Statistics and Its Interface Volume 12 (2019) 387–395

Photographic diary: a new estimation approach to $PM_{2.5}$ monitoring

Ke Xu[*], Jianqiao Wang[‡], Rui Pan[†], and Hansheng Wang[*]

Air pollution is a global environmental problem that has been particularly severe in China over the past few years. Among all air pollutants, $PM_{2.5}$ is one of the most hazardous to human health; therefore, monitoring and reducing $PM_{2.5}$ pollution has become an issue of fundamental importance. Despite the comprehensive air quality monitoring system established by Chinese government, it is still a problem in China. Developing an effective and more economical way to monitor $PM_{2.5}$ has become a pressing challenge. In this study, we explore a promising solution: the possibility of recovering $PM_{2.5}$ values using a new haze indicator known as a photographic diary. Based on the related literature, our method is a cost-effective way to monitor $PM_{2.5}$ at any location and at any point in time with an acceptable accuracy. The government could use our method to conduct data quality monitoring and detect outliers. We also constructed features that the general public could use and interpret directly. Our method allows them to monitor air quality and protect the environment using their cellphones.

Keywords and phrases: $PM_{2.5}$ pollution, Haze indicator, Photographic diary.

1. INTRODUCTION

Air pollution has recently become an important public health problem worldwide (Carvalho-Oliveira et al., 2017). Of all the air pollutants, $PM_{2.5}$ is a key contributor, especially in China (Wang et al., 2017). $PM_{2.5}$ refers to fine particulate matter with aerodynamic diameter of 2.5 μm or less (Zhang et al., 2017). $PM_{2.5}$ is particularly hazardous due to its extremely small size; it is small enough to enter the human bloodstream, become lodged deeply in the lungs, accumulate in the respiratory system, and eventually lead to severe health problems (Chowdhury and Dey, 2016). It also has harmful impacts on agriculture (Burney and Ramanathan, 2014), the climate (Huang et al., 2014), ecosystems (Mahowald, 2011) and other aspects of the environment in general (Guo et al., 2014). Therefore, monitoring air pollutants, especially $PM_{2.5}$ concentrations, has caught the attention of governments and researchers (Liang et al., 2015).

To this end, the Chinese government has established a comprehensive air quality monitoring system. It costs a total of 1.82 billion yuan in investments on the construction of the monitoring system, according to Ministry of Environmental Protection of the People's Republic of China (MEPC) (MEPC, 2015). The system includes several air monitoring stations that typically use a suite of sensors to monitor six air pollutants, i.e., $PM_{2.5}$, SO_2, NO_2, PM_{10}, CO, and O_3. The monitoring stations make relatively accurate measurements; however, they are usually sparsely and preferentially located. Up to January 2015, the MEPC (www.zhb.gov.cn) has $PM_{2.5}$ data for 338 cities (Liang et al., 2016), accounting for only half the total number of cities in China. Furthermore, since $PM_{2.5}$ varies dramatically over short distances (Cheng et al., 2017), the number of monitoring stations within any given city is very insufficient. For example, Beijing–a city with a population of more than 20 million and a land area of 16,410 km^2 (Zhao, 2016), has only 12 state-controlled monitoring stations. Therefore, developing an effective and economical way to monitor $PM_{2.5}$ has become an important and challenging problem.

Various alternatives have been proposed as solutions in past studies; these can be classified as direct and indirect methods. The direct methods aim to develop low-cost sensors to monitor air quality (Alvarado et al., 2015; Gao et al., 2016). Although these kinds of sensors can be deployed on a large scale and record data in real time, the accuracy of their measurements need to be further improved (Chen et al., 2017). For the indirect methods the use of imaging data has been widely adopted. One type of imaging data used are remotely sensed satellite images, which are global images acquired from satellites (Wang and Christopher, 2003). Many researchers rely on them to estimate ground-level $PM_{2.5}$ (Lin et al., 2015). However, such images have relatively low spatial and temporal resolutions (Levy et al., 2013; Crosson et al., 2012). They are also affected by unstable meteorological or geographical conditions (Ma et al., 2014).

Compared to this, ground photos are more easily to obtain at any given time or place. They can be conveniently

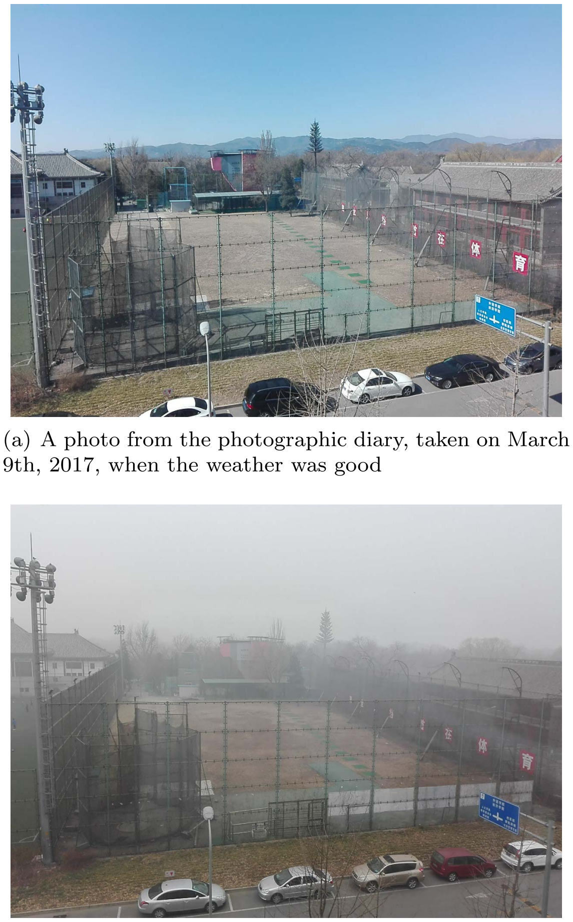
[*] Ke Xu and Hansheng Wang are supported by National Natural Science Foundation of China (Grant No. 11525101, No. 71332006, No. 71532001), and China's National Key Research Special Program (No. 2016YFC0207704).

