

Full Length Research Paper

An air quality composite index based on pollutant concentration factor analysis

Kai-guang ZHANG*, Ming-ting BA, Hong-ling MENG and Yan-min SUN

Zhengzhou Normal University, Zhengzhou 450044, China.

Received 4 May 2019; Accepted 9 September 2019

To overcome the problems of Air Quality Index in describing regional short-term air quality status and changing trend, based on the daily average concentrations of six air pollutants affecting regional air quality in Zhengzhou from 2015 to 2018, this paper analyzes their correlation characteristics and defines an air quality composite index instead of the original index by using factor analysis method, which not only extracts all the information of each Individual Air Quality Index, but also eliminates the correlation among them. At last, it analyzes the difference between the two indices in describing regional short-term air quality status and changing trend.

Key words: Air quality index (AQI), individual air quality index (IAQI), correlation coefficient, partial correlation coefficient, information extraction, factor analysis.

INTRODUCTION

The inefficient resource utilization and rapid population agglomeration have greatly promoted the rapid growth of China's economy and the rapid advancement of urbanization to some certain extent in recent China (Zhang et al., 2019a). At the same time, the excessive emissions of atmospheric pollutants, which have great impact on human health, ecology and environment, such as total suspended particulate matter, sulfur dioxide (SO₂), nitrogen oxide (NO_x), volatile organic compounds (VOCs), photochemical oxides and the greenhouse gases have been causing serious ecological problems (especially atmospheric environmental problems) in many urbanized regions.

The statistical results show that the air pollutants in Chinese cities mainly include six pollutant items such as PM_{2.5}, PM₁₀, SO₂, NO₂, CO and O₃ (Liu et al., 2013). In order to monitor and forecast the air pollution level, the short-term air quality status and its change trends in a

region, the People's Republic of China National Environmental Protection Standards (HJ 633-2012) defines the Air Quality Index (AQI), which according to these six pollutant impacts on human health, ecology and environment, simplifying separately their concentration values into a conceptual index value (Individual Air Quality Index, IAQI), which is the maximum value of the conceptual index values.

In recent years, research by geographers into urban air quality has focused on two aspects, one is the change characteristics of air quality, the other research aspect is one that emphasizes influential factors, giving rise to a lot of valuable results (Lin and Wang, 2016).

In fact, there are certain correlations between the six air pollutants in a region. Although the single maximum could describe the overall distribution characteristics of the main pollutant and correlate the partial distribution characteristics with the other pollutants, it largely ignored

*Corresponding author. E-mail: zzgis@zznu.edu.cn or zzgis@sina.com

Table 1. Individual Air Quality Index and corresponding pollutant concentration extreme value.

IAQI	Pollutant concentration interval					
	PM2.5	PM10	SO ₂	NO ₂	CO	O ₃
0	0	0	0	0	0	0
50	35	50	50	40	2	100
100	75	150	150	80	4	160
150	115	250	470	180	14	210
200	150	350	800	280	24	265
300	250	420	1600	565	36	800
400	350	500	2100	750	48	1000
500	500	600	2620	940	60	1200

the impact of non-correlate partial of other pollutants on the air quality, where there are certain limitations in describing the short-term air quality condition and its change trend. Taking Zhengzhou as an example, this paper analyzes the correlation characteristics of the six air pollutants items using the air pollutant monitoring data from 2015 to 2018, which studies the limitations of AQI and proposes a composite index to describe the air quality conditions, calculates the composite index values of the region in the study period, and then compares the differences of the two indices in described air quality conditions.

The definition of AQI

AQI is the maximum of the six Individual Air Quality Indices ($IAQI$), and takes the corresponding pollutant as the primary pollutant.

The calculation steps are as follows:

(1) According to the equivalent impact degrees of the six air pollutants PM2.5, PM10, SO₂, NO₂, CO and O₃ on human health, ecology and environment, the pollutant daily average concentration of PM2.5, PM10, SO₂, NO₂, CO, and 8-h moving average concentration of O₃ ($\mu\text{g}/\text{m}^3$) is divided into seven levels shown in Table 1 (Pan and Li, 2016).

