, 349–361. https://doi.org/10.1016/j.buildenv.2015.07.019.

Davies, T.D., 1967. Precipitation scavenging of sulphur dioxide in an industrial area. Atmos. Environ. 10, 879–890. https://doi.org/10.1016/0004-6981(76)90143-8, 1976.

Diurnal temperature range. n.d. http://appinsys.com/globalwarming/DTR.htm. accessed September 19, 2018.

Diurnal Temperature Variation, 2018. Wikipedia. https://en.wikipedia.org/w/index.php?title=Diurnal_temperature_variation&oldid=858033209. accessed September 19, 2018.

Feurer, M., Hutter, F., 2019. Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (Eds.), Autom. Mach. Learn. Methods Syst. Chall. Springer International Publishing, Cham, pp. 3–33. https://doi.org/10.1007/978-3-030-05318-5_1.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Geographic Information System, 2018. Wikipedia. https://en.wikipedia.org/w/index.php?title=Geographic_information_system&oldid=853176113. accessed August 6, 2018.

Gurjar, B.R., Butler, T.M., Lawrence, M.G., Lelieveld, J., 2008. Evaluation of emissions and air quality in megacities. Atmos. Environ. 42, 1593–1606. https://doi.org/10.1016/j.atmosenv.2007.10.048.

Haagenson, P.L., 1967. Meteorological and climatological factors affecting Denver air quality. Atmos. Environ. 13, 79–85. https://doi.org/10.1016/0004-6981(79)90247-6, 1979.

Hall, M.A., Smith, L.A., 1997. Feature Subset Selection: a Correlation Based Filter Approach.

Hao, Y., Liu, Y.-M., 2016. The influential factors of urban PM2.5 concentrations in China: a spatial econometric analysis. J. Clean. Prod. 112, 1443–1453. https://doi.org/10.1016/j.jclepro.2015.05.005.

Hewson, E.W., Olsson, L.E., 1967. Lake effects on air pollution dispersion. J. Air Pollut. Control Assoc. 17, 757–761. https://doi.org/10.1080/00022470.1967.10469069.

Indicators, Economic, 2003. Personal Income and Outlays. Investopedia. https://www.investopedia.com/university/releases/personalconsumption.asp. accessed September 21, 2018.

Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. Atmos. Environ. 43, 51–63. https://doi.org/10.1016/j.atmosenv.2008.09.051.

Jun, M.A., Cheng, J.C.P., 2017. Selection of target LEED credits based on project information and climatic factors using data mining techniques. Adv. Eng. Inf. 32, 224–236. https://doi.org/10.1016/j.aei.2017.03.004.

Khan, J., Ketzel, M., Kakosimos, K., Sørensen, M., Jensen, S.S., 2018. Road traffic air and noise pollution exposure assessment – a review of tools and techniques. Sci. Total Environ. 634, 661–676. https://doi.org/10.1016/j.scitotenv.2018.03.374.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T., 2017. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. Environ. Pollut. 231, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114.

Lu, Y., Cohen, I., Zhou, X.S., Tian, Q., 2007. Feature selection using principal feature analysis. ACM 301–304.

Ma, J., Cheng, J.C.P., 2016. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Appl. Energy 183, 193–201. https://doi.org/10.1016/j.apenergy.2016.08.096.

Ma, J., Cheng, J.C.P., 2016. Data-driven study on the achievement of LEED credits using percentage of average score and association rule analysis. Build. Environ. 98, 121–132. https://doi.org/10.1016/j.buildenv.2016.01.005.

Ma, J., Cheng, J.C.P., 2016. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology. Appl. Energy 183, 182–192. https://doi.org/10.1016/j.apenergy.2016.08.079.

Ma, J., Cheng, J.C.P., 2017. Identification of the numerical patterns behind the leading counties in the U.S. local green building markets using data mining. J. Clean. Prod. 151, 406–418. https://doi.org/10.1016/j.jclepro.2017.03.083.

Ma, J., Ding, Y., Gan, V.J.L., Lin, C., Wan, Z., 2019. Spatiotemporal prediction of PM2.5 concentrations at different time granularities using IDW-BLSTM. IEEE Access 7, 107897–107907. https://doi.org/10.1109/ACCESS.2019.2932445.

Monteiro, A., Russo, M., Gama, C., Lopes, M., Borrego, C., 2018. How economic crisis influence air quality over Portugal (Lisbon and Porto)? Atmospheric Pollut. Res. 9, 439–445. https://doi.org/10.1016/j.apr.2017.11.009.

Muleski, G.E., Cowherd, C., Kinsey, J.S., 1995. Particulate emissions from construction activities. J. Air Waste Manag. Assoc. 55, 772–783, 2005.