[†] Rui Pan is supported by National Natural Science Foundation of China (NSFC, No. 11601539, No. 11631003), and Fundamental Research Funds for the Central Universities (No. QL18010).

[‡] Jianqiao Wang is the corresponding author.

obtained using a cellphone or surveillance camera. Therefore, the possibility of estimating $PM_{2.5}$ values from ground photos is worth exploring. Mao et al. (2014) estimate the haze factor using the atmospheric scattering model. Li et al. (2015) make further improvements by considering both the transmission and depth. However, the ground photos they used were taken at different sites; therefore, large visual variations in the scenes might cause significant problems with estimation accuracy. In our work, ground photos are taken automatically by a camera fixed on the window of an office at Peking University; therefore, our photos share a background and only the difference in air quality causes a difference in the photos. The photos we analyzed were taken hourly from 08:00 to 12:00 from December 20th, 2016 to March 9th, 2017. These photos constitute our *photographic diary* data, and some of them are displayed in Figure 1 and Figure 2. As it is expected, photos taken when air quality was good are typically clear; see Figure 2(a). In contrast, photos taken under poor air quality conditions are likely to be vague; see Figure 2(b). This suggests that $PM_{2.5}$ values could be estimated from photographic diary data with reasonable accuracy.

Figure 1. There were 100 photos in our photographic diary. Each photo showed a playground, a close view of cars, and the faraway mountain. The diary can provide an intuitive understanding of air quality in Beijing.



(a) A photo from the photographic diary, taken on March 9th, 2017, when the weather was good

(b) A photo from the photographic diary, taken on December 21st, 2016, when the weather was bad

Figure 2. A view from an office at Peking University, showing different air quality days. They share the same background but are very different in terms of other details. In the Figure 2(a), every detail of the faraway trees and mountains can be seen where the $PM_{2.5}$ value is 5. In the Figure 2(b), hardly any detail can be seen when the $PM_{2.5}$ value is 366.

Therefore, in this work, we propose a linear regression method to estimate $PM_{2.5}$ values with that data and construct several meaningful features derived from it. The first feature is the proportion of blue pixels, where the color was carefully defined according to HSI space (Gonzalez and Woods, 2006). This was done because of the intuition that a blue sky often indicates a good air quality. As Figure 2 shows, the sky is very blue when the $PM_{2.5}$ value is low. The second feature is inspired by the idea of Sure Independence Screening (SIS); further discussion of this idea can be found in Fan and Lv (2008). SIS is a very well-known idea and has many extensions. Here, we take each pixel as a variable and select the one most correlated with $PM_{2.5}$. To the best of our knowledge, very few researchers have estimated $PM_{2.5}$ in this way. The third feature considers air clarity by transmission, calculated according to the atmospheric scattering model (He et al., 2011). This model describes the formation of a hazy image as a result of atmospheric interaction between particulate matter and light. The light from the scene will become very attenuated before reaching the camera due to the scattering and absorption of the haze particles. Therefore, we use the clarity feature to measure the loss of such a visual effect.

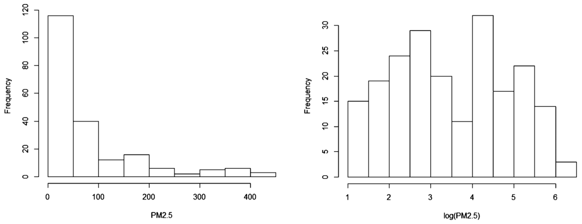
Our paper provides three major contributions. Firstly, our method is cost effective. The data used can be collected easily either by mobile devices (e.g., cellphones) and/or surveillance cameras, any of which are cheap. Secondly, our method uses photographic diary data, which is high-resolution at any location and point in time. This is because the devices that produce the photographic diary data could theoretically cover any place in which human activity occurs. Finally, our method is reasonably accurate to an acceptable level. In terms of real data analysis, our model performs competitively when compared to some state-of-the-art methods. It is the simplest one that offers the best data interpretability.

The rest of this paper is organized as follows. Section 2 introduces notations and features and is then followed by a descriptive analysis of the latter. In Section 3, we present our model results and compare it with those from some other methods. We also demonstrate a potential application of our model. We conclude this article with some interesting topics for future study in Section 4.

2. DATA AND DESCRIPTIVE ANALYSIS

2.1 Introduction to the data

The $PM_{2.5}$ values we collected were measured by the state-owned WanLiu monitoring station (see the Beijing Municipal Environmental Monitoring Center (http://www.bjmemc.com.cn/) for more details). The histogram of $PM_{2.5}$ measurements is given in Figure 3(a). It shows that all $PM_{2.5}$ values are positive and that the distribution is heavily skewed. Therefore, we conducted a log transformation for the $PM_{2.5}$ values as the response variable of interest. The distribution of these values is given in Figure 3(b), which shows an approximately symmetric distribution. We define the response vector as $Y = (Y_1, Y_2, \ldots, Y_T)^\top \in \mathbb{R}^T$ with $T = 206$, and Y_t corresponds to the log transformed $PM_{2.5}$ value measured at time point t.



(a) A histogram of $PM_{2.5}$. (b) A histogram of the response variable.

Figure 3. Left-hand panel: the histogram of $PM_{2.5}$ concentration, showing a heavily right-skewed distribution. Right-hand panel: the histogram of the response variable, showing an approximately symmetric distribution. The skewed distribution in the latter is eliminated using logarithmic transformations.

