PM2.5 预测模型及其特征研究[1]

中山大学　梁植斌、彭颖鸿、王珊珊

摘要

随着现代工业的迅猛发展、雾霾天气在全国大部分地区蔓延，民众逐渐意识到大气污染的巨大危害。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，它只占有空气中很小的一部分，但却对空气质量和能见度有非常大的影响。它比较大的 PM10 的毒性更强，对健康的影响更大。空气质量数据和气象数据均是利用爬虫程序收集，通过收集在天气后报网和全球天气精准预报网上公布的空气质量数据和气象数据，可以对 PM2.5 的一些特征进行分析。本文通过对 PM2.5 的历史数据进行滤波分析，可以看出 PM2.5 的季节性变化的特点，同时与其他相关的指标做一个对比找出其他空气质量指标、气象指标与 PM2.5 之间的关系。利用主成分分析，找出与 PM2.5 浓度最为相关的指标，以及一些与 PM2.5 相关的隐藏变量。最后，基于前一天的空气质量数据和气象数据，建立 LASSO 回归，处理多重共线性问题，对第二天的 PM2.5 的浓度进行预测。

关键词：PM2.5；主成分分析；滤波分析；LASSO 回归

[1] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类本科生组三等奖。

1 引言

1.1 研究背景与意义

随着现代工业的迅猛发展、城市规模的快速扩张以及人口数量的持续增加，大气环境质量日益恶化,城市空气质量不容乐观。如今，雾霾天气在全国大部分地区蔓延，给市民的生活带来了诸多不便，对城市的发展造成恶劣的影响，政府以及民众逐渐意识到大气污染的巨大危害。在许多地方的两会上，雾霾治理更是成为提及频度最高的热点之一，各地纷纷出重拳治霾[1]。

2013 年 7 月 11 日《中国新闻网》报道：“2013 年初以来，中国发生大范围持续雾霾天气。据统计，受影响雾霾区域包括华北平原、黄淮、江淮、江汉、江南、华南北部等地区，受影响面积约占国土面积的 1/4，受影响人口约 6 亿人。（中国国家发展和改革委员会(发改委)11 日公布在官方网站上的一份报告披露了上述信息。）这份题为《节能减排形势严峻产业发展潜力巨大》的报告指出，本轮雾霾天气呈现出三大特点：其一是影响范围广；其二是持续时间长，一月份北京市只有 5 天达到二级标准；三是污染物浓度高。[2]”而现在，雾霾天气仍然持续影响着人名的正常生活；因此，对雾霾的治理十分迫切。

2012 年 2 月 29 日中国环境保护部下发关于实施《环境空气质量标准》的通知，将用空气质量指数（AQI）替代原有的空气污染指数（API）。而 AQI 增加了雾霾的主要成因——PM2.5 作为检测指标，并与 SO_2、NO_2、PM10、O_3、CO 等指标组合成为评价空气质量的主要指标[1]。PM2.5 是指能够进入人体肺泡的空气动力学当量直径不大于 2.5μm 的大气悬浮颗粒。与较粗的大气颗粒物相比，PM2.5 粒径小，面积大，吸附能力强，可在大气中悬浮较长时间，能够通过呼吸道直接进入人体肺部，因而对人体健康和大气环境质量的影响更大[3]。

在我国，PM2.5 的研究尚处于初级阶段，并且由于 PM2.5 对空气质量有着重要的影响，是政府、环保部门以及全国人民关注的热点，因此对 PM2.5 的研究是十分必要的。

1.2 本文研究内容

本文研究的对象是 PM2.5，用到的主要指标是可以从各大网站中找到的空气质量数据和气象相关数据，包括 AQI 指数、PM2.5、PM10、Co、No2、So2、温度、露点、湿度、气压、风速、降雨量等变量以及他们对应的时间地点，选取这些数据，主要是因为他们比较容易获取，在众多的网站中都可以找到相关的数据，同时，可以期望这些数据能在未来的很大一段时间都可以收集到。然后利用这些数据，我们可以进行基于滤波分析的 PM2.5 季节性特点研究，PM2.5 的浓度波动是非常大的，单单看 PM2.5 随着时间的变化，由于波动太大，没有办法看出有明显的规律，经过滤波分析，可以找出在隐藏在波动后的一些趋势。我们收集了与 PM2.5 相关的一些污染物以及一些气象数据，我们试图利用主成分分析找出具体是哪几种指标与 PM2.5 浓度高度相关，以便了解 PM2.5 以及如何防治的对策。最

后，本文建立了一个 LASSO 模型，旨在利用前一天的数据来预测后一天的数据，预测第二天的 PM2.5 浓度对人们的工作生活均有一个指导意义。

下面介绍本文的章节内容安排：

第一章主要介绍了本文研究课题的背景、现实意义，并介绍了本文的主要研究内容和章节安排。

第二章介绍了在本文所用的数据以及数据的采集和预处理。

第三章是利用滤波分析来对 PM2.5 的季节差异性分析。

第四章是通过主成分分析以找出数据中影响 PM2.5 的指标，并根据得出的主成分找出影响 PM2.5 的隐藏指标。

第五章是利用 LASSO 模型的建立，利用 LASSO 对数据进行特征选取，并对 PM2.5 进行预测。

第六章是本文的总结。

第七章是本文的不足和对未来的展望。

2 数据采集

2.1 数据介绍

本文所用数据主要包括空气质量数据、气象数据，其中，空气质量数据包括有日期、AQI 指数、PM2.5、PM10、Co、No2、So2 这几个不同的变量，对应的数值单位为μg/m3(CO 为 mg/m3)，气象数据包括气温，露点，湿度，海平面气压，能见，风速，降雨量，活动等变量。

2.2 数据来源及采集方法

本文数据中的变量都可以查询到实时的记录，实时记录是每小时公布的，按小时公布的记录只有当天的实时记录，对于历史数据，网络上没有找到有相应按小时记录的数据提供。如果想要查看历史的记录，就只能找到单日的记录。所以，这里我们采集了以天为单位的数据，对于空气质量数据，即为该天 24 小时内的平均水平；对于气象数据，由于一天内数值是变化的，所以对于气温，露点，湿度，海平面气压，能见，风速等数据均有最大值，平均值和最小值，对所有这些变量，利用程序全部采集下来。对于数据的来源，我们采集的数据主要来自两个网站，其中，空气质量数据在天气后报网站(http://www.tianqihoubao.com)中收集，而气象数据则在全球天气精准预报网(http://www.wunderground.com/)中收集。由于天气后报网站和全球天气精准预报网中均没有对应的接口可供直接下载数据，所以我们自己写了一个爬虫程序用于收集空气质量数据和天气数据。

爬虫程序主要有三个功能：获取 URL、下载页面和解析页面。首先需要获取天气后报网、全球天气精准预报网中关于空气质量和气象数据的 URL，然后利用程序下载所有这些页面，最后对页面进行解析，利用正则表达式提取我们需要的