Ngo, N.S., Zhong, N., Bao, X., 2018. The effects of transboundary air pollution following major events in China on air quality in the U.S.: Evidence from Chinese New Year and sandstorms. J. Environ. Manag. 212, 169–175. https://doi.org/10.1016/j.jenvman.2018.01.057.

Nobre, J., Neves, R.F., 2019. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Syst. Appl. 125, 181–194. https://doi.org/10.1016/j.eswa.2019.01.083.

Oberndorfer, E., Lundholm, J., Bass, B., Coffman, R.R., Doshi, H., Dunnett, N., Gaffin, S., Köhler, M., Liu, K.K.Y., Rowe, B., 2007. Green roofs as urban ecosystems: Ecological structures, functions, and services. Bioscience 57, 823–833. https://doi.org/10.1641/B571005.

Orun, A., Elizondo, D., Goodyer, E., Paluszczyszyn, D., 2018. Use of Bayesian inference method to model vehicular air pollution in local urban areas. Transp. Res. Part Transp. Environ. 63, 236–243. https://doi.org/10.1016/j.trd.2018.05.009.

9 out of 10 people worldwide breathe polluted air, but more countries are taking action, World Health Organ. n.d. http://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action. accessed September 6, 2018.

Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Uotila, P., Tapper, N.J., 2011. Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modelling. Atmos. Environ. 45, 128–136. https://doi.org/10.1016/j.atmosenv.2010.09.032.

Pillai, A.G., Naik, M.S., Momin, G.A., Rao, P.S.P., Safai, P.D., Ali, K., Rodhe, H., Granat, L., 2001. Studies of wet deposition and dustfall at Pune, India, water. Air. Soil Pollut. 130, 475–480. https://doi.org/10.1023/A:1013862024276.

Ravindra, K., Mor, S., Ameena, Kamyotra, J.S., Kaushik, C.P., 2003. Variation in spatial pattern of criteria air pollutants before and during initial rain of monsoon. Environ. Monit. Assess. 87, 145–153. https://doi.org/10.1023/A:1024650215970.

Reich, S.L., Gomez, D.R., Dawidowski, L.E., 1999. Artificial neural network for the identification of unknown air pollution sources. Atmos. Environ. 33,

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955

3045–3052. https://doi.org/10.1016/S1352-2310(98)00418-X.
Selden, T.M., Song, D., 1994. Environmental quality and development: is there a kuznets curve for air pollution emissions? J. Environ. Econ. Manag. 27, 147–162. https://doi.org/10.1006/jeem.1994.1031.
Simon, D., 2013. Evolutionary Optimization Algorithms. John Wiley & Sons.
Subramani, T., 2012. Study OF air pollution due to vehicle emission IN tourism centre. Int. J. Eng. Res. Afr. 2, 11.
Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M., 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform 4, 159. https://doi.org/10.1007/s40708-017-0065-7.
U.S.. Energy information administration (EIA) - ap (n.d.). https://www.eia.gov/environment/emissions/carbon/. accessed September 14, 2018.
Vardoulakis, S., Kassomenos, P., 2008. Sources and factors affecting PM10 levels in two European cities: implications for local air quality management. Atmos. Environ. 42, 3949–3963. https://doi.org/10.1016/j.atmosenv.2006.12.021.
Wang, X., Westerdahl, D., Chen, L.C., Wu, Y., Hao, J., Pan, X., Guo, X., Zhang, K.M., 2009. Evaluating the air quality impacts of the 2008 Beijing Olympic Games: on-road emission factors and black carbon profiles. Atmos. Environ. 43, 4535–4543. https://doi.org/10.1016/j.atmosenv.2009.06.054.
WHO | Air pollution and health: Summary, WHO, 2018. n.d. http://www.who.int/airpollution/ambient/about/en/. accessed September 13, 2018.
Yang, Z., Tang, M., 2018. Does the increase of public transit fares deteriorate air quality in Beijing? Transp. Res. Part Transp. Environ. 63, 49–57. https://doi.org/10.1016/j.trd.2018.04.020.
Zhang, L., Zhan, C., 2017. Machine learning in rock facies classification: an application of XGBoost. In: Int. Geophys. Conf. Qingdao China 17-20 April 2017. Society of Exploration Geophysicists and Chinese Petroleum Society, pp. 1371–1374. https://doi.org/10.1190/IGC2017-351.
Zhang, H., Qiu, D., Wu, R., Deng, Y., Ji, D., Li, T., 2019. Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model. Appl. Soft Comput. 80, 57–79. https://doi.org/10.1016/j.asoc.2019.03.017.
Zheng, H., Yuan, J., Chen, L., 2017. Short-term load forecasting using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation. Energies 10, 1168. https://doi.org/10.3390/en10081168.

Please cite this article as: Ma, J et al., Identification of high impact factors of air quality on a national scale using big data and machine learning techniques, Journal of Cleaner Production, https://doi.org/10.1016/j.jclepro.2019.118955