Let $\mathcal{D} = \{(D_t^R, D_t^G, D_t^B) : 1 \leq t \leq T\}$ be the photographic diary data, where D_t^R (D_t^G, D_t^B) is the red (green, blue) pixel matrix obtained at time point t, then $D_t^R = (D_{t,mn}^R) \in \mathbb{R}^{M \times N}$, where $D_{t,mn}^R \in [0,1]$ with $1 \leq m \leq M (M = 1280)$ and $1 \leq n \leq N (N = 960)$. D_t^G and D_t^B are defined in a similar manner. We follow Hamilton (2004) and define a gray pixel matrix as $D_t = 0.299 D_t^R + 0.587 D_t^G + 0.114 D_t^B$. Next, we define a hue pixel matrix as $D_t^H = (D_{t,mn}^H) \in \mathbb{R}^{M \times N}$, following Gonzalez and Woods (2006). Here, $D_{t,mn}^H = \theta$ if $D_{t,mn}^B \leq D_{t,mn}^G$, and $D_{t,mn}^H = 360 - \theta$ in other instances, where

$$\theta = \cos^{-1}\left[\frac{\{(D_{t,mn}^R - D_{t,mn}^G) + (D_{t,mn}^R - D_{t,mn}^B)\}}{2\{(D_{t,mn}^R - D_{t,mn}^G)^2 + (D_{t,mn}^R - D_{t,mn}^B)(D_{t,mn}^G - D_{t,mn}^B)\}^{\frac{1}{2}}}\right].$$

The first feature is called the "Blue Pixel Proportion" (BPP). It measures the amount of blue in a photo. A photo taken when air quality is good should generally include a sky with a large number of blue pixels. In contrast, a photo taken when air quality is bad would include a sky with a smaller number of blue-colored pixels. According to the standardized X11 color names (the X.org source code, 1989), we define a pixel (m, n) as "sky blue" if the value of its hue is $D_{t,mn}^H = 197°$. In practical terms, there are few pixels whose hue values equal exactly 197°. Therefore, we define a pixel $(m, n) \in \Omega$ to be a "blue pixel" if $D_{t,mn}^H \in [197° - a, 197° + b]$, where $a, b > 0$ are both tuning parameters selected using the marginal two-fold cross-validation method[1]. Our empirical results suggest the best choice is about $a = 37$ and $b = 35$. The tuning parameters a and b are selected using the cross-validation method. This lead to the first feature, the BPP, to be $BPP = (BPP_1, BPP_2, \ldots, BPP_T)^\top \in \mathbb{R}^T$, where

$$BPP_t = \frac{1}{MN} \sum_{(m,n) \in \Omega} 1\Big\{D_{t,mn}^H \in [197° - a, 197° + b]\Big\},$$

and $1(\cdot)$ is the indicator function.

The second feature is called the "Most Correlated Pixels" (MCP) and is inspired by SIS (Fan and Lv, 2008). We consider each pixel as a variable and look for those that are the most correlated with our response variable of interest. For a pixel $(m, n) \in \Omega = \{(m, n) : 1 \leq m \leq M, 1 \leq n \leq N\}$, we define $D_{mn} = (D_{1,mn}, D_{2,mn}, \ldots, D_{T,mn})^\top \in \mathbb{R}^T$ as the the grey value vector across the time period. Then the absolute value of the sample correlation coefficient between the pixel (m, n) and the response variable Y is

$$|\hat{\rho}(D_{mn}, Y)| = \left| \frac{\sum_{t=1}^T (D_{t,mn} - \overline{D}_{mn})(Y_t - \overline{Y})}{\sqrt{\sum_{t=1}^T (D_{t,mn} - \overline{D}_{mn})^2 \sum_{t=1}^T (Y_t - \overline{Y})^2}} \right|,$$

where $\overline{D}_{mn} = \sum_{t=1}^T D_{t,mn}/T$ and $\overline{Y} = \sum_{t=1}^T Y_t/T$. We define $(m_{(i)}, n_{(i)})$ as the pixel associated with the i-th largest

[1] We first randomly split the data into two equal-sized parts. One was used for training the other for testing. Next we constructed a BPP feature for a given value of a and b. We then conducted a marginal regression on the training dataset with the log transformed $PM_{2.5}$ values as the response variable and BPP as the predictor variable. This lead to a univariate regression model. The corresponding squared prediction error is then computed, and the (a, b) combination with the lowest prediction error is selected.

value of $\{|\hat{\rho}(D_{mn}, Y)| : (m,n) \in \Omega\}$, where $1 \leq i \leq MN$. This leads to $MCP = (MCP_1, MCP_2, \ldots, MCP_T)^\top \in \mathbb{R}^T$, where $MCP_t = \sum_{i=1}^{I} D_{t,m_{(i)}n_{(i)}}/I$ with $I = 500$.