内容，整个过程，我们只需给定天气后报网和天气精准预报网的 URL，爬虫程序就会自动解析，下载相关的内容。利用爬虫程序，我们可以收集到全国各个城市的历史数据，单由于 PM2.5 是最近几年才受到社会的关注，相关的历史数据不是非常的多，基于我们的研究对象，我们收集了广东省 21 个城市从 2013 年 11 月起至今的数据。整个爬虫程序的流程可如下图 2.1 所示。

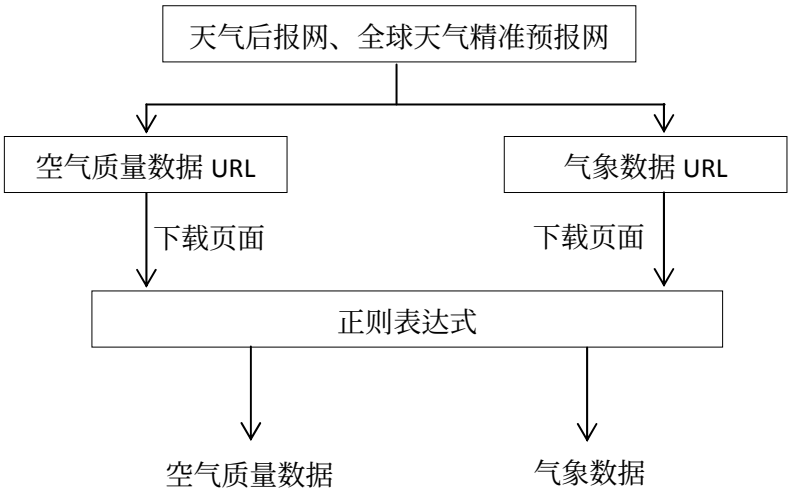

图 2.1　爬虫程序的流程

2.3 数据预处理

在两个网站分别获取到空气质量数据和气象数据之后，需要对数据进行一个预处理，首先，按照日期，对数据集进行合并，合并之后的变量包括有空气质量和气象数据的相关变量，这其中，有部分变量我们选择性地删除了，包括瞬时最大风速，CloudCover 和活动，前两个变量是因为大部分记录都是缺失的，对分析用处不大，而活动则记录了当天的天气情况，如小雨、中雨等，但是，这可以用降雨量来衡量，降雨量是一个连续形的变量，比活动记录了更加多的信息。最后，因为我们要利用前一天的数据去预测当天的 PM2.5 水平，所以，我们需要取当天的 PM2.5 数值与前一天的空气质量和气象数据结合。

3　PM2.5 季节性差异分析

3.1 PM2.5 的季节性分析

随着季节变化，天气参数也会发生相应变化，进而影响到 PM2.5 的浓度，分布等等。我们选取了广州 2014 年 3 月到 2015 年 2 月的 PM2.5 数据进行分析。其中，三、四、五月划分为春天，六、七、八月划分为夏天，九、十、十一月为秋

天，十二、一、二为冬天。我们的目标，是将这一年中 PM2.5 的一些季节性的特征、趋势提取出来进行分析。这里我们可以使用信号处理的相关方法方法，将这一年的 PM2.5 的数据视作一段离散的信号。一般来说，对于一个信号，我们会分离出其高频项与低频项，即“波动”与“趋势”：

$$x(t) = x_h(t) + x_l(t).$$

这样分解的好处是能够去掉“高频项”，或者说“噪声”对数据分析的影响，从而获得更加准确的趋势信息[4][5]。

由于每天的天气参数以及空气参数并无十分明显的规律，所以我们将使用经验模式分解（Empirical Mode Decomposition，EMD）将原始信号进行拆解。EMD 是一种自适应的滤波分析方法。相对于传统的傅立叶分析或小波分析，它能自适应地分依据数据自身的时间尺度特征来进行分解，无须预先设定任何基函数，对信号的波形要求低，理论上可以被用在任何类型的信号上，特别实在非平稳及非线性的数据处理上有明显的优势。主要的算法如下[4][5]：

1．寻找信号$x(t)$中的所有极大值和极小值，然后采用三次样条插值进行处理，分别获得信号$x(t)$的上包络线$x_u(t)$和下包络线$x_d(t)$，信号所有的数据点都位于上下包络线之间。计算：

$$m(t) = \frac{x_u(t) + x_d(t)}{2}$$

2．计算

$$x_1(t) = x(t) - m(t)$$

若满足 IMF 条件，则记$x_1(t)$为第一个 IMF 分量IMF_1，否则，令$x(t) = x_1(t)$，重复第 1 步和第 2 步，直到得到第一个 IMF 分量；

3．记$r_1(t) = x(t) - IMF_1$，并令$x(t) = r_1(t)$，重复以上步骤，直到剩余的信号$r_n(t)$为非震荡的函数或误差小于给定值时停止，原始信号最终分解为n个 IMF 分量和一个残差项$r_n(t)$，即：

$$x(t) = \sum_{i=1}^{n} IMF_i + r_n(t)$$

其中，IMF 条件为：

1．信号的极大极小值数目与过零点数目相等或最多相差一个；

2．信号的任一点上，由极大值确定的包络线与由极小值确定的包络线的均值在 0 附近。

原始数据的信号图如图 3.1。原始信号由于波动较大，阻碍了对季节性差异分析。我们需要将这些波动过滤掉，得到相应的季节性系。对一年的数据进行 EMD 分解，得到其“趋势项”$r_n(t)$和“波动项”$\sum_{i=1}^{n} IMF_i$如图 3.2 和图 3.3。