(2) Calculation of Individual Air Quality Index ($IAQI_p$) of pollutant P by using interpolation method, the calculating formula of interpolation method is

$$IAQI_p = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_p - BP_{Lo}) + IAQI_{Lo} \quad (1)$$

Where is the Individual Air Quality Index of single pollutant, C_p is the concentration of pollutant P , BP_{Hi} and BP_{Lo} respectively are the how threshold and low threshold of C_p , $IAQI_{Hi}$ and $IAQI_{Lo}$ respectively are the air quality indices of BP_{Hi} and BP_{Lo} . the air quality index grading standards

are shown in Table 1 (Pan and Li, 2016). The $IAQI_p$ of pollutant P describes the influence degree of pollutant P on the short-term air quality status and its change trends.

(3) Calculation of AQI using

$$AQI = \max\{IAQI_{PM2.5}, IAQI_{PM10}, IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{CO}, IAQI_{O_3}\} \quad (2)$$

and takes the corresponding pollutant as the primary pollutant.

CORRELATION ANALYSIS OF POLLUTANT ITEM

Correlation analysis

The correlation coefficient is a statistical indicator used to describe the relationships between two variables by using the sum of the product of normalized deviations of the two variables (Xia et al., 2014). For the two variables X and Y , (x_i, y_i) $i = 1, 2, 3 \dots n$ are their observed values, and the correlation coefficient of the two variables is defined as

$$r_{XY} = \left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right) / \sqrt{D(X)D(Y)}, \quad (3)$$

where \bar{x} and \bar{y} , $D(X)$ and $D(Y)$ are respectively their means and variances. r_{XY} belongs to $[-1, 1]$ using the statistic $t = r\sqrt{(n-2)/(1-r^2)} \sim t(n-2)$ as the test function of significant level, under the confidence degree of 0.05, the greater $|r|$ means the stronger correlation between X and Y , $|r|=1$ means X and Y is linear correlation; also, $r=0$ means X and Y is independence (Cao et al., 2014; Lunev and Neknitkin, 2019; Zhang et al., 2019b).

In order to study the correlation between pollutant items, taking Zhengzhou as an example, Formula 3 is used in calculating the correlation coefficient between the IAQIs of the six pollutant items in the study period. The results are shown in Table 2.

Table 2. The correlation coefficients of the pollutant items in Zhengzhou.

Correlation	PM2.5	PM10	SO ₂	NO ₂	CO	O ₃
PM2.5	1.000	0.867	0.542	0.618	0.844	-0.328
PM10	0.867	1.000	0.584	0.639	0.741	-0.210
SO ₂	0.542	0.584	1.000	0.621	0.571	-0.365
NO ₂	0.618	0.639	0.621	1.000	0.573	-0.297
CO	0.844	0.741	0.571	0.573	1.000	-0.384
O ₃	-0.328	-0.210	-0.365	-0.297	-0.384	1.000

Table 3. The partial correlation coefficients of the pollutant items in Zhengzhou.

Correlation	PM2.5	PM10	SO ₂	NO ₂	CO	O _{3_8 h}
PM2.5	1.000	0.643	-0.141	0.095	0.561	-0.136
PM10	0.643	1.000	0.229	0.178	***	0.242
SO ₂	-0.141	0.229	1.000	0.351	0.184	-0.211
NO ₂	0.095	0.178	0.351	1.000	***	***
CO	0.561	***	0.184	***	1.000	-0.159
O ₃	-0.136	0.242	-0.211	-0.069	***	1.000

***Partial correlation coefficients is non-significant.

Under the confidence degree of 0.01 (both sides), there are significant correlations among them. The positive correlations appear among PM2.5, PM10, SO₂, NO₂ and CO, with all correlation coefficients greater than 0.500, indicating that the increase of one pollutant concentration in the region would inevitably lead to the increase of the other four pollutant concentrations, and controlling the emission of one pollutant item would reduce the atmosphere pollution by the remaining four pollutant items. The negative correlations appear between O₃ and PM2.5, PM10, SO₂, NO₂, and CO showing the increase of the concentration of PM2.5, PM10, SO₂, NO₂ and CO would inevitably cause the decrease of the concentration of O₃ and vice versa.

