浙江财经大学

2015 年全国大学生
统计建模大赛
参赛论文

题　目：关于杭州市 PM2.5 的影响因素分析及预测研究[1]

参赛单位：浙江财经大学

参赛队员：杨文青　徐璇　徐雅伦

指导教师：李云霞

2015 年 6 月 28 日

[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”大数据统计建模类研究生组三等奖。

关于杭州市 PM2.5 的影响因素分析及预测研究

摘要

近年来，雾霾是政府和民众关心的热点话题，而雾霾的主要产生原因是空气中存在的悬浮颗粒物即 PM2.5.因此对空气中的 PM2.5 浓度的研究及预测变得十分重要。本文收集了自 2013 年 11 月 1 日至 2015 年 4 月 30 日杭州市的 PM2.5 浓度、其他空气污染物以及天气指标等数据，通过主成分分析、多元回归、向量自回归移动平均模型（VARMAX 模型）探寻了 PM2.5 浓度与其他观测指标之间的关系，并利用 VARMAX 模型和状态空间模型进行预测，最后通过模型比较检验，最后选择 VARMAX 模型作为最终预测模型。

关键词：雾霾　PM2.5　主成分分析　多元回归 VARMAX 模型

一、研究的背景及意义

1.1 背景

PM2.5 又称为细颗粒物，是指环境空气中空气动力学当量直径小于等于 2.5 微米的颗粒物。它能较长时间悬浮于空气中，其在空气中含量浓度越高，就代表空气污染越严重。细颗粒物的标准，是由美国在 1997 年提出的，主要是为了更有效地监测随着工业化日益发达而出现的、在旧标准中被忽略的对人体有害的细小颗粒物。与较粗的大气颗粒物相比，PM2.5 粒径小，面积大，活性强，易附带有毒、有害物质（例如，重金属、微生物等），且在大气中的停留时间长、输送距离远，因而对人体健康和大气环境质量的影响更大。近年来，我国政府也越来越重视 PM2.5 的危害，环保部发布的《环境空气 PM10 和 PM2.5 的测定重量法》从 2011 年 1 月 1 日开始实施。PM2.5 值已经成为一个重要的测控空气污染程度的指数。

近些年来，随着工业化与城市化的快速发展，化石燃料不断地消耗，从而导致了空气质量的严重恶化。雾霾已经由局部的环境因素影响变为全国范围内的环境灾害。大范围与高频率的雾霾天气，给人们的正常生活带来了诸多的不便，同时也对人们的身体健康造成不良影响。这成为政府与民众共同关注的焦点，也亟需研究与解决。

今年，一个关于雾霾的调查视频《苍穹之下》引发了群众的热议，人们再次深刻认识到雾霾危害的严重性以及治理雾霾的紧迫性。而杭州作为全国著名的旅游城市之一，以环境优美著称，却也逃不出雾霾的阴影。据统计，杭州全年有 200 多天是雾霾天气，雾霾的防治已迫在眉睫。雾霾产生的主要原因就是人类活动排放的大量细颗粒物（PM 2.5）超过了大气循环能力和承载度，导致细颗粒物浓度的持续积聚。雾霾不仅严重影响人们的身体健康，而且给出行带来了诸多不便，对城市的发展造成了恶劣的影响，不少地区已将“雾霾天气”作为灾害性天气现象进行预警预报。因此治理雾霾义不容缓。

“治污先治源”、“治病先治根”，要想减少雾霾天气，首先需要找出引起雾霾天气的根本原因，这样才能快速有效的解决雾霾带来的各种问题。 PM2.5 来源广泛，成因复杂，主要是人为排放，包括化石燃料（煤、汽油、柴油、天然气）和生物质（秸秆、木柴）等燃烧、道路和建筑施工扬尘、工业粉尘等污染源直接排放的颗粒物，也包括由一次排放出的气态污染物（主要有二氧化硫、氮氧化物、挥发性有机物、氨气等）转化生成的二次颗粒物。因此，PM2.5 与其他污染物（NO，SO2，CO 等）之间存在内在联系。此外，PM2.5 还受到天气状况（温度，湿度，气压等）的影响。根据历史数据，我们可以研究 PM2.5 指数的影响因