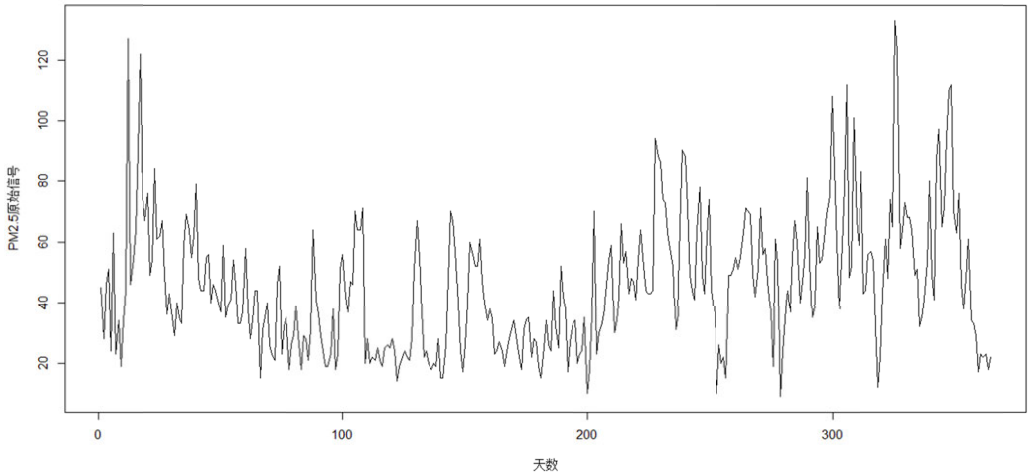

图 3.1 PM2.5 原始信号图

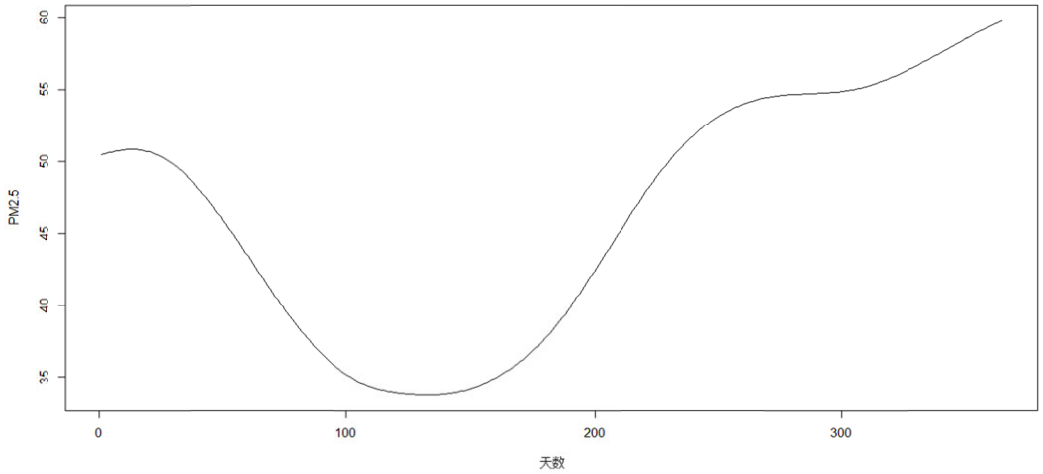

图 3.2 PM2.5 趋势项走势图

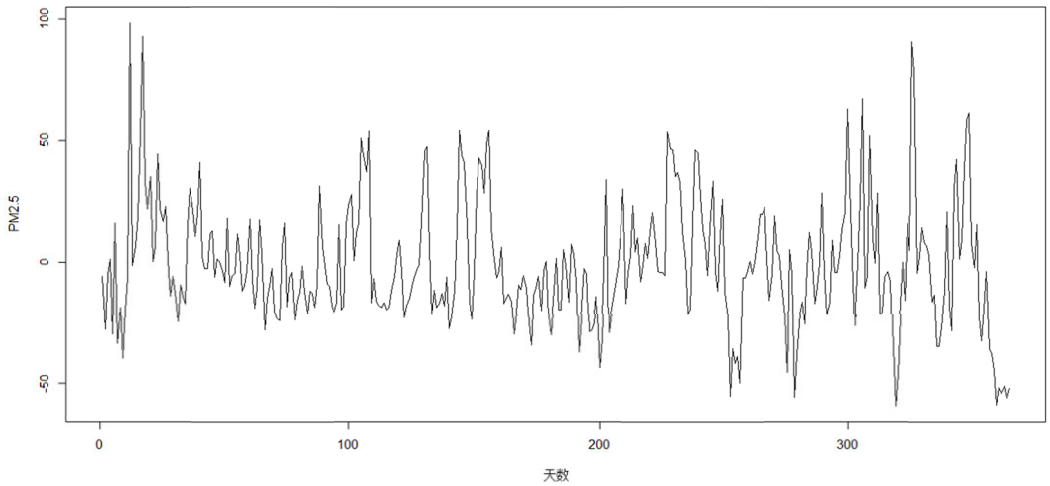

图 3.3 PM2.5 波动项走势图

由上面的图，我们可以得出，高频项中波动极为频繁，难以分析其规律。但低频项中，很明显地，夏季的 PM2.5 处于一个较低的水平，而其它季节的 PM2.5 相对较为严重。十一，十二，一，二四个月非常突出，这说明冬天是 PM2.5 的一个爆发期。

3.2 PM2.5 与季节相关变量的联系

如果我们将几个明显与季节相关的变量，如降水量，温度，SO_2浓度，PM2.5 浓度等也纳入研究，会得出更多有用的结论。

（1） 降水量的季节性变化。

我们对相同时间段的降雨量进行 EMD 分解，并将其趋势项与 PM2.5 的趋势项进行对比（如图 3.4），发现波形的走向明显是“ 异向 ”的——降雨量增大，PM2.5 呈减小的趋势；降雨量变小，PM2.5 就有抬头的倾向，并且这都是季节分明的。夏季降水量大，能“ 清洗 ” 空气中的 PM2.5。

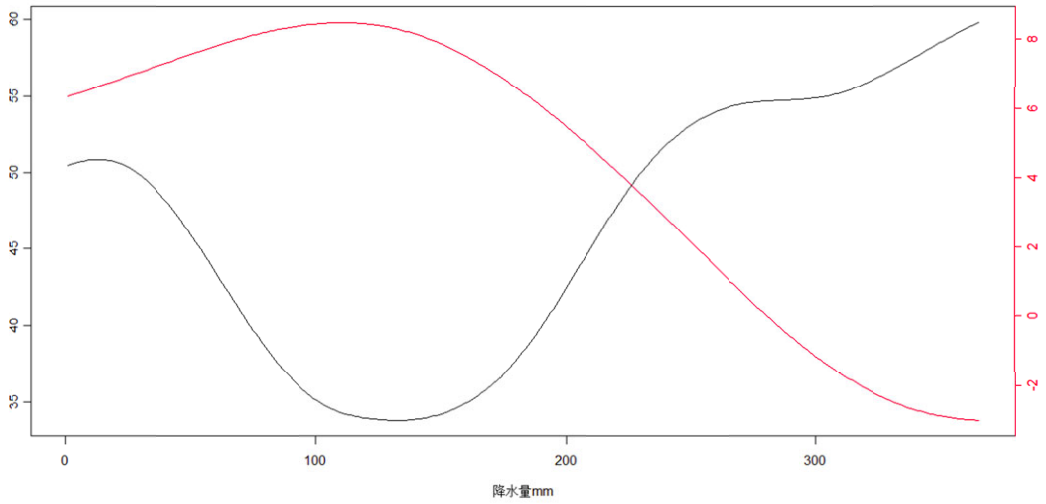

图 3.4 降水量趋势项与 PM2.5 趋势项对比图

（2） 温度的季节性变化。

我们对温度的数据进行 EMD 分解，并将之与 PM2.5 进行比较，如图。发现，温度与 PM2.5 也是呈现“ 异向 ” 的现象 ;在夏季, 温度升高, PM2.5 的水平降低。其背后的原因, 很可能就是温度越高, PM2.5 中的二次颗粒越难形成, 二次颗粒, 指的是大气中某些污染组分之间, 或这些组分与大气成分之间发生反应而产生的颗粒物。它在 PM2.5 中占了相当大的一部分。