The third feature is called the "Air Clarity Quantile" (ACQ) and it concerns air clarity. We follow He et al. (2011) and, for each pixel $(m,n) \in \Omega$, define air clarity as

$$T_{t,mn} = 1 - \min_{(x,y)\in\Phi(m,n)} \left\{ \min_{C\in\{R,G,B\}} \left(\frac{D^C_{t,xy}}{A} \right) \right\},$$

where

$$A = \max \left\{ \min_{(x,y)\in\Phi(m,n)} \left(\min_{C\in\{R,G,B\}} D^C_{t,xy} \right) \right\} \in \mathbb{R}^{M\times N}.$$

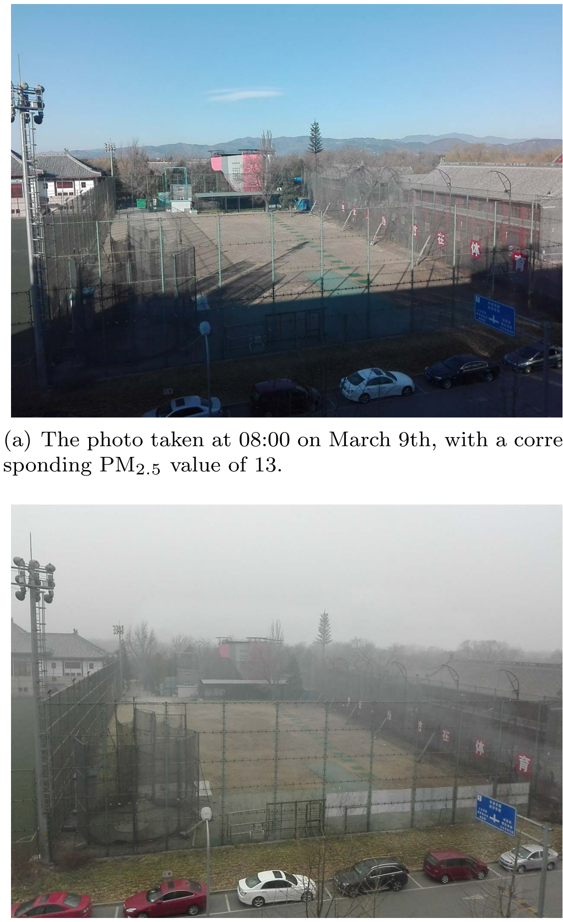
A is the atmospheric background light and $\Phi(m,n) = \{(x,y) : |x-m| \leq 15, |y-n| \leq 15, x \geq 1, y \geq 1\}$ is a local patch centered at (m,n). Air clarity refers to the portion of light that reaches the camera without being scattered and helps distinguish cloudy days from hazy days. On hazy days, the increased $\mathrm{PM}_{2.5}$ concentration can change the light refraction and reduce $T_{t,mn}$. In contrast, $T_{t,mn}$ is high on a cloudy day. This leads to the third feature becoming $ACQ = (ACQ_1, ACQ_2, \ldots, ACQ_T)^\top \in \mathbb{R}^T$, where $ACQ_t = \hat{Q}_q(T_{t,mn})$, the q-th sample quantile of $\{T_{t,mn} : (m,n) \in \Omega\}$ with $q = 0.3$. The quantile q is a tuning parameter selected using the marginal two-fold cross-validation method[2].

2.2 Descriptive analysis of the features

Descriptive statistics for the features are shown in Table 1. For the BPP, the mean is 0.35, which is reasonably close to its median of 0.36. The maximum is 0.60, which was recorded at 08:00 on March 9th, as Figure 4(a) shows. The corresponding $\mathrm{PM}_{2.5}$ value is 13, which suggests that the air quality was good. The minimum value is 0.03, which was recorded at 08:00 on January 5th, as Figure 4(b) shows. The corresponding $\mathrm{PM}_{2.5}$ value is 280, which suggests the air quality was bad. Next, we consider the MCP feature. The maximum value is about three times the minimum. We present a heatmap in Figure 5 to show the positions of the most correlated pixels. It uses white to black to represent different $|\hat{\rho}(D_{mn}, Y)|$ values from 0 to 1. The positions of the most correlated pixels, which are around trees, are plotted in the darkest color. The MCP of Figure 6(a) is 0.28, which is much smaller than that of Figure 6(b) (i.e., 0.47). This reveals that severe air pollution leads to a large MCP. Furthermore, if the grey values of the pixels are larger, the corresponding pixels would appear more white. This conclusion is consistent with the visual intuition from Figure 6. A similar phenomenon is also observed for the third feature, ACQ. The ACQ of Figure 4(a) is 0.39, which is much larger than the ACQ of Figure 4(b) (i.e., 0.20). This indicates that the ACQ is larger on a better air quality day. The range of the ACQ is from 0.09 to 0.55, suggesting that the ACQ changes significantly over time.

Table 1. Summary statistics: the mean, standard deviation (SD), minimum (Min), median, and maximum (Max) for each feature

Feature	Mean	SD	Min	Median	Max
Blue Pixels Proportion (BPP)	0.35	0.13	0.03	0.36	0.60
Most Correlated Pixels (MCP)	0.29	0.08	0.18	0.26	0.54
Air Clarity Quantile (ACQ)	0.28	0.10	0.09	0.27	0.55



(a) The photo taken at 08:00 on March 9th, with a corresponding $\mathrm{PM}_{2.5}$ value of 13.

(b) The photo taken at 08:00 on January 5th, with a corresponding $\mathrm{PM}_{2.5}$ value of 280.

Figure 4. The upper photo corresponds to the largest BPP, 0.60. The lower photo corresponds to the smallest BPP, 0.03. The ACQ of the upper panel, 0.39, is much larger than the ACQ of the lower panel, 0.20.

[2] First we randomly split the data into two equal-sized parts. One was be used for training and the other for testing. Next, we constructed an ACQ feature for a given q. We then conducted a marginal regression on the training dataset, with the log transformed $\mathrm{PM}_{2.5}$ concentration as the response variable and ACQ as the predictor variable. This lead to a univariate regression model. Then corresponding squared prediction error is computed, after which the tuning parameter q with the lowest prediction error is selected.

