第 28 卷第 1 期　　四川理工学院学报(自然科学版)　　Vol. 28　No. 1
2015 年 2 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Feb. 2015
文章编号:1673-1549(2015)01-0071-05　　DOI:10.11863/j.suse.2015.01.17

基于偏最小二乘法的 PM2.5 相关因素分析研究

卢　鹏[1], 何　杰[2], 彭丛笑[3]

(1. 西南交通大学峨眉校区基础课部, 四川　峨眉　614202;2. 西南交通大学土木工程学院, 成都　610031;
3. 四川理工学院建筑工程学院, 四川　自贡　643000)

摘　要:利用偏最小二乘法,主要分析了 PM2.5(含量)与 SO_2、NO_2 和 CO 等指标的关联度以及具体的关系式,并对距离分析和典型分析得到的结果进行了对比分析。最后在结果的基础上,分析了 PM2.5(含量)与 SO_2、NO_2 和 CO 等指标的具体函数表达式,为如何更好的控制、治理该污染物提供了依据。

关键词:距离分析;典型分析;偏最小二乘分析

中图分类号:O213　　**文献标志码**:A

引　言

细颗粒物已经被列为了影响我国各城市空气质量的主要大气污染物之一,其不仅影响气候、城市能见度,同时对人体的健康有巨大的影响,这主要是因为细颗粒物(PM2.5)能够被人吸入呼吸系统,甚至能穿透肺细胞而进入血液循环,最终对人体健康造成影响。鉴于此,我国已经将细颗粒物(PM2.5)作为了首要污染物,对其做深入的研究有利于制定有效的控制治理方案。PM2.5 的主要来源有两个方面[1],即自然源与人为源,且主要成分包括水溶性离子、颗粒有机物和微量元素等。有相关学术研究[2-4]认为:AQI 监测指标中的 SO_2、NO_2 和 CO 在一定的条件下能通过化学反应生成 PM2.5。基于此,本文以西安市各地区所采集的数据[5]进行分析,分析的主要内容包括 PM2.5(含量)与 SO_2、NO_2 和 CO 等指标的关联度以及具体的关系式。

1　分析方法理论研究

为了分析 PM2.5(含量)与 SO_2、NO_2 和 CO 等指标之间的相关程度和关系,本文首先采用距离分析法进行分析各指标之间的相关性与独立性。为进一步深层次研究各指标之间的相关性,在考虑两组变量相关性时,同时考虑其他变量的影响,于是采用典型相关分析方法进行分析,得到更加合理的相关性关系。考虑到偏最小二乘回归能够提供一种多对多线性回归模型的方法,该方法集中了主成分分析、典型相关性分析和线性回归分析方法的特点,不仅能提供一个更为合理的回归模型,同时还能够完成一些类似于主成分分析和典型相关分析的研究内容,提供一些更丰富、深入的信息。所以本文采用偏最小二乘回归分析方法建立 PM2.5 与其他各指标的关系模型,利用 MATLAB 等数学工具就可以获得偏最小二乘回归分析模型中的各参数值,然后对计算结果进行检验。

1.1　距离分析

采用 Person 相关系数统一的表征相关程度[6],两组变量 X 和 Y 的 Person 相关系数计算:

$$\rho = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\overline{X})^2\sum_{i=1}^{N}(Y_i-\overline{Y})^2}} \tag{1}$$

式中,$\overline{X}$ 和 $\overline{Y}$ 分别为 X_i 和 Y_i 的平均值,N 表示 (X_i,Y_i)

收稿日期:2014-11-13
基金项目:中央高校基本科研业务费专项资金(2682014BR039)
作者简介:卢 鹏(1983-),男,四川贡贡人,讲师,主要从事数学建模理论与方法,粗糙集理论与应用方面的研究,(E-mail)1983lupeng@163.com

数对的数目。

1.2 深入分析

在距离相关分析的基础上,考虑到采用距离相关分析[7]仅能得到两组变量之间简单的相关系数,这样的结果不能抓住问题的本质,如果能够采用类似于主成分的思想,分别找出两组变量的各自的某个线性组合,讨论线性组合之间的相关关系,会使结果更加符合实际情况。基于此,本文采用典型相关分析对 6 个指标做进一步的相关与独立分析,这种方法更加便捷且能突显问题的本质。