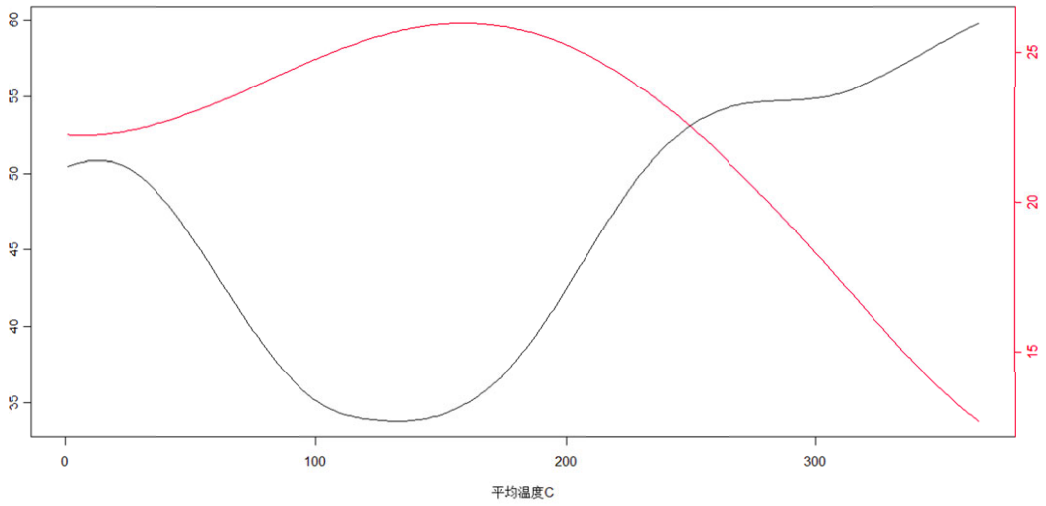

图 3.5 温度趋势项与 PM2.5 趋势项对比图

(3) SO_2的季节性变化。

我们可以看到，SO_2与 PM2.5 的波形并不同步，甚至有相当一部分是异向的。这很可能是由于二次颗粒主要是由SO_2与大气其它成分反应而来，导致某些时候 PM2.5 偏高时，SO_2反而偏低。但并不总是这样的情况。在污染比较严重的时候，SO_2和 PM2.5 也可能一起上升。

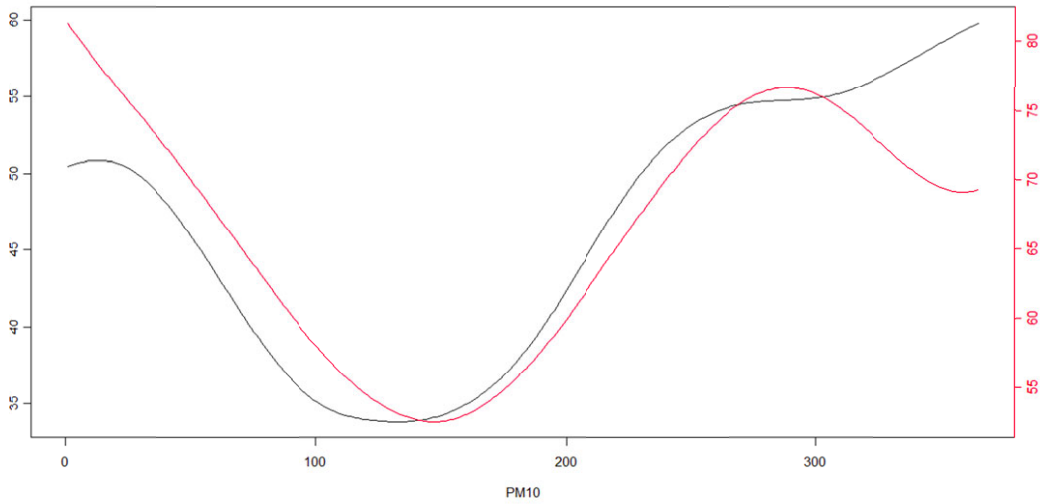

图 3 6 SO_2趋势项与 PM2.5 趋势项对比图

(4) PM10 的季节性变化。

从图像可以看出，PM10 和 PM2.5 几乎是同步的，PM10 也有夏季偏偏低，其它季节偏高的情况。这个结果是符合逻辑的，因为 PM10 和 PM2.5 仅仅是颗粒测量半径的标准有所区别。

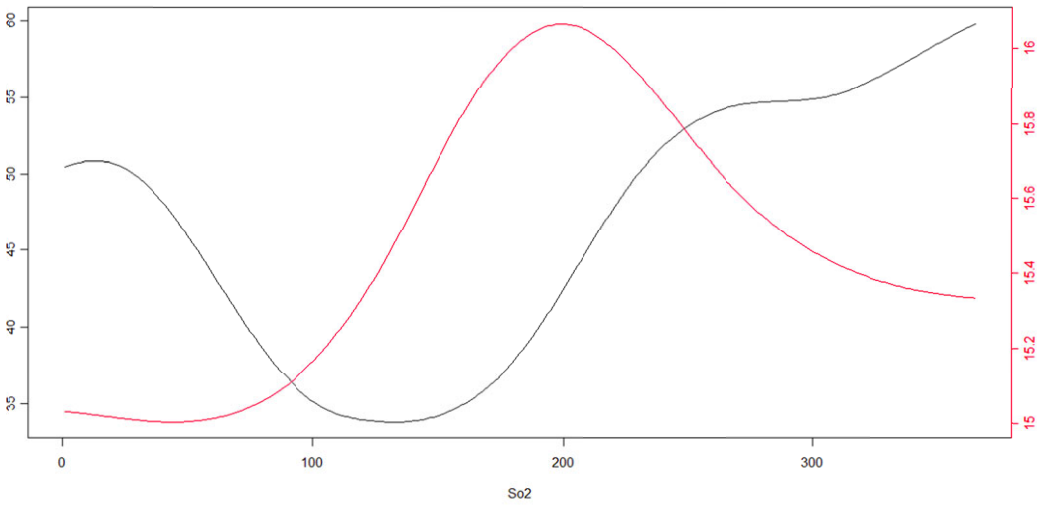

图 3.7 PM10 趋势项与 PM2.5 趋势项对比图

4 PM2.5 成因分析

4.1 主成分分析

上一节我们通过用前一天的各变量来预测第二天的 PM2.5，以得到前一天各变量对第二天 PM2.5 之间的关系。然而，现在我们希望研究每日的各变量对当日 PM2.5 的影响。因此我们选择了主成分分析法。值得指出的是 AQI 变量由 PM2 5、CO、SO_2等污染指标计算而成，而现在我们希望研究当日个影响因素对 PM2.5 的影响,即此时 PM2.5 属于未知值，因此，在这一节中我们将删除 AQI 变量和上一届作为自变量的 PM2.5 变量。

主成分分析是一种将多个变量化为几个综合指标的统计分析方法。这些综合指标通常为原始指标的线性组合，并且它们能够较好的反应原变量所代表的信息。假设我们有数据矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=\begin{bmatrix} X_{(1)}^{T} \\ X_{(2)}^{T} \\ \vdots \\ X_{(n)}^{T} \end{bmatrix}=\left[X_1, X_2, \cdots, X_p\right]$$

