

基于R语言的城市PM2.5影响因素分析

解 蕾¹, 狄光智²

(1.运城师范高等专科学校数计系, 山西 运城 044000;

2.西南林业大学计算机与信息学院, 云南 昆明 650224)

摘 要:以R语言为数据分析的工具,基于相关分析和回归分析方法,对太原市PM2.5的影响因素进行分析。研究PM2.5与其他气态污染物之间的关系,探讨各气态污染物在PM2.5二次合成中的贡献;建立PM2.5和PM10的回归模型,方便通过PM10对PM2.5进行预测。结果显示:(1)太原市区空气污染物中,PM2.5和PM10相关性最强;(2)PM2.5和PM10回归分析得到回归模型为 $PM2.5=0.63PM10-11.76(R^2=0.8427)$,回归方程拟合度较好;(3)PM2.5和其他气态污染物多元线性回归模型为 $PM2.5=0.24SO_2+20.08CO+11.54(R^2=0.4844)$,拟合度检验效果一般,考虑三者之间还有其他因素的影响。

关键词: R语言; 数据分析; 相关分析; 回归分析

中图分类号: TP315.69 **文献标识码:** A

An Analysis of the Influencing Factors for PM2.5 in Cities Based on R Language

XIE Lei¹, DI Guangzhi²

(1. Department of Mathematics and Computer Science, Yuncheng Advanced College, Yuncheng 044000, China;

2. School of Computer and Information Technology, Southwest Forestry University, Kunming 650224, China)

Abstract: Using R language as a tool for data analysis, based on correlation analysis and multiple linear regression analysis methods, the paper analyzes the influencing factors of PM2.5 in Taiyuan City, studies the relationship between PM2.5 and other gaseous pollutants, and explores the contribution of gaseous pollutants in secondary synthesis. The regression model of PM2.5 and PM10 is established to facilitate the prediction of PM2.5 through PM10. The results show: the correlation between PM2.5 and PM10 is the strongest among the air pollutants in Taiyuan; the regression model achieved through the regression analysis on PM2.5 and PM10 is $PM2.5=0.63PM10-11.76(R^2=0.8427)$, with high fitting degree; the PM2.5 and other gaseous pollutants multivariate linear regression model is $PM2.5=0.24SO_2+20.08CO+11.54(R^2=0.4844)$. The test results are general, considering the influence of other factors.

Keywords: R language; data analysis; correlation analysis; regression analysis

1 引言(Introduction)

PM2.5是指大气中直径小于或等于2.5微米的颗粒物,也称为可入肺颗粒物。它的直径还不到人的头发丝粗细的1/20。虽然PM2.5只是地球大气成分中含量很少的组成,但它对空气质量和能见度等有着重要的影响。与较粗的大气颗粒物相比,PM2.5由于粒径小,富含大量有毒、有害物质,且在大气中停留时间长、输送距离远,因而对人体健康和大气环境质量影响更大,是导致心脑血管系统和呼吸系统的损伤,增加患癌症的风险,以及雾霾产生的主要因素^[1-3]。

本研究通过python编写接口程序,获取太原市九处国控监测点,2014年3月至2018年2月的PM2.5及其他污染指标数据。通过对近四年的数据进行统计分析,探讨太原市PM2.5形成的影响因素。此研究将为太原市相关管理部门和人员,对城市规划建设、城市环境治理和生态文明建设工作提供合理有效的科学参考。

2 R语言简介(Introduction to R language)

R是一种用于统计分析计算及图形化展示的开源软件,同时也是一种编程语言,它广泛应用于农业、林业、商业、工

业、政府部门、医药和科研等众多涉及数据分析的领域,随着近几年数据挖掘、大数据等概念不断发展,R作为数据分析的利器也越来越多地被人关注^[4-6]。

3 方法及原理(Methods and principles)

为了研究PM2.5和其他大气污染物和气象条件之间的关系,根据变量类型,在本研究中主要用到了二元定距变量的相关分析。二元定距变量的相关分析通过计算定距变量间两两相关的相关系数,对定距变量两两相关程度进行分析^[7]。

Pearson简单相关系数计算式:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

式(1)中, r 成为观测量相关系数,在统计学中,一般用观测量相关系数来推断总体相关系数。 X_i 、 Y_i 分别为随机变量 X 与随机变量 Y 的第 i 个样本值,总共样本容量为 n 。

相关系数取值范围在-1和+1之间,即 $-1 \leq r \leq +1$ 。其中:

- 若 $0 < r < 1$,表明变量之间存在正相关关系。
- 若 $-1 > r > 0$,表明变量之间存在负相关关系。
- 当 $|r|=1$ 时,其中一个变量的取值完全取决于另一个变量。
- 当 $r=0$ 时,说明两变量之间不存在线性相关关系。