Partial correlation analysis

The correlation coefficient reflects the degree of dependence between pollutant items as a whole. To quantitatively describe the dependence degree between two pollutant items, the partial correlation coefficient between two pollutant items is defined as:

$$Pr_{ij} = (R_{i(0 < k \leq 6; k \neq i)}^2 - R_{i(0 < k \leq 6; k \neq i, j)}^2) / (1 - R_{i(0 < k \leq 6; k \neq i, j)}^2), \quad (4)$$

where $R_{i(0 < k \leq 6; k \neq i)}^2$ is the proportion of the linear regression part of the other five pollutant items in the variance of pollutant item i , in other words, it is the proportion of the linearly explained part by the other five pollutant items. $R_{i(0 < k \leq 6; k \neq i, j)}^2$ is the proportion of the linear regression part of the other pollutant items except j in the variance of

pollutant item i . Pr_{ij} is the influence of the pollutant concentration of j on the pollutant concentration of i when controlling influence of the other four pollutant concentrations. The value of the partial correlation coefficient ranges from +1 to -1. Using the statistic $t = (Pr_{ij} \sqrt{(n-q-2)}) / (1 - Pr_{ij}^2) \sim t(n-q-2)$ as the test function of significant level. Under the confidence degree of 0.05, the greater Pr_{ij} means the higher influence of j on i . If the value is equal to zero, there is no relationship between the variables, if the value is positive, this means that the two variables have changed, if the value is negative, then the two variables have changed in opposite directions (Cao et al., 2014; Lunev et al., 2019).

Taking Zhengzhou as an example, Formula 4 was used to calculate the partial correlation coefficient between IAQI of pollutant items. The results are shown in Table 3.

Under the confidence degree of 0.01 (both sides), there is non-significant correlation between the pollutant items PM10 and CO, NO₂ and CO, and CO and O₃. The strongest partial correlations appear between PM2.5 and PM10 and PM2.5 and CO, with their partial correlation coefficients greater than 0.500. There is a negative partial correlation between PM2.5 and SO₂.

PROBLEMS ANALYSIS IN AQI

Analysis of problems in calculating of AQI

AQI is the maximum of the six IAQIs, and the corresponding pollutant item is the primary pollutant.

Table 4. Days distribution of air quality grade in Zhengzhou (2015-2018).

Correlation	PM2.5	PM10	SO ₂	NO ₂	CO	O ₃ _8 h
PM2.5	446	8	0	1	0	1
PM10	8	300	0	6	0	4
SO ₂	0	0	0	0	0	0
NO ₂	1	6	0	111	0	0
CO	0	0	0	0	1	0
O ₃	1	4	0	0	0	356

Using Formula 1 to calculate the AQIs of the six pollutant items, based on the daily average concentration values of the six air pollutants in Zhengzhou from 2015 to 2018 (where O₃ is the 8-h average concentration value), the day distributions of primary pollutants is shown in Table 4. SO₂ has never been the primary pollutant. Two of the six pollutant items in 20 days are the primary pollutants, that is, the corresponding concentration values of the two pollutant items simultaneously reaches the extreme high-values of the same air quality level, their IAQIs are equal, but air quality description of AQI only uses one of them as the primary pollutant, description value is only equal to the IAQI, and the air quality status is obviously worse than only one reaching the extremes high-values of the air quality level.

When the primary pollutant item remains unchanged, AQI only shows the change pattern of the primary pollutant concentration, which ignores the change characteristics of non-primary pollutant items, especially when the concentration of one or more non-primary pollutant items increases significantly, but does not reach the extremes high-values of the air quality level. In this case, the real air quality situation changes obviously, but this change could not be reflected by AQI. For example, in the lightly polluted weather 2015.10.14 and 2016.3.20, their pollutant concentrations, respectively are (89, 185, 54, 81, 1.3, 125) and (7, 185, 40, 54, 1.9, 96), corresponding individual air quality indices, respectively are (118, 118, 52, 101, 33, 71) and (78, 118, 40, 68, 48, 48), AQIs are all 118, the primary pollutant items are all PM10, but the air quality status in the former is obviously better than that in the latter.