其中 X_j 为矩阵的各列，$X_{(k)}=\left(x_{k1}, x_{k2}, \cdots, x_{kp}\right)^{T}(k=1,2,\cdots,n)$ 为总体 X 的样本。

由于我们采用样本的相关矩阵作主成分分析, 因此我们只介绍用样本相关矩阵求主成分的理论知识。由上，样本的相关矩阵为

$$R=\frac{1}{n-1}\sum_{k=1}^{n}X_{(k)}^{*}X_{(k)}^{*\ T}=\left(r_{ij}\right)_{p\times p},$$

其中

$$X_{(k)}^{*}=\left[\frac{x_{k1}-\overline{x_1}}{\sqrt{s_{11}}},\frac{x_{k2}-\overline{x_2}}{\sqrt{s_{22}}},\cdots,\frac{x_{kp}-\overline{x_p}}{\sqrt{s_{pp}}}\right],$$

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}},\quad i,j=1,2,\cdots,p\ ,$$

$$\left(\overline{x_1},\overline{x_2},\cdots,\overline{x_p}\right)^{T}=\frac{1}{n}\sum_{k=1}^{n}X_{(k)}$$

$$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_{ki}-\overline{x}_i\right)\left(x_{kj}-\overline{x}_j\right)$$

设 $\lambda_1^{*}\geq\lambda_2^{*}\geq\cdots\geq\lambda_p^{*}\geq 0$ 为样本的相关矩阵 R 的特征值， $a_1^{*},a_2^{*},\cdots,a_p^{*}$ 为相应的单位特征向量，并且它们彼此正交。令

$$Z_{(i)}^{*}=Q^{T}X_{(i)}^{*}\ ,$$

其中 $Q=\left(a_1^{*},a_2^{*},\cdots,a_p^{*}\right)$,则样本主成分为

$$Z^{*}=\begin{bmatrix}z_{11}^{*} & z_{12}^{*} & \cdots & z_{1p}^{*}\\ z_{21}^{*} & z_{22}^{*} & \cdots & z_{2p}^{*}\\ \vdots & \vdots & & \vdots\\ z_{n1}^{*} & z_{n2}^{*} & \cdots & z_{np}^{*}\end{bmatrix}=\begin{bmatrix}Z_{(1)}^{*\ T}\\ Z_{(2)}^{*\ T}\\ \vdots\\ Z_{(n)}^{*\ T}\end{bmatrix}=\begin{bmatrix}X_{(1)}^{*\ T}Q\\ X_{(2)}^{*\ T}Q\\ \vdots\\ X_{(n)}^{*\ T}Q\end{bmatrix}=X^{*}Q$$

$$=\left[X^{*}a_1^{*},X^{*}a_2^{*},\cdots,X^{*}a_p^{*}\right]=\left[Z_1^{*},Z_2^{*},\cdots,Z_p^{*}\right]$$

其中 $Z_{(k)}^{*\ T}$ 为主成分各行， Z_j^{*} 为主成分各列[6]。

4.2 主成分分析结果

我们采用 R 软件对数据进行主成分分析，最终得到 23 个主成分。图 4.1 给出了主成分的碎石图。

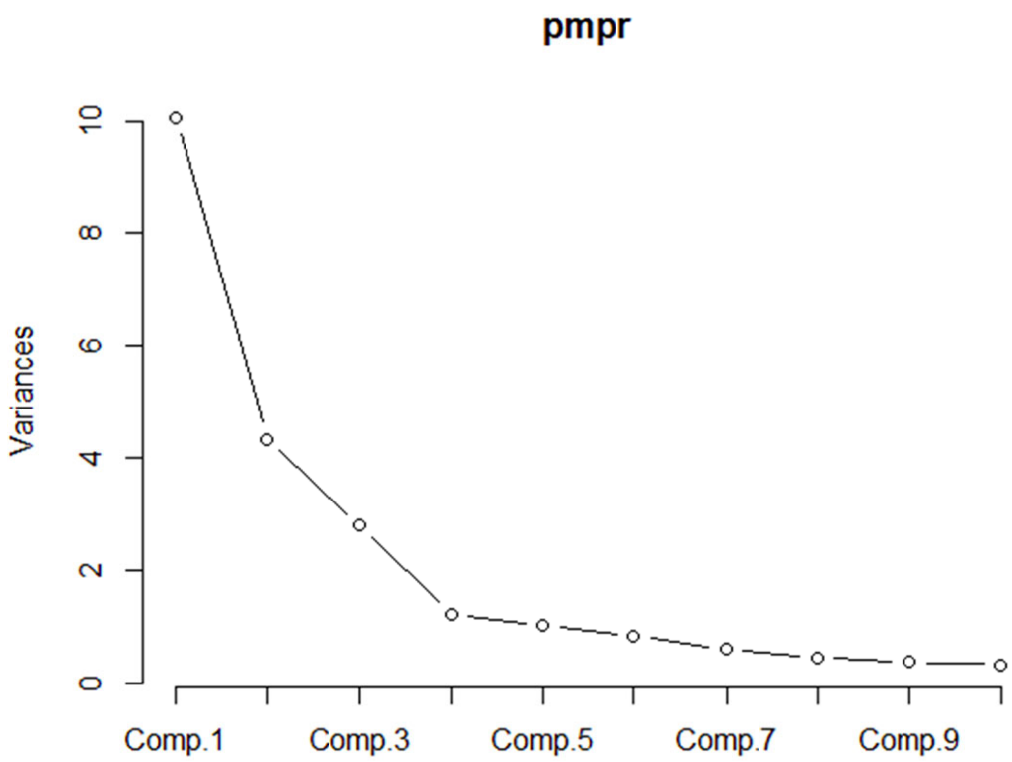

图 4.1 主成分的碎石图

由图 4.1 可得，成分 1 至成分 4 较为陡峭，说明它们包含了大部分信息；而成分 4 之后的变化都较平缓；且从表 4.1 可看出前 4 个主成分的方差累积贡献率已达到 80.19%。因此，我们选择成分 1 至成分 4 作分析，舍去剩下的 19 个主成分。表 4.2 给出了 4 个主成分的详细情况。需要指出的是，有的影响因素有多个指标，为解释方便，我们将以其统称代替。如温度有最高温度、最低温度、平均温度；我们将统称为温度；若有单独解释其详细变量时，会再直接指出。

表 4.1 所得主成分的重要信息

成分 / 重要信息	成分 1	成分 2	成分 3	成分 4
标准差	3.1699336	2.0867272	1.6749658	1.11121857
方差贡献率	0.4368904	0.1893231	0.1219787	0.05368725
方差累积贡献率	0.4368904	0.6262135	0.7481922	0.80187943