为了进一步描述PM2.5与PM10的关系,进一步利用回归分析进行分析。一元线性回归模型只包含一个解释变量和一个被解释变量,设有两个变量 x 和 y ,变量 y 的取值随变量 x 取值的变化而变化,称 y 为因变量, x 为自变量^[8-10]。相对于自变量和因变量,通过观察和实验可以得到若干组数据,记作 (x_i, y_i) ($i=1, 2, \dots, n$)。用直线方程的方法来表示他们之间的关系:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2)$$

式(2)称为一元线性总体回归模型。式中, β_0 和 β_1 是未知参数, β_0 称为回归常数, β_1 称为回归系数。 ε 称为随机扰动项,通常满足:

$$\begin{cases} \text{var}(\varepsilon) = \sigma^2 \\ E(\varepsilon) = 0 \end{cases} \quad (3)$$

式(3)中 $E(\varepsilon)$ 表示 ε 的数学期望, $\text{var}(\varepsilon)$ 表示 ε 的方差, σ^2 称为总体方差。根据均值的思想

$$E(y) = \beta_0 + \beta_1 x \quad (4)$$

可以由式(4)得到一元线性总体回归方程:

$$y = \beta_0 + \beta_1 x \quad (5)$$

一元线性回归模型的样本回归方程可以表示为:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (6)$$

式(6)中的 $\hat{Y}$ 是样本回归直线上与 x 相对应的 y 值,可视为 $E(y)$ 的估计值; $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是未知常数,作为总体回归常数 β_0 、 β_1 的估计值。

实际测量的因变量 y 值,并不完全等于 $\hat{y}$,其二者之差为 $e=y-\hat{y}$ 。由此得到一元线性样本回归模型为:

$$y = \beta_0 + \beta_1 x + e \quad (7)$$

式(7)中 e 称为残差,它与总体回归模型中的随机扰动项 ε 在概念上相对应。所不同的是 ε 是 y 与未知的总体回归直线之间的纵向距离,它是不可直接观测的。而 e 是 y 与样本回归直线之间的纵向距离,当根据样本观测之拟合出回归直线后,可以计算出 e 的具体数值。

在研究具体问题时,对于组样本观测 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 来说,式(7)还可以有以下样本回归模型表达:

$$y_i = \beta_0 + \beta_1 x_i + e_i, i=1, 2, \dots, n \quad (8)$$

e 满足:

$$\begin{cases} \sum_{i=1}^n e_i = 0 \\ \sum_{i=1}^n x_i e_i = 0 \end{cases} \quad (9)$$

为了由样本数据得到总体回归参数 β_0, β_1 的 $\hat{\beta}_0, \hat{\beta}_1$,一般利用普通最小二乘法估计,确定回归方程 $\hat{\beta}_0, \hat{\beta}_1$ 的取值,从而确定一条与总体回归直线最为拟合的样本回归直线。

选择 $\hat{\beta}_0, \hat{\beta}_1$,使得

$$\sum e^2 = \sum (y - \hat{y})^2 \quad (10)$$

为最小。

由于残差平方和

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad (11)$$

当 $\sum e^2$ 为最小时, $\sum e^2$ 对 $\hat{\beta}_0, \hat{\beta}_1$ 的偏导数等于零。从而可导出 $\hat{\beta}_0, \hat{\beta}_1$ 的式:

$$\hat{\beta}_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \quad (12)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (13)$$

4 分析过程(Analysis process)

4.1 PM2.5与其他污染物的关系

表1 PM2.5与各污染物相关性检验

Tab.1 Correlation test of PM2.5 and other pollutants

污染物	R	t	p-value	95%置信区间
PM2.5和PM10	0.92	62.492	2.2e-16	0.91—0.93
PM2.5和SO ₂	0.66	23.613	2.2e-16	0.62—0.71
PM2.5和CO	0.66	23.127	2.2e-16	0.61—0.70

为了了解PM2.5和其他各污染物之间的关系,采用Pearson相关系数检验,从表1可以看出,PM2.5和PM10,PM2.5和SO₂,PM2.5和CO之间有较强的线性相关性,相关系数分别为0.92、0.66、0.66,相伴概率P值均小于0.05,说明两两之间相关性很显著。PM2.5和PM10线性相关性最强,说明二者具有相似的污染源。除臭氧之外PM2.5和其他污染物的相关系数介于0.56至0.92,说明臭氧污染对PM2.5影响不大,太原市的PM2.5污染除一次排放之外,气态污染物的二次化学合成也占一定比重。

4.2 PM2.5与PM10的回归分析

我国对PM10的监测研究起步较早,由于空气质量的恶化

和雾霾事件的频发,对PM2.5的关注是从近几年才开始,对PM2.5的监测区域和监测点数都有所限制,所以研究PM2.5和PM10之间的关系,探索分析二者之间的回归模型,对于PM2.5的预测和研究有重要的意义。