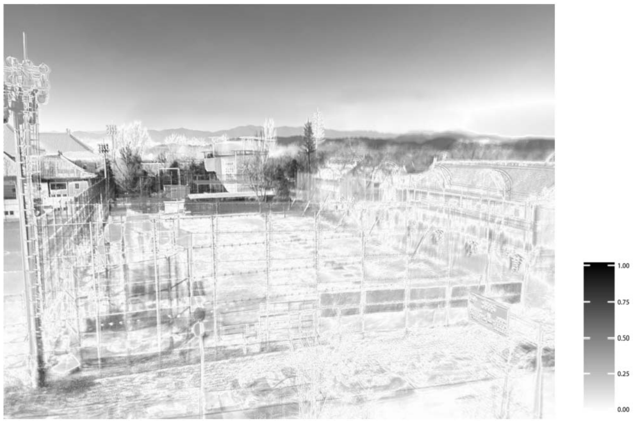
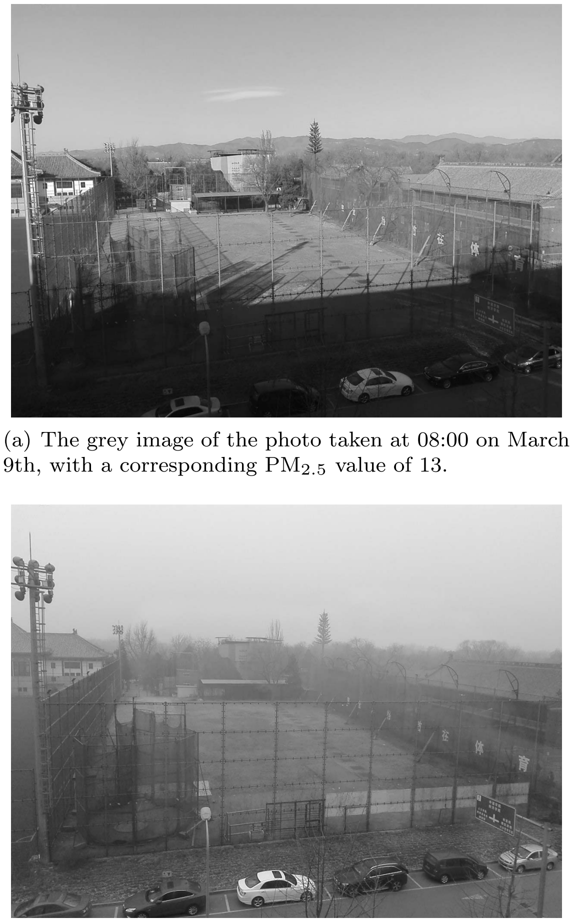


Figure 5. The heatmap uses different colors to represent varying $|\hat{\rho}(D_{mn}, Z)|$ values for different pixels. It shows the color and the position of the most correlated pixels.



(a) The grey image of the photo taken at 08:00 on March 9th, with a corresponding $PM_{2.5}$ value of 13.

(b) The grey image of the photo taken at 08:00 on January 5th, with a corresponding $PM_{2.5}$ value of 280.

Figure 6. The upper photo, taken when the air quality was good, corresponds to an MCP of 0.28; the lower photo, taken when the air quality was bad, corresponds to an MCP of 0.47. The view of the corresponding pixels in the lower photo is whiter and more likely to disappear than it in the upper one.

According to the Technical Regulation on Ambient Air Quality Index published by the MEPC, we classified the continuous $PM_{2.5}$ values into four different categories A, B, C, and D, representing good, moderate, unhealthy for sensitive groups, and very bad air quality, respectively. Further detailed descriptions are given in Table 2. We used boxplots to further explore the potential relationship between each feature and $PM_{2.5}$ value; see Figure 7.

For Figure 7(a), the median of BPP decreases monotonically as the air quality worsens from A to D. Among all the categories, category D has the least variability. This shows that when photos are taken when air quality is very bad, the proportion of blue pixels in these photos are very similar. In Figure 7(b), the MCP is positively correlated with the $PM_{2.5}$ value; specifically, there is a monotonically increasing trend between the median of the MCP and the categories A to D. Figure 7(c) shows a monotonically decreasing pattern between the median of the ACQ and the different categories A to D. Among all the categories, category A has a much larger median ACQ value than the other categories do. This suggests that a considerable proportion of light can be easily scattered by $PM_{2.5}$ on a hazy day. Therefore, ACQ is effective at distinguishing between haze-free and hazy weather.

3. THE MODEL AND POTENTIAL APPLICATIONS

3.1 Regression analysis

To model the relationship between the $PM_{2.5}$ value and the model features, we used the following linear regression:

$$Y = \beta_0 + \beta_1 BPP + \beta_2 MCP + \beta_3 ACQ + \varepsilon, \tag{1}$$

where β_0 is the intercept, β_i is the coefficient corresponding to the i-th feature with $1 \leq i \leq 3$, and ε is the random error term. All the features were standardized to have a mean of 0 and a variance of 1. We conducted an ordinary least squares estimation, and the estimated coefficients are shown in Table 3. The corresponding standard errors, t-values, and p-values are also reported. The adjusted R-squared value of our model is 0.79.

All coefficients are significant at the 0.05 level. The signs of the estimated coefficients also met our expectations. The estimated coefficient of the first feature, BPP, in particular, is -0.37 (SE=0.08). This means that a rise in BPP is often accompanied by a decline in $PM_{2.5}$ value when the effects of the other features are controlled for, because a larger proportion of blue pixels often indicates better air quality. The estimated coefficient of the second feature, MCP, is 0.82 (SE=0.07). This suggests that MCP is positively correlated with $PM_{2.5}$ values when all the other independent variables are held constant; in other words, the $PM_{2.5}$ values increase

Table 2. $PM_{2.5}$ Value Scale

Level	$PM_{2.5}$ Concentration in China (μg/m3)	Air Quality Grade	Category	Health Implications	Cautionary Statement (for $PM_{2.5}$)
1	0∼35	Good	A	Air quality is considered satisfactory, and air pollution poses little or no risk.	None.
2	35∼75	Moderate	B	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory conditions such as asthma, should limit prolonged outdoor exertion.
3	75∼115	Unhealthy for Sensitive Groups	C	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory conditions such as asthma, should limit prolonged outdoor exertion.
4	115∼150	Unhealthy	D	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.	Active children and adults, and people with respiratory conditions such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
5	150∼250	Very Unhealthy		Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory conditions such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
6	More than 250	Hazardous		Health alert: everyone may experience more serious health effects.	Everyone should avoid all outdoor exertion.