首先研究任意两组指标随机变量之间的相关关系,第一组指标 X 共 5 个(包含:SO_2、NO_2、可吸入颗粒物、CO、O_3),第二组指标为 Y(包含:PM2.5),可用复相关系数。其思想是先将每一组指标随机变量作线性组合,成为两个随机变量,

$$\begin{cases} u = c^T X = \sum_{i=1}^{P} c_i x_i \\ v = \gamma^T Y = \sum_{j=1}^{q} \gamma_j y_j \end{cases} \tag{2}$$

式中,P 表示 5 个指标,q 表示 1 个指标。

由于 u,v 与投影向量 c,γ 有关,所以相关系数矩阵 r_{yu} 与 c,γ 有关,$r_{yu} = r_{yu}(c,\gamma)$。取在 $c^T \sum_{xx} c = 1$,$\gamma^T \sum_{YY} \gamma = 1$ 的条件下使 r_{yu} 达到最大的 c,γ 作为投影向量,得到的相关系数为复相关系数,

$$r_{yu} = \max_{c^T \sum_{xx} C=1} r_{yu}(c,\gamma)$$

将两组变量的协方差矩阵分块得:

$$Cov\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(Y,X) & Var(Y) \end{bmatrix} = \begin{bmatrix} \sum_{XX} & \sum_{XY} \\ \sum_{YX} & \sum_{YY} \end{bmatrix}$$

此时

$$\rho = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2 \sum_{i=1}^{N} (Y_i - \overline{Y})^2}} \tag{3}$$

典型相关系数计算结果检验公式参考文献[8]。

1.3 偏最小二乘回归分析

由分析结果可知,PM2.5 与其他 4 个指标具有较强的相关性,所以采用偏最小二乘法[9]建立 PM2.5 与其他 4 个指标(SO_2、NO_2、可吸入颗粒物、CO,将指标编号为 1-4)之间的关系模型。

用 x_i^m(i 表示时间,m 表示指标编号)表示实测的 AQI[10]监测指标浓度值;y_i 表示实测 PM2.5 浓度值。4 个指标浓度的数据阵记为 $A = (a_{ij})_{238\times4}$,实测 PM2.5 浓度的数据矩阵记为 $B = (b_{ij})_{238\times1}$,即为:

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,m} \end{bmatrix}, B = \begin{bmatrix} b_{1,1} & \cdots & b_{1,p} \\ \vdots & & \vdots \\ b_{n,1} & & b_{n,p} \end{bmatrix}$$

具体的求解流程:

(1)分别提取两变量组的第一对成分,并使之相关性达到最大。

(2)建立 $\gamma_1,\cdots,\gamma_p$ 对 u_1 的回归及 $x_1,\cdots,x_m$ 对 u_1 的回归。

(3)用残差阵 A_1 和 B_1 代替 A 和 B,重复以上步骤。

(4)设 $n \times m$ 数据阵 A 的秩为 $r \leqslant \min(n-1,m)$,则存在个成分 $u_1,u_2,\cdots,u_r$,使得

$$\begin{cases} A = \sum_{i=1}^{r} \hat{u}_i \sigma^{(i)T} + A_r \\ B = \sum_{i=1}^{r} \hat{u}_i \tau^{(i)T} + B_r \end{cases} \tag{4}$$

(5) p 个因变量的偏最小二乘回归方程式为

$$\gamma_j = c_{j1}x_1 + \cdots + c_{jm}x_m \quad j = 1,2,\cdots,p \tag{5}$$

2 基于相关分析的实验结果

2.1 距离分析实验结果

将西安市各地区采集的数据,经过处理后代入(1)式进行计算,得到 6 个指标的相关系数矩阵(表 1)。

表 1 指标的相关性系数表

	SO_2	NO_2	可吸入颗粒物	CO	O_3	PM2.5
SO_2	1	0.807	0.678	0.659	-0.179	0.726
NO_2	0.807	1	0.727	0.626	-0.063	0.734
可吸入颗粒物	0.678	0.727	1	0.586	-0.069	0.779
CO	0.659	0.626	0.586	1	-0.381	0.822
O_3	-0.179	-0.063	-0.069	-0.381	1	-0.352
PM2.5	0.726	0.734	0.779	0.822	-0.352	1