表 4.2 所得主成分的详细信息

成分 变量名	成分 1	成分 2	成分 3	成分 4
PM10	-0.168	0.284	0.284	0.188
Co	-0.163	0.281		0.198
No2	-0.129	0.354	0.176	
So2		0.209	0.390	0.241
最高温度 C	0.252		0.329	
平均温度 C	0.283		0.235	
最低温度 C	0.297		0.117	
露点 C	0.307			
MeanDew.PointC	0.308			
Min.DewpointC	0.305			
Max.湿度	0.185	0.223	-0.223	-0.200
Mean.湿度	0.210	0.163	-0.366	
Min.湿度	0.171		-0.436	
Max.海平面气压 hPa	-0.294		-0.120	-0.112
Mean.海平面气压 hP	-0.293		-0.119	-0.124
Min.海平面气压 hPa	-0.290		-0.131	-0.134
Max.能见度 Km		-0.320	0.100	-0.252
Mean.能见度 Km		-0.412	0.127	-0.158
Min.能见度 kM		-0.353	0.171	-0.105
Max.风速 Km.h		-0.261	-0.102	0.612
Mean.风速 Km.h		-0.336	-0.151	0.422
降水量 mm	0.121		-0.201	0.335
WindDirDegrees				

从表 4.2 中我们可以看出,成分 1 大部分所反映的是湿度的影响,如降雨量、最大湿度、最小湿度、平均湿度、露点以及温度对湿度都有影响。除了它包含于湿度相关的变量多以外，我们还可以看出露点的系数都较高，由此可以判定成分 1 较多地反映的是湿度。而我们可以注意到气压和 PM10、CO、NO_2、SO_2的系数都是负的,而剩下的变量的系数都大于 0,因此成分 1 表示的是温度、湿度与压强、空气污染度之差，反映的是天气状况，如天晴、大雨等。

成分 2 只有风速、能见度、湿度和 PM10、CO、NO_2、SO_2有系数，反映的是 PM10、CO、NO_2、SO_2等污染物以及湿度与风速和能见度之差。值得注意的是成分 2 是唯一一个将所有污染物都包含的主成分，且其系数较大，在 0.209~0.354 之间。因此它更多地是反应空气的污染程度和污染的扩散程度。

成分 3 表示的是能见度、温度、污染物与湿度、风速、降水量、气压之差。成分 3 中系数绝对值较大的是温度和湿度，且两者符号相反。因此我们认为成分 3 较多的反应的是空气的干湿情况以及天气的湿热情况。

成分 4 表示的是风速、降水量污染物与能见度、海平面、湿度之差。成分 4 中最大风速的系数十分高，为 0.612；而与其符号相反且系数绝对值较高的变量为能见度。因此，成分 4 反映的是能见度与风速之间的关系。值得注意的是，风

速中以最大风速的系数最高，降雨量也达到 0.335，所以，我们猜想成分 4 反映的是极端天气，例如：沙尘暴、台风等。

综上所述，4 个主成分较好地反映了所有变量的信息。

4.3 主成分回归

5.2 中我们得出的主成分较好地表示了变量的完整信息。然而，它与 PM2.5 是否有显著的关系？因此，我们将用上节主成分分析所得到的 4 个主成分，对 PM2.5 做主成分回归.最终，回归所得到的 R^2 为 0.9。回归结果如表 4.3。

表 4.3 主成分回归结果

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	47.9693	0.3457	138.77	<2e-16	***
V2	-4.5105	0.1090	-41.36	<2e-16	***
V3	7.7380	0.1656	46.71	<2e-16	***
V4	6.1652	0.2064	29.88	<2e-16	***
V5	5.8326	0.3111	18.75	<2e-16	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

表 4.3 中我们可以看出，由主成分所进行的回归分析，其回归系数和回归方程均通过检验，且效果显著。由此，验证了我们所得到的主成分的有效性。

结合 5.2 的结论，我们可以得出，对 PM2.5 有较大影响的变量有 PM10、CO、NO_2、SO_2、温度、风速、湿度（这里包含降雨量）、能见度等。而根据所得主成分，我们所分析出对 PM2.5 有较大影响的有天气状况、极端天气、天气湿热状况、空气污染情况和扩散情况以及空气的干湿情况等。

5 LASSO 模型介绍与建立。

5.1 特征选取

用于分析的数据有着各式各样的属性，然而，一些属性可能与分析不相关或是冗余的。因此，在做分析之前(数据预处理)，需要对数据属性进行挑选，去掉不相关的变量，以使得构造出来的模型更好，这个过程被称为特征选取[7]。

LASSO(The Least Absolute Shrinkage and Selectionator operator)是Tibshirani于1996年提出的特征选取算法。LASSO是一种能够实现属性集合精简的算法,并能够准确且稳定地进行特征选取，它通过构造一个惩罚函数获得一个较为精炼的模型；最终使得一些变量的系数为零。其基本思想是在回归系数的绝对值之和小于一个指定常数的条件下，使残差平方和最小化，从而使一些回归系数为0，竟而达到特征选取的目的[8]。

假设有数据$\left(X^i, y_i\right), i=1,2,\cdots,N$，$X^i=\left(x_{i1},\cdots,x_{ip}\right)^T$和$y_i$分别是第$i$个观测值对应的自变量和响应变量，考虑线性回归模型

$$y=X\beta+\varepsilon, \varepsilon\sim N\left(0,\sigma^2 I\right),$$

其中$y=\left(y_1,y_2,\cdots,y_n\right)^T$, $x_i=(x_{1j},x_{2j},\cdots,x_{nj})^T$, $j=1,2,\cdots,d$, $X=(x_1,x_2,\cdots,x_d)$, β为d维列向量，为待估参数。则LASSO的估计为[8]

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\|y-X\beta\|^2, s.t.\sum_{i=1}^{d}\left|\beta_i\right|\le t$$

记为惩罚函数形式为：

$$\min\left\|y-X\hat{\beta}\right\|^2+\lambda\left\|\hat{\beta}\right\|.$$

在本文中，我们选择Lasso来对数据进行特征选取，以得到和PM2.5相关的变量。并用选取得到的变量对数据做回归分析，以得到最终预测结果。

5.2特征选择结果与解释

LASSO对变量的选择有k-折交叉验证（k-fold CV）和C_P两种方法。但k-折交叉验证的随机性，使得其每次的运行结果很难相同。因此，本文选择C_P方法进行变量选择。C_P统计量也是用来评价回归的一个准则。其定义为：如果从k个自变量中选取p个（k>p）参与回归，则

$$C_P=\frac{SSE_P}{S^2}-n+2p, SSE_p=\sum_{i=1}^{n}(Y_i-Y_{pi})^2.$$

据此，选择使C_P最小时系数不为0的变量为最终变量。

我们按照时间顺序，选取所有数据的前4/5作为训练集，后1/5的数据作为测试集；即2013年11月1日至2015年2月3日的数据作为训练集，2015年2月4日至2015年5月31日的数据为测试集。值得提出的是，我们采取的预测方式是用前一天的各指标预测第二天的PM2.5取值，因此，上述训练集和测试集的时间跨度是针对自变量数据而言的。本文通过R软件利用LASSO对数据进行特征选取（如图5.1），