When the primary pollutant item changes, the original primary pollutant item concentration decreases significantly, and the other one or more pollutant item concentration increases exactly to the corresponding concentration value in the original air quality level, the real air quality composition changes significantly, but the AQI remains unchanged.

Analysis of AQI information extraction

When the primary pollutant item is determined, AQI is the IAQI. Due to the correlation among pollutant items, the

IAQI of the primary pollutant item also contains the information of the IAQIs of other pollutant items, described by the partial correlation coefficient between the two IAQIs.

For example, the primary pollutant item is PM2.5, the partial correlation coefficients between PM2.5 and the other five pollutant items, respectively are 0.643, -0.141, 0.095, 0.561 and -0.136, implying that $AQI = IAQI_{PM2.5}$ respectively contains the information of other five IAQIs 64.3, 14.1, 9.5, 56.1 and 13.6% information. The un-described information in the other five IAQIs, respectively are 35.7, 85.9, 90.5, 45.9, and 86.4%.

If the amount of information contained in each IAQI is regarded as one overall unit, the amount of information described by AQI is 42.9%. Similarly, when the other pollutant items are as the primary pollutant item, the amount of information described by AQI, respectively are 38.3, 35.3, 28.4, 31.9 and 30.28%. Combining with Table 4, the amount of information extracted by AQI from the six IAQIs in the study period is only 36.8%.

In general, from the monitoring data of Zhengzhou, the continuity of AQI with time is poor, and the average daily change rate (the slope of AQI) is 36.52. Less information extraction means AQI does not fully describe the change pattern of air quality, meanwhile the application of maximum conceals the other IAQIs and their change pattern.

AIR QUALITY COMPOSITE INDEX

In order to compensate for the shortcomings of AQI in describing the overall air quality of the region, the Air Quality Composite Index is defined as the weighted average of IAQIs.

$$AQCI = a_1 IAQI_{PM2.5} + a_2 IAQI_{PM10} + a_3 IAQI_{SO_2} + a_4 IAQI_{NO_2} + a_5 IAQI_{CO} + a_6 IAQI_{O_3} \quad (5)$$

where $a = (a_1, a_2, a_3, a_4, a_5, a_6)$ is the weighted average of the factor score vector in the factor analysis with the factor contribution rate as weights, which reflects all the information of the original six IAQIs, and could effectively eliminate the correlation among them, describing the air

Table 5. Air quality factors and factor contribution rates in Zhengzhou (2015-2018).

Variable	β_1	β_2	β_3	β_4	β_5	β_6
β_{1i}	0.449	0.074	-0.218	-0.297	-0.875	-2.329
β_{2i}	0.405	0.059	-0.207	-0.275	-0.780	1.964
β_{3i}	-0.051	0.098	1.345	-0.284	-0.086	-0.283
β_{4i}	-0.031	0.039	-0.240	1.408	0.123	-0.058
β_{5i}	0.279	0.071	-0.163	-0.173	1.814	0.613
β_{6i}	0.063	1.074	0.152	0.047	0.240	-0.233
V_i	55.728	16.590	11.840	10.102	3.903	1.837

quality status as a whole. The calculation principles and steps are shown in the following.

Assume

$$X_i = (IAQI_{PM2.5,i}, IAQI_{PM10,i}, IAQI_{SO_2,i}, IAQI_{NO_2,i}, IAQI_{CO,i}, IAQI_{O_3,i}), i = 1, 2, 3, \dots, n$$

is observed values of $X = (IAQI_{PM2.5}, IAQI_{PM10}, IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{CO}, IAQI_{O_3})$

, using the linear combination $Z_i, i = 1, 2, \dots, 6$, as common factors to substitute the six IAQIs. The common factors are mutually orthogonal, and their coefficient vectors are unit vector,

$$(IAQI_{PM2.5}, IAQI_{PM10}, IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{CO}, IAQI_{O_3})^T = (a_{ij})_{6 \times 6} (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)^T$$

$$(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)^T = (a_{ij})_{6 \times 6}^{-T} (IAQI_{PM2.5}, IAQI_{PM10}, IAQI_{SO_2}, IAQI_{NO_2}, IAQI_{CO}, IAQI_{O_3})^T$$