素，进而建立统计预测模型，对未来一段时期 PM2.5 的值进行预测。对 PM2.5 的有效预测有利于政府及公众及时采取相应的措施，避免造成一些不必要的损失，减少其带来的危害。我国尚未研究开发出预报 PM2.5 浓度这方面的应用软件，因此，对 PM2.5 进行预测有重要的实际意义。

由于众多因素影响着 PM2.5 浓度，统计分析中的多元线性回归模型比较适合处理这种关系。并且国内的诸多文献已将其用于各方面的研究。本文建立常规的多元回归模型分析 PM2.5 与空气中的化学成分、温度、湿度等的关系，发现残差存在着自回归现象。结合时间序列分析，建立向量自回归移动平均模型（VARMAX 模型）以及状态空间模型分析 PM2.5 的影响因素，并给出杭州市 PM2.5 日均值浓度的预测。这是本文的创新点，具有一定的理论意义和实际意义。

1.2 国内外研究现状

20 世纪 90 年代以来，对可吸入颗粒物的研究主要集中在以下几个方面：颗粒物的化学组成、物理特征、存在状态等及其在大气中的变化、大气颗粒物对人体健康的影响以及大气颗粒物的气候效应、能见度的影响、源解析等。近年来，大多数城市的雾霾天气越来越严重，不仅严重影响人们的身体健康，而且给出行带来了诸多不便，对城市的发展造成了恶劣的影响，不少地区已将“雾霾天气”作为灾害性天气现象进行预警预报。国内外学者越来越重视空气颗粒物的有效预测研究。

对空气颗粒物进行有效测度首先必须了解其基本特性，包括数量浓度、质量浓度、大小及形状、颗粒的聚集特性、有机和无机化学组分、矿物组成、可溶性等。其次，对其来源的探究是了解其成因的根本依据。此外，颗粒物的物理和化学性质与粒径密切相关，所以空气颗粒物的时空分布规律也称为人们的关注焦点。不同地区不同时间的污染源不同，其粒度分布规律也不相同。

近年来，国内外对空气颗粒物的来源进行了大量的研究，根据污染源不同的分布特征，确定颗粒物与污染源的关系，为治理空气污染提供科学依据。

目前对 PM2.5 来源的研究以受体模型和扩散模型为主。受体模型是通过分析空气环境受体以及污染源的性质，对受体的污染源进行定性识别，对储污染源的分担率进行定量计算。受体模型无需考虑污染源的地形、气象及排放条件等因素，气颗粒物的迁移过程也无需追踪，因此被广泛运用。学者们对受体模型进行了大量研究，探究了多元线性回归，主成分分析，因子分析，投影寻踪回归法，化学质量平衡等多种方法。Chan 等通过受体模型分析得出澳大利亚布里斯班的 PM1025%来自扬尘，10%来自二次污染物和碳元素，13%来自机动车尾气，12%来自海盐，11%来自钛化合物和富钙。张晶等采用化学质量平衡模型探究了北京市空气颗粒物源解析，结果表明燃煤、汽车尾气和尘土是北京市的三大污染源。王灿

星等利用同样的方法研究了杭州市区空气颗粒物 PM10 来源，其主要的污染源为汽车尾气、燃煤尘、道路建筑尘和二次粒子，它们占到了 PM10 总量浓度的 90%。韩力慧等运用同位素示踪法，估算出北京矿物气溶胶中本地源与外来源的相对贡献值，探究出在春节、冬季外来源对北京地区的矿物气溶胶的贡献要高于夏季。