由表 1 可知,PM2.5 与其他 5 个指标之间具有较强的相关性,除 O_3 是负相关,其他各指标对 PM2.5 均为正相关,且相关系数均大于 0.7,这说明 PM2.5 浓度变化与其他 5 个指标密切相关。

分析 O_3 与其他指标的相关系数可以发现,O_3 与其

他指标的相关性较弱,且大部分是负相关,说明其他指标对 O_3 浓度的变化影响不大。

2.2 典型分析实验结果

将处理后的数据代入编好的程序式进行计算,可以得到 6 个指标的典型相关系数及检验表(表 2)。

表 2 典型相关系数

序　号	1	2
典型相关系数	0.9992	0.9769

由表 2 可知,2 个典型相关系数均较高,表明 PM2.5 与其他 5 个指标之间密切相关。但要确定典型变量相关性的显著程度,尚需要进行相关系数 χ^2 统计量检验[11],具体做法是:比较统计量 χ^2 计算值与临界值的大小,据比较结果判定典型变量相关性的显著程度,结果见表 3。

表 3 相关系数检验表

序号	自由度	χ^2 计算值	χ^2 临界值(显著水平 0.05)
1	10	72.845	4.1608×10^{-7}
2	4	42.984	2.5801×10^{-4}

从表 3 知这两对典型变量均值通过了 χ^2 统计量检验,表明相应典型变量之间相关关系显著,能够用其他 5 个指标来分析 PM2.5 的变换。因此表 4 的第一组相关性系数是可靠的。

表 4 结构分析(相关系数)

序号	指标名称	SO_2	NO_2	可吸入颗粒物	CO	O_3
1	PM2.5	0.6350	0.8177	0.9966	0.8534	0.0201
2	PM2.5	0.6250	0.7988	0.9737	0.8337	0.0196

表 5 给出了两种分析方法的计算结果,可以看出两种分析方法分析结果较为一致,典型相关性分析表明:可吸入颗粒物与 PM2.5 密切相关,相关性达到 0.9966,O_3 与 PM2.5 不相关。

表 5 两种相关性分析结果对比表

方法名称	指标名称	SO_2	NO_2	可吸入颗粒物	CO	O_3
典型相关性分析	PM2.5	0.6350	0.8177	0.9966	0.8534	0.0201
距离相关性分析	PM2.5	0.726	0.734	0.779	0.822	−0.352

典型相关分析考虑了更多的成分影响,典型相关性分析比简单的距离相关性分析更适合于研究 PM2.5 与其他 5 个指标的相关性。

3 基于偏最小二乘回归分析法的实验结果

3.1 实验结果

将标准化后的数据代入编写好的偏最小二乘回归程序[12],得到的实验结果包括 PM2.5 与 4 个指标之间的相关系数矩阵(表 6)、回归方程和回归系数直方图(图 1)。

PM2.5 与 4 个指标之间的偏最小二乘回归方程:

$$y = -51.3374 + 0.2820x_1 + 0.2973x_2 + 0.8271x_3 + 2.6519x_4$$

表 6 相关系数矩阵

	PM2.5	SO_2	NO_2	可吸入颗粒物	CO
PM2.5 y	1	0.8066	0.6804	0.6587	0.726
SO_2 x_1	0.8066	1	0.7289	0.6263	0.7342
NO_2 x_2	0.6804	0.7289	1	0.5868	0.7791
可吸入颗粒物 x_3	0.6587	0.6263	0.5868	1	0.8224
CO x_4	0.726	0.7342	0.7791	0.8224	1