According to principal component analysis theory, $Var(Z_i) = \alpha_i^T \Sigma \alpha_i = \lambda_i$, $Cov(Z_i, Z_j) = 0, (i \neq j)$, Σ is the correlation coefficient matrix of the six IAQIs, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_6$ are the eigenvalues of Σ , $\alpha_1, \alpha_2, \dots, \alpha_6$ are corresponding eigenvectors. $\sum_{i=1, \dots, 6} \lambda_i$ is equal to the total information of the six IAQIs (the variance sum of the six IAQIs), λ_i is the information Z_i contains, while the percentage of total information is

$$V_i = \frac{\lambda_i}{\sum_{i=1, \dots, 6} \lambda_i} \times 100\% \quad (6)$$

which effectively rotates the common factors $\alpha_1, \alpha_2, \dots, \alpha_6$ to $\beta_1, \beta_2, \dots, \beta_6$, makes the common factor load coefficients closer to 1 or 0, and then calculates the information contribution rate of the new factor V_i .

The Air Quality Composite Index is defined as:

$$AQCI = \sum_{i=1, 2, \dots, 6} V_i \beta_i = \sum \beta_{1i} V_i IAQI_{PM2.5} + \sum \beta_{2i} V_i IAQI_{PM10} + \sum \beta_{3i} V_i IAQI_{SO_2} + \sum \beta_{4i} V_i IAQI_{NO_2} + \sum \beta_{5i} V_i IAQI_{CO} + \sum \beta_{6i} V_i IAQI_{O_3} \quad (7)$$

Air quality grade

In order to grade and release the air quality, the AQCI values corresponding to each level of AQI is calculated by referring to the classification standard of AQI.

Assume T_k is the IAQI maximum value of k level, $(T_{IAQI_{PM2.5}}, T_{IAQI_{PM10}}, T_{IAQI_{SO_2}}, T_{IAQI_{NO_2}}, T_{IAQI_{CO}}, T_{IAQI_{O_3}})$ is the mean value of the observation values of k level, and the upper limit of AQCI of k level is defined as:

$$CAQI_k = \min\{FAQI(x_1, x_2, \dots, x_6 \mid x_i = T_k, x_j = T_j, i \neq j)\} \quad (8)$$

If $AQCI \in (AQCI_{k-1}, AQCI_k]$, then the air quality of the day is called k level.

Air quality analysis of Zhengzhou

Performing factor analysis on the six IAQIs in the study periods of Zhengzhou, the results β_i after α_i by maximum quadratic rotation and the information contribution rate V_i of β_i are shown in Table 5.

In descending order of the information contribution rates, the six common factors, respectively describe the characteristics of PM10 and PM2.5, O₃, SO₂, NO₂, CO, and conjugate feature of PM10 and PM2.5. Overall, the first common factor presents that as the concentration of PM2.5 is higher, the corresponding concentrations of PM10 and CO also increase, but the concentrations of NO₂ and SO₂ decrease sequentially. The second common factor shows that the concentrations of the six pollutant items have the synchronous characteristics. When the concentration of PM2.5 is lower, the concentrations of NO₂ and SO₂ decrease sequentially as described by the sixth common factor. The empirical formula of AQCI is

$$FAQI = 0.130IAQI_{PM2.5} + 0.189IAQI_{PM10} + 0.110IAQI_{SO_2} + 0.107IAQI_{NO_2} + 0.212IAQI_{CO} + 0.241IAQI_{O_3} \quad (9)$$

Table 6. The grading values and the grading results of air quality composite index.

Air quality grade	I	II	III	IV	V	VI
Air quality level	Good	Moderate	Lightly polluted	Moderately polluted	Heavily polluted	Severely polluted
Air quality index	0-50	51-100	101-150	151-200	201-300	301-500
Composite index	0-29.95	29.96-51.48	51.49-72.97	72.98-86.70	86.71-100.49	100.49-151.43
Air quality index days	27	528	406	137	90	27
Composite index days	19	413	555	119	62	47

Table 7. The transfer matrix of AQI grading and AQCI grading.