为了更好地做好空气颗粒物的预报问题，国内外学者开展了长期且广泛的研究，以传统的统计模型和神经网络模型为主。其中，统计模型中的自回归移动平均模型(ARIMA)和多元线性回归模型(MLR)被广泛的应用于空气污染预测分析，但是由于这两种模型均为线性模式，从而导致非线性的关系很难被准确预测。然而，神经网络模型作为一种非线性工具，也被广泛应用于污染物预测领域。在空气污染物预测模型中，对 PM2.5、PM10 进行预测的文献却并不是很多，但也有部分研究取得了一定的成果。Grivas 等利用 NN 模型以希腊雅典地区前一天的 PM10 浓度、风速、温度、相对湿度等作为解释变量，对每小时的 PM10 浓度进行预测，预测的平均误差大约在 25%到 30%之间。Perez 等利用 NN 模型对圣地亚哥、智利的每小时 PM2.5 浓度进行预测，该模型以风速以及湿度作为解释变量，预测误差在 30%到 60%之间。

对于中国来说，到目前为止有关 PM2.5 浓度的研究还比较少，对其进行定量分析预测的文章更是稀少。本文将探究 PM2.5 的影响因素，并利用收集到的数据对 PM2.5 的值进行定量的分析与预测，提出可靠的预测模型，为 PM2.5 浓度的预警工作提供依据。

二、研究目标与内容

2.1 目标

1、探寻杭州市 PM2.5 与温度、湿度、可见度以及空气污染物浓度 SO_2 浓度、CO 浓度、NO_2浓度之间的关系；

2、建立预测模型以预测未来几天的杭州市每天的 PM2.5 浓度。

2.2 内容

1、采用主成分分析消除温度、湿度、可见度、以及空气污染物 SO_2 、CO、NO_2浓度的强相关性；

2、建立简单的多元线性回归模型拟合 PM2.5 浓度与温度、湿度、可见度以及空气污染物浓度 SO_2 浓度、CO 浓度、NO_2浓度之间的关系，检验残差自相关现象；

3、结合时间序列相关知识，建立 VARMAX 模型分析 PM2.5 浓度并给出预测；

4、建立空间状态模型预测 PM2.5 浓度并给出预测；

5、最后比较两种预测模型的预测性能。

三、符号说明

为了便于问题的表达和研究，我们运用一些符号来代替问题中涉及的一些基本变量，其他一些变量我们将在文中陆续说明。

表 1 符号说明

符号	意义
$X_{ij}\quad i=1,\cdots,8$	第 i 个观测指标在第 j 天的观测值
P_j	杭州在第 j 天的 PM2.5 浓度
$Y_{ij}\quad i=1,2,\cdots,4$	第 i 个新解释变量（主成分 i）在第 j 天的观测值
L	滞后算子
X_1	日均温度
X_2	日均湿度
X_3	大气压强
X_4	日均可见度
X_5	日均风速
X_6	日均 CO 浓度
X_7	日均 NO2 浓度
X_8	日均 SO2 浓度

四、数据的描述及处理

4.1 数据来源

本文从相关网站上收集了 2013 年 11 月 1 日至 2015 年 4 月 30 日杭州市的每日 PM2.5 浓度以及影响 PM2.5 浓度的影响因素的观测数据。数据包括了 9 个变量，变量的名称分别为：PM2.5、日均温度、日均湿度、大气压强、日均可见度、日均风速、日均 CO 浓度、日均 NO2 浓度和日均 SO2 浓度。

4.2 数据预处理

首先探索 PM2.5 与收集到的其他空气污染物以及气象因素的关系，其散点图如图 4-2-1 所示，从图 4-2-1 中可以初步发现杭州的 PM2.5 浓度与空气污染物 SO2,CO,NO2 浓度之间有着较强的正相关，与湿度、风速之间有着较强的负相关。