表2 PM2.5和PM10线性模型系数表

Tab.2 Correlation coefficient table of PM2.5 and PM10

linear model				
coefficient	Estimate	Std.Error	t value	Pr(> t)
Intercept	-11.763629	1.298264	-9.061	<2e-16 ***
PM10	0.623910	0.009984	62.492	<2e-16 ***

对PM2.5和PM10做回归分析,从线性模型系数表(表2)中可以得到回归方程 $PM2.5=0.63PM10-11.76$,回归模型图如图1所示,PM10对应的 t 统计量值为62.492,相伴概率 $<2e-16$,远小于系统默认值0.05,所以PM10在回归模型中控制作用非常显著。方差分析结果显示, F 检验量的值为3905,相伴概率是 $2.2e-16$,说明回归方程很显著,可决系数 $R^2=0.8427$,修正可决系数为0.8425,说明以上回归方程拟合度较好。

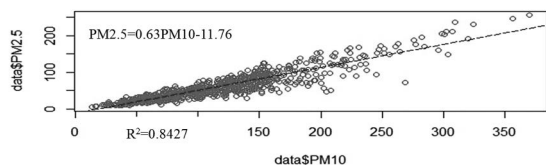


图1 PM2.5与PM10回归模型

Fig.1 Regression model of PM2.5 and PM10

4.3 PM2.5与其他气态污染物的关系

PM2.5的来源比较复杂,除了土壤扬尘、植物花粉等自然源外,太原市的PM2.5污染只要是人为源,从之前的区域监测点数据对比,以及太原市空气污染物主成分分析可以看出,太原市的PM2.5污染源主要是以工业源和交通源。按照形成方式来看,PM2.5又分为一次排放和二次合成,由于工业过程和燃烧排放的气态前体污染物通过大气化学反应生成二次颗粒物。

通过相关性检验发现PM2.5和SO₂、CO两种气态污染物之间关系最大,说明SO₂和CO在PM2.5的二次合成中贡献较大,根据两年的样本数据,对PM2.5和CO、SO₂进行多元回归,得出回归方程。

表3 PM2.5与SO₂、CO多元回归系数表

Tab.3 Multiple regression coefficient table of PM2.5 and SO₂、CO

Coefficients	Estimate	Std.Error	t	Pr(> t)
(Intercept)	11.5444	2.9308	3.939	8.97e-05 ***
data\$SO ₂	0.2435	0.0262	9.298	<2e-16 ***
data\$CO	20.0814	2.3656	8.489	<2e-16 ***

回归结果(表3)可以看出,回归方程为 $PM2.5=0.24SO_2+$

$20.08CO+11.54$ 。回归系数表显示,SO₂的回归系数 t 统计量为9.298,相伴概率 $<2e-16$,CO回归系数 t 统计量为8.489,相伴概率 $<2e-16$,说明两变量在回归模型中的作用显著。 F 统计量为342,相伴概率 $<2e-16$,说明回归方程很显著,判定系数 $R^2=4844$,拟合度检验(图2)效果一般,说明三者之间还有其他因素的影响。

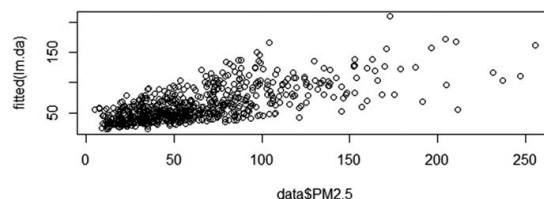


图2 PM2.5和回归值的拟合效果图

Fig.2 Fitting plot of regression value and PM2.5

5 实例分析代码(Analysis code)

(1)PM2.5和PM10回归分析

```
getwd()
folder<-"E:/数据/官网数据"
setwd(folder)
data<-read.csv("data-day.csv",header=TRUE,
sep=",")
data
plot(data$PM2.5~data$PM10,data=data,pch=1,col='red')
x<-data$PM2.5
y<-data$PM10
data2<-data.frame(PM2.5=x,PM10=y)
data2
test.pr<-princomp(data2,cor=TRUE)
summary(test.pr,loadings=TRUE)
model=lm(PM2.5~PM10,data=data2)
abline(model,lty=5)
summary(model)
(2)PM2.5和SO2、CO多元回归
getwd()
folder<-"E:/数据/官网数据"
setwd(folder)
data<-read.csv("data-day.csv",
header=TRUE,sep=",")
data
lm.da=lm(data$PM2.5~data$SO2+data$CO)
summary(lm.da)
plot(data$PM2.5,fitted(lm.da),type="p")
```

6 结论(Conclusion)

太原市区空气污染物中,PM2.5和PM10相关性最强,另外PM2.5和SO₂、CO两气态污染物之间也有较强的相关性,

(下转第8页)