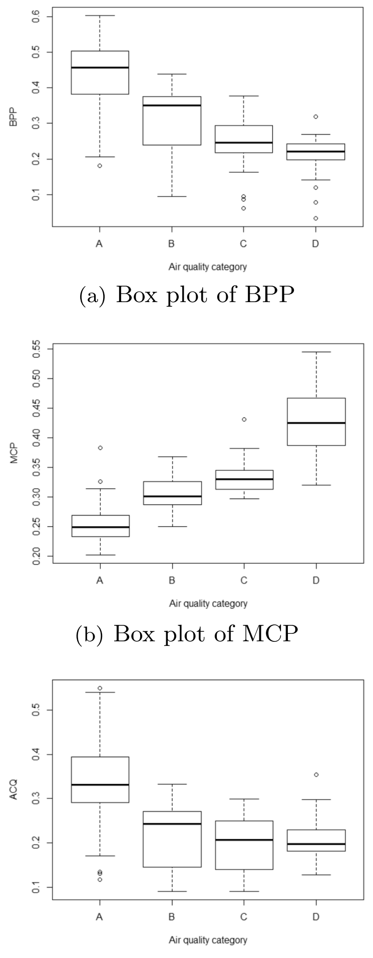
when the MCP values increase. For the third feature, ACQ, the estimated coefficient is -0.16 (SE=0.07). This suggests a negative correlation between ACQ and $PM_{2.5}$ value when the effect of other features was adjusted for. As the ACQ value increases, the corresponding $PM_{2.5}$ value decreases, revealing that clearer air often suggests better air quality.

Table 3. Regression coefficients of the model, estimated with a small sample of T=206 and the corresponding standard errors, t-values, and p-values

Feature	Estimated Coefficient	Standard Error	t-value	p-value
Intercept	3.52	0.04	79.35	< 0.001
Blue Pixels Proportion (BPP)	-0.37	0.08	-4.73	< 0.001
Most Correlated Pixels (MCP)	0.82	0.07	12.57	< 0.001
Air Clarity Quantile (ACQ)	-0.16	0.07	-2.28	0.024
Adjusted R squared	0.79			
P-value of F-Test	< 0.001			

3.2 Competing methods

For comparison purposes, this study also considered several competing methods. Prediction accuracies were evaluated using 5-fold cross validation. The dataset was divided equally into five subsets and the method was repeated five times. Each time, four subsets were put together to form a training set while the other was used as a test set. The observed values in the test set were defined as $Y^* = (Y_1^*, Y_2^*, \ldots, Y_{T^*}^*)^\top = (Y_{t_1}, Y_{t_2}, \ldots, Y_{t_{T^*}})^\top \in \mathbb{R}^{T^*}$, where $1 \le T^* \le T$, and $1 \le t_1 \le t_2 \le \cdots \le t_{T^*} \le T$. Then we defined the corresponding model predicted values as $\hat{Y^*} = (\hat{Y}_1^*, \hat{Y}_2^*, \ldots, \hat{Y}_{T^*}^*)^\top = (\hat{Y}_{t_1}, \hat{Y}_{t_2}, \ldots, \hat{Y}_{t_{T^*}})^\top \in \mathbb{R}^{T^*}$. Therefore, we defined the out-sample R squared value as $1 - \sum_{t=1}^{T^*}(\hat{Y}_t^* - Y_t^*)^2 / \sum_{t=1}^{T^*}(Y_t^* - \bar{Y}_t^*)^2$, where $\bar{Y}_t^* = \sum_{t=1}^{T^*} Y_t^*/T^*$. The average value of all five trials was denoted R^2_{out}. We repeated the process randomly a total of 500 times and compare the R^2_{out} values of the different methods. The specific competing methods tested were gradient boosting, support vector regression, random forest, and a neural network. The R^2_{out} values are summarized in the boxplot in Figure 8, which shows that the performances of the different methods are comparable. However, the linear model (1) is clearly the simplest one with the best interpretability.



(a) Box plot of BPP

(b) Box plot of MCP

(c) Box plot of ACQ

Figure 7. Boxplots of the three features for the different air quality categories A, B, C, and D, which represent good, moderate, unhealthy for sensitive groups, and very bad air quality, respectively.

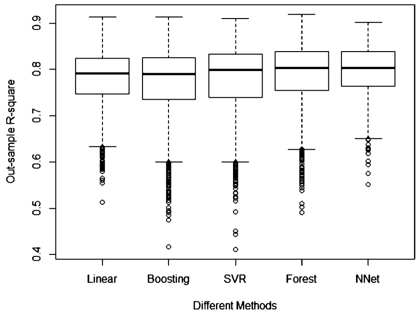


Figure 8. The boxplot of the out-sample R squared value for different models. From left to right, the methods on the horizontal axis are linear regression, gradient boosting, support vector regression, random forest, and neural network, respectively.