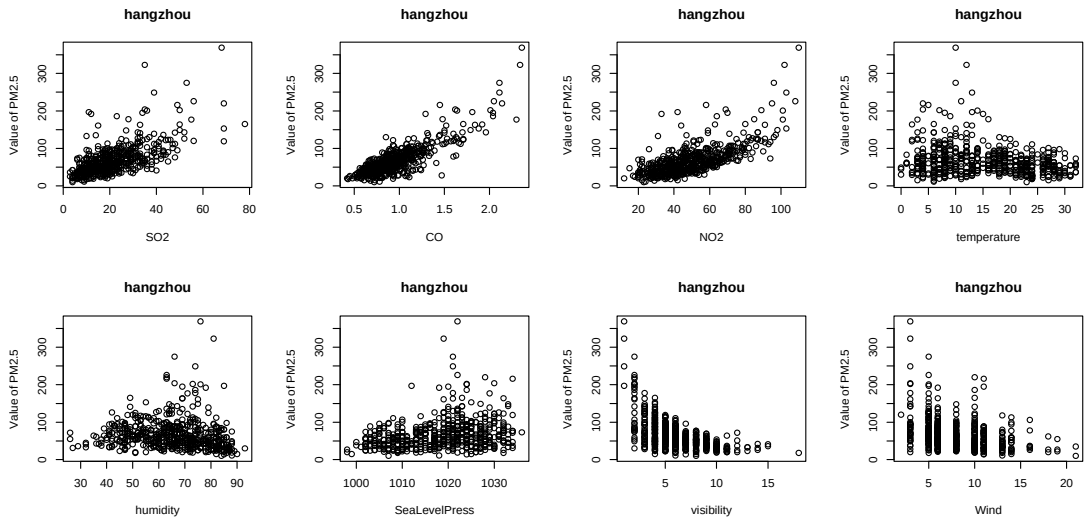

图 4-2-1 PM2.5 与其他空气污染物以及气象因素的散点图

然后，通过观察因素的两两相关图如图 4-2-1,可以发现这些解释变量之间存在着一定的相关性，这可能是由于燃煤、汽车尾气等都会同时产生 SO_2、CO_2、NO，解释变量之间存在着一定的相关性，这一现象很符合常理。 因此，当变量选取较多时，很有可能出现多重共线性的情况。我们可以通过条件数来检测多重共线性。条件数的定义为：

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

其中，λ 为 $X^T X$ 的特征值(代表自变量矩阵)。一般地认为，当 κ>15时，则有共线性问题，而 $\kappa > 30$ 时，则有严重的共线性问题。我们通过 R 统计软件计算出的

数据的条件数$\kappa = 4846.305$,结果表明自变量之间的共线性问题相当的严重,所以接下来采用主成分分析方法要对这些解释变量进行处理以消除共线性, 使得新的解释变量之间的相关性降至最低。

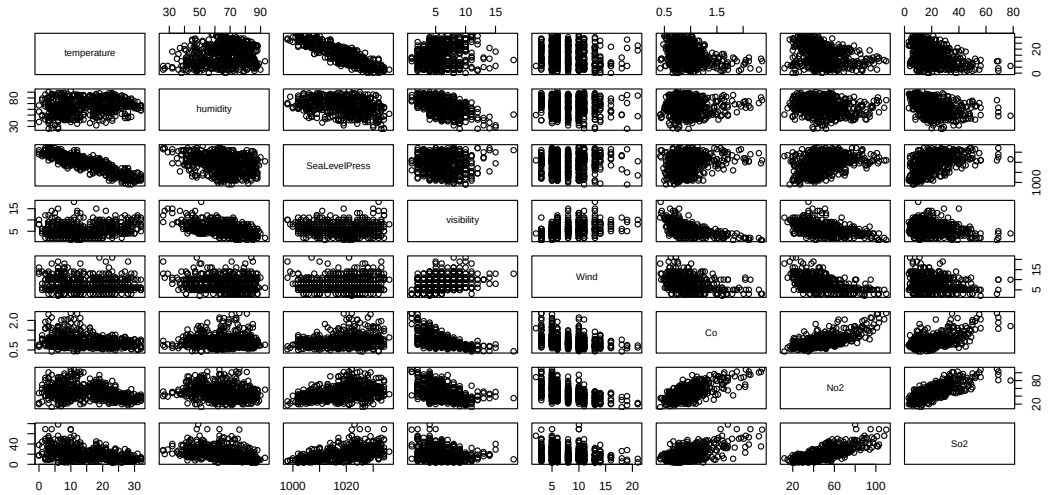

图 4-2-2 两两相关图

4.3 主成分分析

主成分分析[1]时间的观测值的思想是设法将原来众多具有一定相关性的指标, 重新组合成一组新的相互无关的综合指标, 并代替原来指标。数学上的处理是对原来的 p个指标做线性组合。设主成分分析的成分 Y_i与原来变量 X_i之间的关系：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \vdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{cases}$$