Air quality grade	I	II	III	IV	V	VI
I	15	12				
II	4	373	151			
III		28	368	10		
IV			33	79	25	
V			3	30	37	20
VI						27

Using Formula 7 to calculate the upper limit of AQCI of each levels, the results are shown in Table 6. The last two lines in Table 6 are the graded days in the study periods of Zhengzhou. Table 7 is the transfer matrix of the graded days between AQI grading and AQCI grading.

Due to the comprehensive consideration of the impact of the six pollutant concentrations on air quality, the day distributions of air quality are significant difference between AQI grading and AQCI grading. There are 79.1% of the days in the study period, and the description level of AQI and AQCI is the same. The IAQI values are similar, slightly less than the grading maximum value, and the air quality of 17.94% days in the study period are seriously underestimated, in which the overall air pollution levels should be high. The air quality of 2.96% days in the study period is overestimated, in which the concentration of one or more pollutant suddenly increase, whereas the other pollutant concentrations maintain its original state (even decrease).

Conclusions

Based on the daily average concentrations of six atmospheric pollutants from 2015 to 2018, the correlation characteristics of pollutant concentrations are discussed in Zhengzhou using correlation analysis and partial correlation analysis methods. The results show that there are significant positive correlations between PM_{2.5} and PM₁₀, PM_{2.5} and CO, PM₁₀ and O₃, and SO₂ and NO₂, certain negative correlations between O₃ and PM_{2.5}, and O₃ and SO₂. For a low amount of information extraction from IAQIs, AQI could not fully describe short-term air quality status and change trend in the region.

Using the factor analysis method, the paper defines an air quality composite index by using a linear combination of six IAQIs, which not only extracts all the information of each IAQI, but also eliminates the correlation among them. Zhengzhou case analysis shows that there are 17.94% days in which the air qualities are seriously underestimated by AQI, and 2.96% days in which the air qualities are overestimated by AQI.

AQCI, based on the correlation of the daily average concentration of six atmospheric pollutants in a region, is an empirical formula obtained by the factor analysis method. For the different correlations of the six atmospheric pollutants in different regions, AQCI empirical formula and grading values are different in different regions.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES

- Cao R, Jiang W, Yuan L, Wang W, Chen Z (2014). Inter-annual variations in vegetation and their response to climatic factors in the upper catchments of the Yellow River from 2000 to 2010. *Journal of Geographical Sciences* 24(6):963-979. <http://dx.doi.org/10.1007/s11442-014-1131-1>.
- Lin X, Wang D (2016). Spatiotemporal evolution of urban air quality and socioeconomic driving forces in China. *Journal of Geographical Sciences* 26(11):1533-1549. <http://dx.doi.org/10.1007/s11442-016-1342-8>.
- Liu X, Zhong Y, He Q, Yang X, Huo X, Ali M, Huo W (2013). Vertical distribution characteristics of dust aerosol mass concentration in the Taklimakan Desert hinterland. *Sciences in Cold and Arid Regions* 5(6): 0745-0754. <http://dx.doi.org/10.3724/SP.J.1226.2013.00745>.

- Pan B, Li L (2016). Comparison of the calculating method and classifying program of Air Quality Index among some countries. *Environmental Monitoring in China* 32(1):13-17. <http://dx.doi.org/10.19316/j.issn.1002-6002.2016.01.003>.
- Lunev S, Nekritkin VV (2019). A remark on certain classic criteria of mathematical statistics. *Vestnik St. Petersburg University Mathematics* 52(2):154-161.
- Xia K, Wang B, Li L, Shen S, Huang W, Xu S, Dong L, Liu L(2014). Evaluation of snow depth and snow cover fraction simulated by two versions of the flexible global ocean–atmosphere–land system model. *Advances in Atmospheric Sciences* 31(2):407-420. <http://dx.doi.org/10.1007/s00376-013-3026-y>.
- Zhang K, Ba M, Meng H, Sun Y (2019a). Spatial distribution characteristics and evolution pattern of air quality in Henan Province. *Journal of Progressive Research in Mathematics* 15(1):2464-2468.
- Zhang K, Ba M, Meng H, Sun Y (2019b). Spatial Correlation Analysis of Urban Air Quality in Henan Province, *SCIREA Journal of Geosciences*,3(1):1-12.