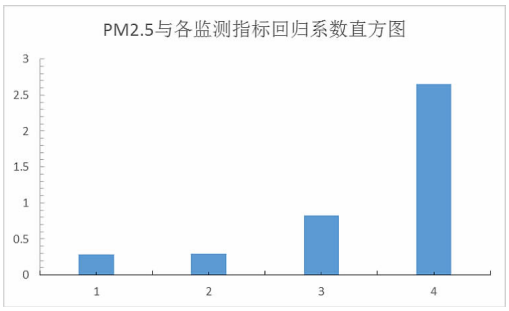

图 1 回归系数直方图

3.2 实验结果的分析及验证

根据偏最小二乘法回归模型的求解及回归系数图(图 1)可以观察到,可吸入颗粒物和 CO 指标对 PM2.5 浓度指标存在较大的正相关性。即它在空气中的含量成分越多,PM2.5 含量也就越多。SO_2 和 NO_2 对 PM2.5 存在较小的正相关。

为了考察偏最小二乘法回归方程的模型精度[13],以 $(\hat{y}_i, y_i)$ 为坐标值,对所有的样本点绘制预测图。$\hat{y}_i$ 是 PM2.5 指标在第 i 个样本点(y_i) 的预测值。在预测图上,如果所有点都能在图的对角线附近均匀分布,则方程的拟合值与原值差异很小,这个方程的拟合效果就令人满意。图 2 为 PM2.5 浓度预测图,图 3 为 PM2.5 实测与预测值析线图,图 4 为 PM2.5 实测与预测值百分比分析图。

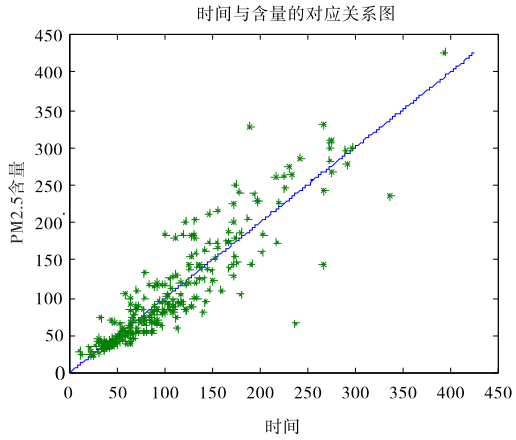

图 2 PM2.5 浓度预测图

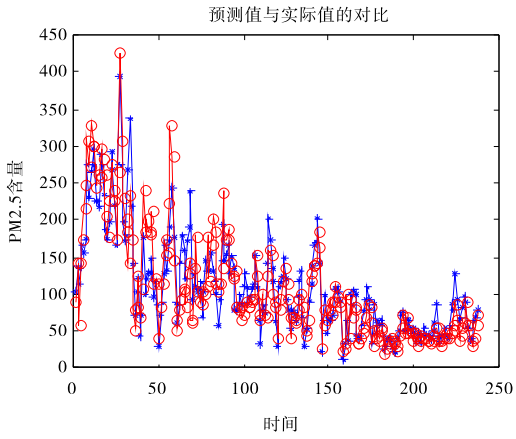

图 3 PM2.5 实测与预测值折线图

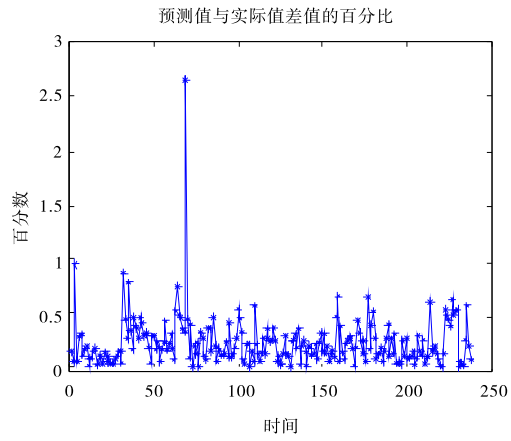

图 4 实测与预测值百分比分析图

由图 2 可知,所有点都在图的对角线附近均匀分布,由图 3 和图 4 可知拟合值与原值差异很小,这些方程的拟合效果令人满意。故偏最小二乘法回归分析 PM2.5 污染物浓度的效果较好。