这里， u_{ij}为第 i个成分 Y_i和第 j个成分 X_j之间的线性相关系数。且 Y_i与 Y_j是互不线性相关的。

4.3.1 主成分分析建模的步骤

Step1:将原始数据标准化，得到标准化数据矩阵：

$$X^{*}=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{t1} & X_{t2} & \cdots & X_{tp} \end{bmatrix}$$

Step2:建立变量的相关系数阵：$R=\left(r_{ij}\right)_{p*p}=X'X$

Step3:求 R的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$以及相应的单位特征向量：

$$u_1=\begin{bmatrix} u_{11} \\ u_{21} \\ \vdots \\ u_{p1} \end{bmatrix},u_2=\begin{bmatrix} u_{12} \\ u_{22} \\ \vdots \\ u_{p2} \end{bmatrix},\cdots,u_p=\begin{bmatrix} u_{1p} \\ u_{2p} \\ \vdots \\ u_{pp} \end{bmatrix}$$

Step4:通过各主成分累计贡献率确定主成分，其中前 m 个主成分包含的数据信息量不低于 80%时，可取 m 个主成分来反映原指标。通过 R 软件求解得到各主成分累计贡献率如表 4-3-1 所示。

表 4-3-1 主成分累计贡献率

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
累计贡献率	0.639	0.890	0.975	0.992	0.997	0.998	0.999	1

由表 4-3-2 中的累计贡献率可知，主成分为 4 时，累计贡献率达到了 99.2%，所以本文选取 4 个主成分，采取这 4 个新的解释变量来解释原来的所有变量。同时也给出原变量在新的解释变量所占权重矩阵如表 4-3-2。

表 4-3-2 原变量在新的解释变量所占权重矩阵

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
X_1	-0.359	0.356	0.437	-0.224	-0.123	-0.109	-0.109	-0.607
X_2	-0.128	0.486	-0.547	0.221	-0.181	0.345	-0.365	-0.339
X_3	0.352	-0.415	-0.319	0.318	0	-0.110	0.396	-0.574
X_4	-0.278	-0.480	0.357	0.177	0.268	0.317	-0.512	-0.314

X_5	-0.226	-0.381	-0.405	-0.713	-0.212	-0.180	-0.183	-0.143
X_6	0.416	0.247	0	-0.455	0.700	0.195	0	-0.163
X_7	0.476	0.131	0.232	0	-0.226	-0.558	-0.546	-0.203
X_8	0.450	-0.108	0.247	-0.230	-0.540	0.613	0	0

根据表 4-3-2 ，把第一主成分命名为空气质量因素；第二主成分命名为天气湿雾程度 ;第三主成分命名为其他天气综合因素 ;第四组成份为天气对流强度。这样我们就得到了新的四个解释变量，并将这四个新的解释变量记为$Y_1,\cdots,Y_4$。

五、模型的建立与求解

首先, 对数据建立常规的多元回归模型来分析杭州市 PM2.5 与空气中的化学成分、温度、湿度等的关系，发现残差项存在着自相关，因此建立多元回归模型是不合理的。然后，结合多元时间序列的相关知识，建立向量自回归移动平均模型（VARMAX 模型）分析杭州市 PM2.5 浓度与其影响因素之间的关系，并运用状态空间模型预测杭州市 PM2.5 浓度。

5.1 多元回归模型

记P_j为第 j 天的 PM2.5 浓度，$Y_i, i=1,2,3,4$为第 i 个新的解释变量。则多元回归模型记为：$P_j=\beta_0+\beta_1 Y_{1j}+\beta_2 Y_{2j}+\ldots+\beta_4 Y_{4j}+e_j$,其中,$e_j i.i.d.\left(0,\sigma^2\right)$。