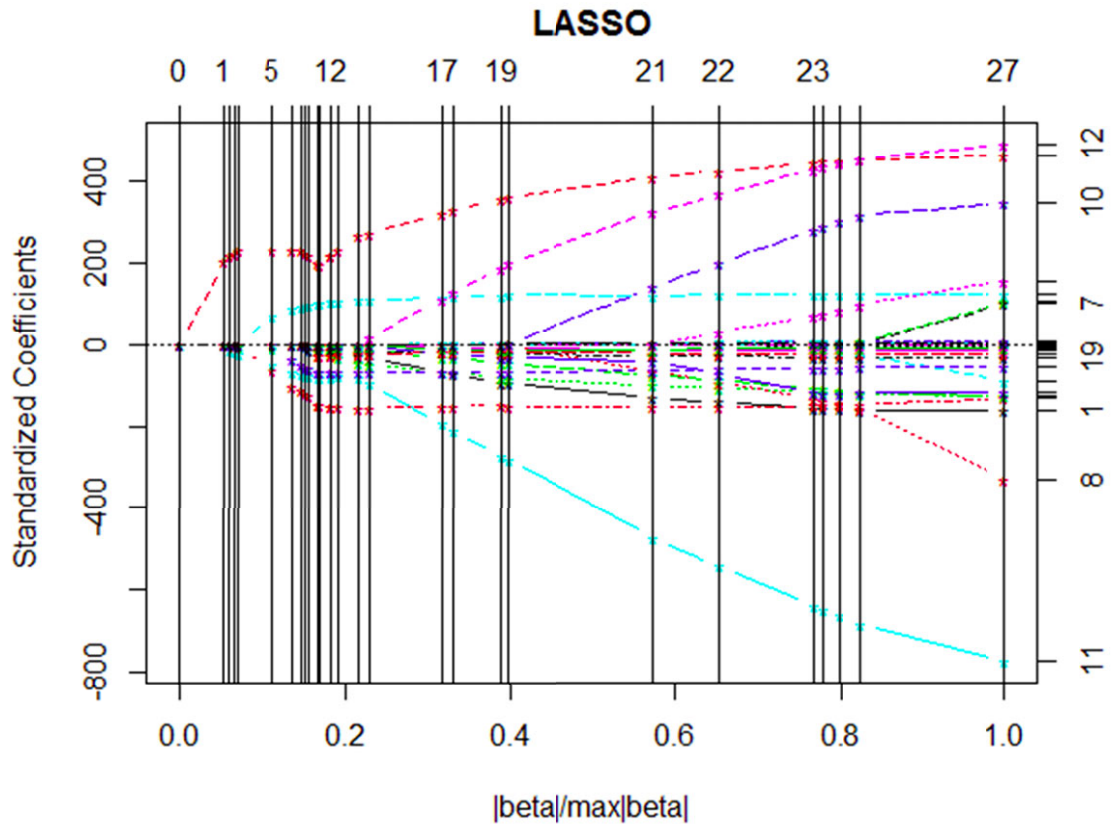

图 5.1 LASSO 特征选取变量结果图

结果见表 5.1，其中系数为 0 的变量(下划线)是被筛除的变量。

表 5.1 LASSO 特征选取变量结果

变量名	系数	变量名	系数
Mean.海平面气压 hPa	0.000000000	最低温度 C	0.000000000
Min.海平面气压 hPa	0.194029924	露点 C	1.224979958
Max.能见度 Km	-0.207139645	MeanDew.PointC	-2.940057947
Mean.能见度 Km	-0.359372509	Min.DewpointC	1.857420813
Min.能见度 kM	-0.220367815	Max.湿度	0.000000000
Max.风速 Km.h	-0.354549727	Mean.湿度	-0.478822318
Mean.风速 Km.h	0.055996188	Min.湿度	-0.216296816
降水量 mm	0.000000000	Max.海平面气压 hPa	-0.464619163
WindDirDegrees	0.001557667	AQI 指数	-0.203848057
PM2.5	0.747146003	Co	0.766170418
PM10	-0.148992644	No2	0.283490149
最高温度 C	0.000000000	平均温度 C	-0.637332337
So2	-0.081914298		

若直接对所有这些变量做一个线性回归，由于变量间存在一定的线性关系，势必会造成多重共线性的问题，而 lasso 则会筛选掉部分变量，使其系数变为 0，从而减少这种共线性，使得变量的系数更加显著。我们可以从表中看出来，系数为 0 的变量有：Mean.海平面气压 hPa，最低温度 C，Max.湿度，降水量 mm 和最高温度 C。

PM2.5 是模型中非常重要的一个指标，从直观上看，前一天的 PM2.5 的浓度，肯定与当天的 PM2.5 的浓度有非常大的关系。在 lasso 中，PM2.5 的系数为 0.747146003，从这里我们可以认为，前一天空气中的 PM2.5 微粒，经过一天的扩散之后，仍然有大约 74%的微粒在空气中，所以，若是前一天 PM2.5 浓度较高，可以预期，第二天 PM2.5 的浓度也将不会显著得减少，需要注意。

对于其他变量，我们可以分成几类来看：

第一类是能见度，可以看出来，能见度的系数为负数，所以说前一天的能见度越高，第二天的 PM2.5 浓度就会较低。这也是可以理解的，PM2.5 是造成雾霭的原因之一，雾霭严重时，能见度就会降低。所以显然能见度与 PM2.5 的浓度时是负相关的关系，而这里的变量是前一天的能见度，从系数来看也是符合这样的规律的。

第二类变量是温度、湿度，可以看出来，温度、湿度的系数是负的，所以说天气潮湿、温度较高时候，PM2.5 的浓度会降低，这也可以印证前面我们做 PM2.5 季节性分析的结论，夏季 PM2.5 的浓度时最低的，而冬季 PM2.5 的浓度水平的是最高的，相对而言，夏季的湿度会大于冬季的湿度。所以可以说，湿度是形成了夏季 PM2.5 浓度低，冬季 PM2.5 浓度高重要原因之一。

第三类是风速，可以看到风速的影响没有像前面几类变量那么明显，在 lasso 中，最高风速的系数是负的，而平均风速的系数是正的，但是，实际上，平均风速的系数约为 0.055，相比于最高风速的系数-0.354，绝对值上要相差一个数量级，所以实际上，前一天风速越高，后一天 PM2.5 的浓度就越低，而且，一阵强风带来的影响要大于持续的微风。

第四类变量是空气中的其他污染物如 Co、No2 等，这些污染物与第二天 PM2.5 的浓度的影响从直观上来看应该是正相关的，然而 So2、PM10 的系数却是负数，究其原因，是因为 So2、PM10 带来的影响体现在 PM2.5 的系数上了，从图 3.6 和图 3.7 中可以看出来，So2、PM10 有非常高的相关性，所以，在回归分析的时候，他们所带来的影响体现在 PM2.5 的系数上了，实际上他们系数的绝对值也是比较小的。

5.3 回归建立与预测结果

利用筛选得到的变量，我们对其进行回归分析。再次指出，我们采取的预测方式是用前一天的各指标预测第二天的 PM2.5 取值，以得出较好的预测结果。**错误!未找到引用源。**给出了预测结果与真实值的对比图。其中**错误!未找到引用源。**中红色所示的是预测值。