3.3 Detecting outliers

One potential application of this model is in the detection of outliers, for which we used a residual plot. As shown in Figure 9, we found that there was one potential outlier, which is a value that seems to have been recorded incorrectly. This outlier was generated by the 89th photo, which was taken at 11:00 on January 17, 2017 (see Figure 10). The $PM_{2.5}$ value actually recorded was 7, but the model estimated a $PM_{2.5}$ value of 105; there was a clear discrepancy between the two, and we were inspired to investigate the reason for it.

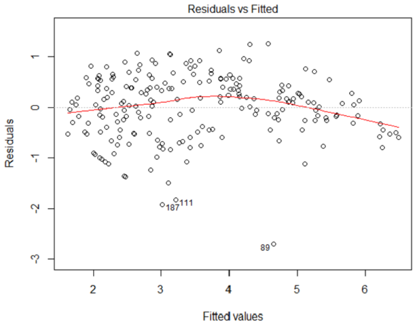


Figure 9. The residual plot of the linear model. The 89th photo appears to have been estimated incorrectly, as there is a clear discrepancy between the model estimated $PM_{2.5}$ value and the actual value recorded.

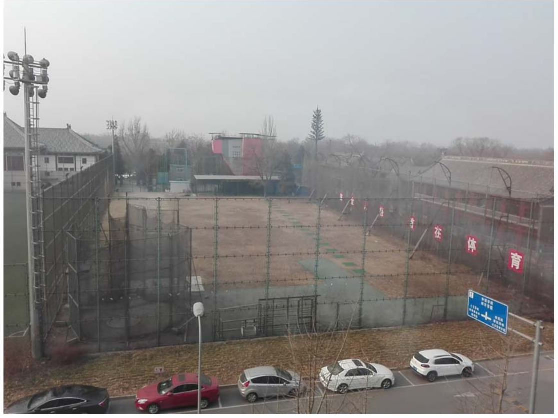


Figure 10. The photo taken at 11:00 on January 17, 2017.

Firstly, we accessed a weather report website (www.tianqi.com) to check the weather on that particular day. The data on this site comes from the China Meteorological Administration. We found that this day was a hazy day, as recorded by the weather report. Secondly, we constructed a time series plot for the hourly $PM_{2.5}$ value recorded at WanLiu station on that day; see Figure 11, and noticed that all recorded $PM_{2.5}$ values were far above 100 except for the time point in question. Finally, we studied all the stations within 30 km around WanLiu Station. We found that all the $PM_{2.5}$ values recorded were far above 100, with a minimum of 129 measured at ZhiWuYuan station; see Table 4 for details. Therefore, all evidence suggests that the estimated $PM_{2.5}$ value seems realistic.

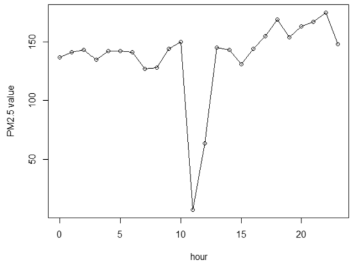


Figure 11. A time series plot for the $PM_{2.5}$ value recorded for 24 hours on January 17, 2017. There is an outlier at 11:00.

Table 4. $PM_{2.5}$ values from air quality monitoring stations around WanLiu Station. Letters "a" to "p" represent the following different stations, respectively: BeiBuXinQu, ZhiWuYuan, AoTiZhongXin, XiZhiMenBei, MenTouGou, GuanYuan, DongSi, NongZhanGuan, DongSiHuan, GuCheng, FengTaiHuanYuan, WanShouXiGong, YongDingMenNei, TianTan, and NanSanHuan.

Site	PM2.5	Site	PM2.5	Site	PM2.5
a	134	g	144	l	169
b	129	h	155	m	161
d	163	i	143	n	149
e	159	j	147	o	142
f	133	k	140	p	154

4. CONCLUSION

In this study, we explored an approach to estimate and recover $PM_{2.5}$ values from a photographic diary. Our method provides a cost-effective way to monitor $PM_{2.5}$ values at any location and any time, with acceptable accuracy, when compared to those found in the related literature. The government can use our method to detect outliers and manage data quality. The general public can also directly interpret the features we constructed, giving them a chance to monitor air quality and protect the environment using their cellphones.

To conclude this paper, we discuss the limitations of our work, which could also direct a number of interesting topics for future study. The firstly limitation is that our photos were taken only in the morning. This was because the camera was fixed on a west-facing window, and light reflections from the window glass may have affected the quality of photos taken in the afternoon. In the future, we may add a polarizing mirror to the camera lens to reduce the reflection. Then we can use more photos taken at more points in time. Secondly, our model only analyzed photos taken at one fixed location; however, in the future, it can be extended to analyze those taken at different locations nearby. Thirdly, we used only three explanatory variables here; in the future, we may consider using additional explanatory variables related to $PM_{2.5}$. For example, meteorological conditions, weather data, and seasonal factors have been proven important in previous studies (Liang et al., 2015), and we may integrate these variables into our model for better accuracy. Lastly, we only estimated $PM_{2.5}$ in this work. Our method has actually provided a promising approach to estimate the concentrations of other pollutants. We believe it could still work as an effective evaluation method in situations where the main pollutant is PM_{10}. Furthermore, this method could provide a promising way to build a system to comprehensively evaluate several air pollutants.

Received 19 June 2018

REFERENCES

ALVARADO, M., F. GONZALEZ, A. FLETCHER, AND A. DOSHI (2015). Towards the development of a low cost airborne sensing system to monitor dust particles after blasting at open-pit mine sites. *Sensors 15*(8), 19667–19687.

BURNEY, J. AND V. RAMANATHAN (2014). Recent climate and air pollution impacts on indian agriculture. *Proceedings of the National Academy of Sciences 111*(46), 16319–16324.