表 5-1 多元回归的估计系数表

变量	系数	标准误	T 值	P 值
截距	538.059	92.292	5.47	6.74e-08
Y_1	2.686	0.176	15.27	<2e-16
Y_2	1.545	0.117	13.20	<2e-16
Y_3	-0.933	0.206	-4.53	7.4e-06
Y_4	3.628	0.517	-7.02	6.8e-12

从表 5-1 中，我们可以看出检验的 p 值都是远远小于 0.05 的，也就是回归系数是非常显著的。且 R^2 为 0.5937，说明建立多元回归方程比较显著，回归方程可记为：

$$P_j = 538.059 + 2.686Y_{1j} + 1.545Y_{2j} - 0.933Y_{3j} - 3.628Y_{4j} + \varepsilon_j$$

接下来考虑残差序列之间是否满足独立同分布于正态性。为了检验正态性，我们绘制了残差序列的直方图，如图 5-1-1 所示。

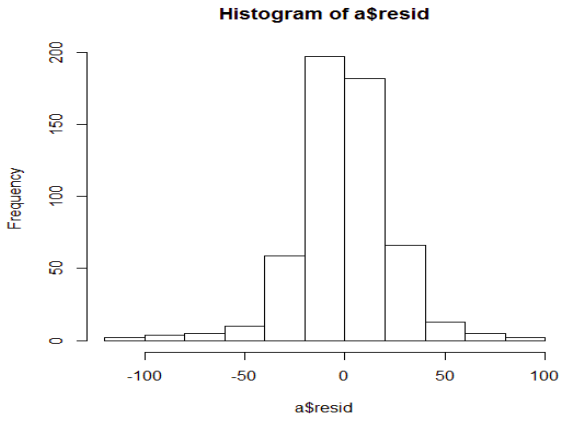

图 5-1-1 残差序列的直方图

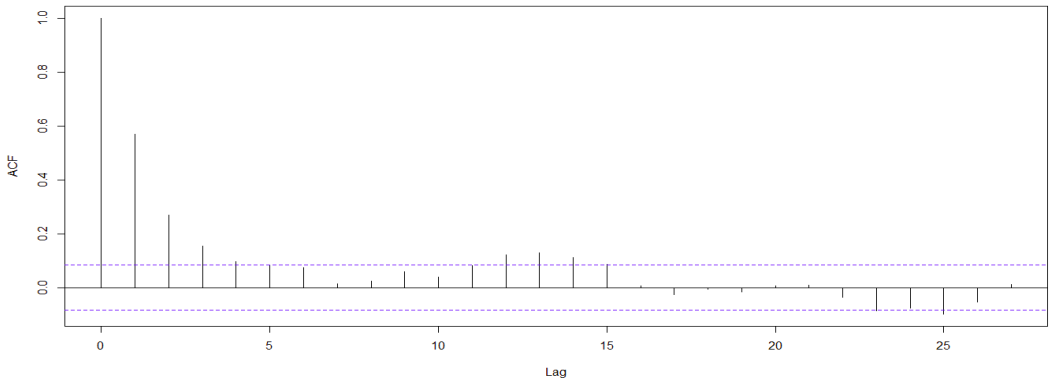

图 5-1-2 残差自相关图

从图 5-1-1 可以看出，残差序列基本上是服从正态分布的。接下来我们检验残差序列是否独立。图 5-1-2 为残差序列自相关图，直观上我们可以看出，在延迟阶数为 25 下的残差序列并没有很快地衰减为 0，所以可以初步判断残差序列不满足独立性的假设。为了进一步判断残差序列是自相关的，我们对残差序列进行 Box-Ljung 检验，得出的 p 值为 1.38e-07，显著地小于 0.05，所以可以判断残差序列不满足独立性的条件。