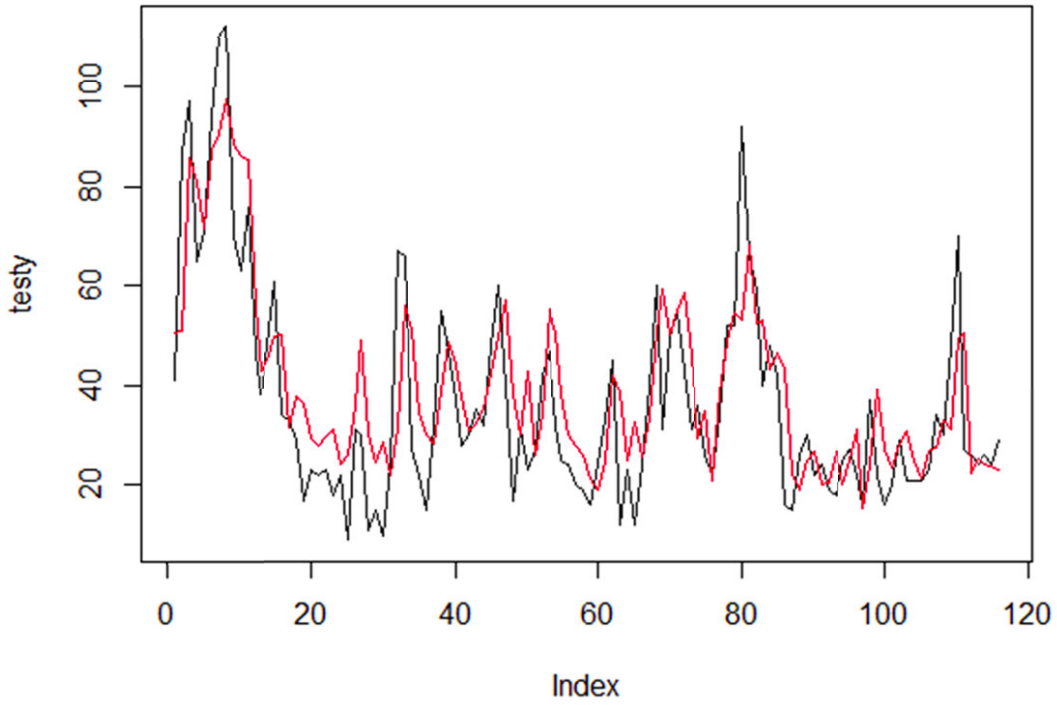

图 5.2 预测值和真实值对比图

从图 5.2 中可以看出来，lasso 模型预测结果大概能够反映 PM2.5 变化的趋势且拟合的较好，能够提早一天的时间去判断 PM2.5 的变化趋势，即根据前一天的空气质量数据、气象数据能够比较准确地去判断第二天 PM2.5 浓度是升高还是降低。

6 总结

在这一部分简要说明一下本文的结论：

1、夏季的 PM2.5 处于一个较低的水平，而其它季节的 PM2.5 相对较为严重。十一，十二，一，二四个月非常突出，冬天是 PM2.5 的一个爆发期。

2、对 PM2.5 有较大影响的有天气状况、极端天气、天气湿热状况、空气污染情况和扩散情况以及空气的干湿情况等

3、对 PM2.5 建立了预测模型，根据模型，可以利用前一天的空气质量数据和气象数据来预测后一天的 PM2.5 浓度。

7 不足与展望

在数据上，由于空气质量数据没有办法找到按照小时记录的历史数据，如果需要按照小时记录下来的空气质量历史数据，就需要设计一个程序，长期地收集数据才可以实现，这次论文写作过程没有办法做到这一点，所以只能退而求其次，选择了以天为单位记录下来的空气质量数据，这样的话，数据包含的信息量大大减少，一天之内 PM2.5 的浓度变化也不可能知道，只能用一个平均数去替代。如果需要做到更加精确的结果，接下来，需要更加多的数据，就是需要用程序长期收集全国的空气质量数据。另外，即使按照天来收集数据，由于 PM2.5 数据最近年来才受到关注，所以在 13 年之前的数据基本上是找不到的，只能从 13 年后半年开始找到数据。这对我们研究 PM2.5 的季节性变化有一定的问题，因为只有一年半左右的数据，数据有可能出现偶然的情况，没有多年的数据作为支撑，得出来的结果可能具有偶然性，不具有通用性。

由于我们收集的变量较多，不可避免的存在一些多重共线性的问题，我们采用了 lasso 模型，lasso 可以处理回归模型的多重共线性问题，但是在最后从模型的系数上看，仍然有一些变量的系数不符合预期的情况，比如说一些其他污染物如 PM10 和 So2 的系数为负，不符合预期，证明系数的显著性还不够高，想要再提高预测的精准度的话，就需要提高系数的显著性，可以再通过其他手段减少变量尝试。

在我们的模型中，预测变量考虑的是前一天的空气质量数据和气象数据。这样做的话有两个因素没有考虑得到，第一个因素是 PM2.5 的排放强度，第二个因素是周边城市 PM2.5 微粒扩散强度。下一步可以把第二个因素考虑在内，对于城市的距离，可以用经纬度来计算，可以根据周边城市 PM2.5 的浓度、风向、风速建立一个扩散模型，如高斯扩散模型等来描述周边城市 PM2.5 微粒扩散到广州的强度，通过把这个因素考虑在内，应该可以使得模型预测起来更加精准。

最后，由于作者的水平有限，需要在限定时间内完成论文，论文中还存在很多的缺陷和漏洞，这些都需要进一步完善。

参考文献：

[1]陈军，高岩，张烨培，杨阳，刘璧婷.PM2.5 扩散模型及预测研究.数学的实践与认识,2014,44(15):16-27.
[2]周锐.中国四分之一国土现雾霾 近半数国人受影响.http://www.chinanews.com/gn/2013/07-11/5032645.shtml
[3]PM2.5.http://baike.baidu.com/link?url=4l0GjtbWhfptpZ77G28DCtxAeresRd4NUrMiiFjo8eSTC2LW4IlQUiFnQSSgYTF_y7hMUQcdWeQtDiQXx3Lm5c2o9lvuczlOX752pZNzabJ3LDjz_KZvciDQVzyhGircmywM2vw8Rm3NBW2IBQXPZ_
[4]徐晓刚，徐冠雷，王孝通，秦绪佳.经验模式分解(EMD)及其应用.电子学报.2014:03.
[5]王婷.EMD 算法研究及其在信号去噪中的应用.哈尔滨工程大学,2010.
[6]薛毅，陈丽萍.统计建模与 R 软件.北京：清华大学出版社，2007.4

[7]韩家炜.数据挖掘概念与技术.北京：机械工业出版社.201
[8]龚建朝，Lasso 及其相关方法在广义线性模型模型选择中的应用，中南大学,2008.
[9] 吴喜之.复杂数据统计方法——基于 r 的应用.北京：中国人民大学出版社,2012.