CARVALHO-OLIVEIRA, R., L. F. AMATO-LOURENÇO, T. C. MOREIRA, D. R. R. SILVA, B. D. VIEIRA, T. MAUAD, M. SAIKI, AND P. H. N. SALDIVA (2017). Effectiveness of traffic-related elements in tree bark and pollen abortion rates for assessing air pollution exposure on respiratory mortality rates. *Environment international 99*, 161–169.

CHEN, L.-J., Y.-H. HO, H.-C. LEE, H.-C. WU, H.-M. LIU, H.-H. HSIEH, Y.-T. HUANG, AND S.-C. C. LUNG (2017). An open framework for participatory pm2. 5 monitoring in smart cities. *IEEE Access 5*, 14441–14454.

CHENG, N., D. ZHANG, Y. LI, X. XIE, Z. CHEN, F. MENG, B. GAO, AND B. HE (2017). Spatio-temporal variations of pm 2.5 concentrations and the evaluation of emission reduction measures during two red air pollution alerts in beijing. *Scientific Reports 7*(1), 8220.

CHOWDHURY, S. AND S. DEY (2016). Cause-specific premature death from ambient pm 2.5 exposure in india: Estimate adjusted for baseline mortality. *Environment international 91*, 283–290.

CROSSON, W. L., M. Z. AL-HAMDAN, S. N. HEMMINGS, AND G. M. WADE (2012). A daily merged modis aqua–terra land surface temperature data set for the conterminous united states. *Remote Sensing of Environment 119*, 315–324.

FAN, J. AND J. LV (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911. MR2530322

GAO, Y., W. DONG, K. GUO, X. LIU, Y. CHEN, X. LIU, J. BU, AND C. CHEN (2016). Mosaic: A low-cost mobile sensing system for urban air quality monitoring. *The 35th Annual IEEE International Conference on Computer Communications*, 1–9.

GONZALEZ, R. C. AND R. E. WOODS (2006). *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

GUO, S., M. HU, M. L. ZAMORA, J. PENG, D. SHANG, J. ZHENG, Z. DU, Z. WU, M. SHAO, L. ZENG, ET AL. (2014). Elucidating severe urban haze formation in china. *Proceedings of the National Academy of Sciences 111*(49), 17373–17378.

HAMILTON, E. (2004). Jpeg file interchange format.

He, K., J. Sun, and X. Tang (2011). Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence 33*(12), 2341–2353.

Huang, R.-J., Y. Zhang, C. Bozzetti, K.-F. Ho, J.-J. Cao, Y. Han, K. R. Daellenbach, J. G. Slowik, S. M. Platt, F. Canonaco, et al. (2014). High secondary aerosol contribution to particulate pollution during haze events in china. *Nature 514*(7521), 218–222.

Levy, R., S. Mattoo, L. Munchak, L. Remer, A. Sayer, F. Patadia, and N. Hsu (2013). The collection 6 modis aerosol products over land and ocean. *Atmospheric Measurement Techniques 6*(11), 2989.

Li, Y., J. Huang, and J. Luo (2015). Using user generated online photos to estimate and monitor air pollution in major cities. *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 79.

Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen (2016). Pm2. 5 data reliability, consistency, and air quality assessment in five chinese cities. *Journal of Geophysical Research: Atmospheres 121*(17).

Liang, X., T. Zou, B. Guo, S. Li, H. Zhang, S. Zhang, H. Huang, and S. X. Chen (2015). Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A 471*(2182), 20150257.

Lin, C., Y. Li, Z. Yuan, A. K. Lau, C. Li, and J. C. Fung (2015). Using satellite remote sensing data to estimate the high-resolution distribution of ground-level pm 2.5. *Remote Sensing of Environment 156*, 117–128.

Ma, Z., X. Hu, L. Huang, J. Bi, and Y. Liu (2014). Estimating ground-level pm2. 5 in china using satellite remote sensing. *Environmental science & technology 48*(13), 7436–7444.

Mahowald, N. (2011). Aerosol indirect effect on biogeochemical cycles and climate. *Science 334*(6057), 794–796.

Mao, J., U. Phommasak, S. Watanabe, and H. Shioya (2014). Detecting foggy images and estimating the haze degree factor. *Journal of Computer Science & Systems Biology 7*(6), 1.

MEPC (2015). National data. "http://www.chem17.com/news/Detail/69405.html".

the X.org source code (1989). Current official rgb.txt. https://cgit.freedesktop.org/xorg/app/rgb/tree/rgb.txt.

Wang, J. and S. A. Christopher (2003). Intercomparison between satellite-derived aerosol optical thickness and pm2. 5 mass: implications for air quality studies. *Geophysical research letters 30*(21).

Wang, J., S. Li, H. Li, X. Qian, X. Li, X. Liu, H. Lu, C. Wang, and Y. Sun (2017). Trace metals and magnetic particles in pm2. 5: Magnetic identification and its implications. *Scientific Reports 7*(1), 9865.

Zhang, S., B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen (2017). Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A 473*(2205), 20170457.

Zhao, P. (2016). Planning for social inclusion: The impact of socioeconomic inequities on the informal development of farmland in suburban beijing. *Land Use Policy 57*, 431–443.

Ke Xu
School of Statistics
University of International Business and Economics
Beijing, 100871
P. R. China
E-mail address: xk0566@163.com

Jianqiao Wang
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA 19104
USA
E-mail address: wangjq@upenn.edu

Rui Pan
School of Statistics and Mathematics
Central University of Finance and Economics
Beijing, 100081
P. R. China
E-mail address: panrui_cufe@126.com

Hansheng Wang
Guanghua School of Management
Peking University
Beijing, 100871
P. R. China
E-mail address: hansheng@pku.edu.cn