建立的 PM2.5 与 SO_2、NO_2、可吸入颗粒物和 CO 四种指标的数学模型,能够很好的反映 PM2.5 与各指标的相关关系。

4 结束语

本文利用两种相关分析方法,分析了 PM2.5 含量与 SO_2、NO_2、可吸入颗粒物、CO 以及 O_3 含量之间的相关性,并对比分析了这两种方法的结果,最终确定了 PM2.5 与这 5 个指标之间的相关性。

在此基础上,进一步分析了 PM2.5 与这些指标之间的具体关系,简单的回归分析无法体现 PM2.5 与多指标间的相互依赖关系,并且回归分析的结果较差,不能反映 PM2.5 与其他监测指标间的关系。因此,采用了偏最小二乘回归分析法,该方法能够提供一种多对多线性回归模型的方法,且在模型建立过程中集中了主成分分析、典型相关性分析和线性回归分析的方法和特点,因此在分析结果中,除了可以提供一个更为合理的回归模型外,还可以同时完成一些类似于主成分分析和典型相关分析的研究内容,比纯粹的运用灰色关联度分析[14-15]得到的结果更为可信,同时也提供一些更丰富、深入的信息。最后通过将实际值与预测值进行对比,检验了该关系式具有一定的可行性。

参 考 文 献:

[1] 王帅,杜丽等.国内外环境空气质量指数分析和比较[J].中国环境监测,2013,29(6):58-65.

[2] 卢鹏,何杰.PM2.5 的时间分布与演变扩散研究[J].西南民族大学学报:自然科学版,2014,40(1):66-71.

[3] 郑永杰,刘佳,田景芝.齐齐哈尔市大气细粒子 PM2.5 单颗粒研究[J].安全与环境学报,2014,14(1):273-277.

[4] 皮帅帅,程金平.上海市霾与非霾期间 PM2.5 中水溶性阳离子污染特征对比[J].上海交通大学学报:农业科学版,2014,32(3):27-32.

[5] 李勇,宋慧.西安市空气 PM2.5 问题研究[J].黑龙江大学自然科学学报,2014,31(2):233-237.

[6] 司守奎,孙玺清.数学建模算法与应用[M].北京:国防工业出版社, 2012.5.

[7] 韩忠庚. 数学建模方法及其应用(第二版)[M]. 北京:高等教育出版社,2009.

[8] 张文彤.SPSS 统计分析高级教程[M].北京:高等教育出版社,2013.

[9] 姜启源,谢金星,叶俊.数学模型[M].4 版.北京:高等教育出版社,2011.

[10] 白爱民.AQI vs API—新老空气质量标准之对比[J].环境工程学报,2013,32(6):95-97.

[11] 盛骤.概率论与数理统计[M].4 版.北京:高等教育出版社,2008.

[12] 王桂增,叶昊.主元分析与偏最小二乘法[M].北京:清华大学出版,2012.

[13] 欧阳俊强.长春市环保局大气污染模拟系统的设计与实现[D].吉林:吉林大学,2013.

[14] 周颖璇.基于灰色关联度分析法的 PM2.5 影响因素分析[J].管理观察,2014,15(5):14-16.

[15] 毛毳,孙宇.空气中 PM2.5 浓度的灰色预测与关联因素分析[J].宁夏大学学报:自然科学版,2014,35(3):284-288.

Analysis and Research on Correlative Factors of PM2.5 Based on Partial Least Square Method

LU Peng[1], *HE Jie*[2], *PENG Congxiao*[3]

(1. Emei Campus, Southwest Jiaotong University, Emei 614202, China; 2. School of Civil Engineering, Southwest Jiaotong University, Chengdu, 610031, China; 3. School of Architecture and Engineering, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: by using partial least squares method, the relevancy of PM2.5 (content) and sulfur dioxide (SO_2), nitrogen dioxide (NO_2), correlation monoxide (CO) and other indicators as well as the specific relationships is mainly analyzed, and the results obtained by distance analysis and typical analysis are compared and analyzed. Finally, based on the results, specific function expressions of PM2.5 (content) and sulfur dioxide (SO_2), nitrogen dioxide (NO_2), monoxide (CO) and other indicators are analyzed, which provides a basis for that how to better control and govern the pollutants.

Key words: distance analysis; typical analysis; partial least squares analysis