综上说明采用普通的多元回归模型不足以很好的拟合原始数据之间的关系。

为了更好的说明 PM2.5 与其影响因素之间的关系，我们选择建立多元自回归移动平均模型。

5.2 向量自回归移动平均模型（VARMAX 模型）

5.2.1 VARMAX 模型的简介

当 VARMAX 模型是针对向量时间序列而言,则称为向量自回归模型(VAR),自回归模型中的回归系数通常采用最小二乘的方法进行求解得到。另一类自回归模型为自回归滑动平均模型（ARMA）,在此类回归问题中由于考虑到了残差项。最后，在 VARMA 模型的基础上，引入外生变量即 VARMAX 模型。本文就是运用 VARMAX 模型考察空气中的化学成分、温度、湿度对 PM2.5 浓度的影响。VARMAX 模型建立的步骤如下：

Step1: 检验所有的变量构成的时间序列是否平稳。根据绘制各个变量的自相关图初步判断稳定性，然后通过 Box-Ljung 检验进一步判断稳定性。图 5-2-1 为 PM2.5 浓度的时序图，图 5-2-2 为四个新解释变量的自相关图。表 5-2-1 为 Box-Ljung 检验的结果。

Step2: 对原始序列进行差分处理，将其转为平稳序列。我们对四个新的解释变量进行差分处理后，进行单位根检验，即检验这些变量构成的时间序列是否平稳。得到的单位根检验结果如表 5-2-2 所示。

Step3: 对差分后平稳的时间序列进行协整检验，即差分后的时间序列之间是否存在长期关系。结果如表 5-2-1 所示。

Step4: 对差分后的序列建立恰当的 VARMAX 模型：首先 P_t 是输出变量（处于应变量的地位），也就是将杭州 PM2.5 浓度时间序列作为因变量，ΔY_t 是 4 维的解释变量即 $\Delta Y_t=\left(\Delta Y_{1t},\cdots,\Delta Y_{4t}\right)$，$e_t$ 是 k 维不可观测的扰动过程。A,B,C 为适当维的后移滞后算子 L 的矩阵。因此本文的向量自回归移动平均模型（VARMAX 模型）可以用下式表达

$$A\left(L\right)\Delta P_t=B\left(L\right)e_t+C\left(L\right)\Delta Y_t$$

或写成

$$A\left(L\right)\Delta P_t=B\left(L\right)e_t+C_1\left(L\right)\Delta Y_{1t}+C_2\left(L\right)\Delta Y_{2t}+C_3\left(L\right)Y_{3t}+C_4\left(L\right)\Delta Y_{4t}$$

其中后移滞后算子的阶数可以通过 AIC 信息准则以及残差自相关图来判断。

5.2.2 模型的求解与分析

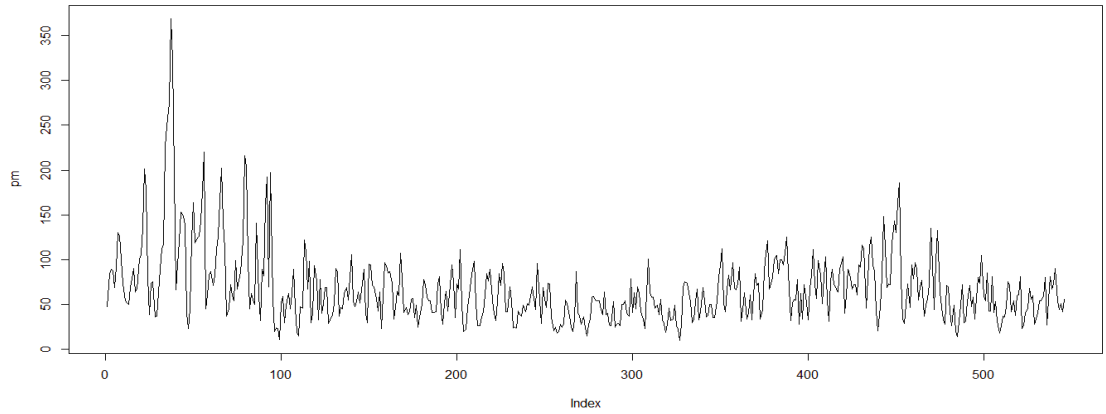

图 5-2-1 PM2.5 浓度的时序图

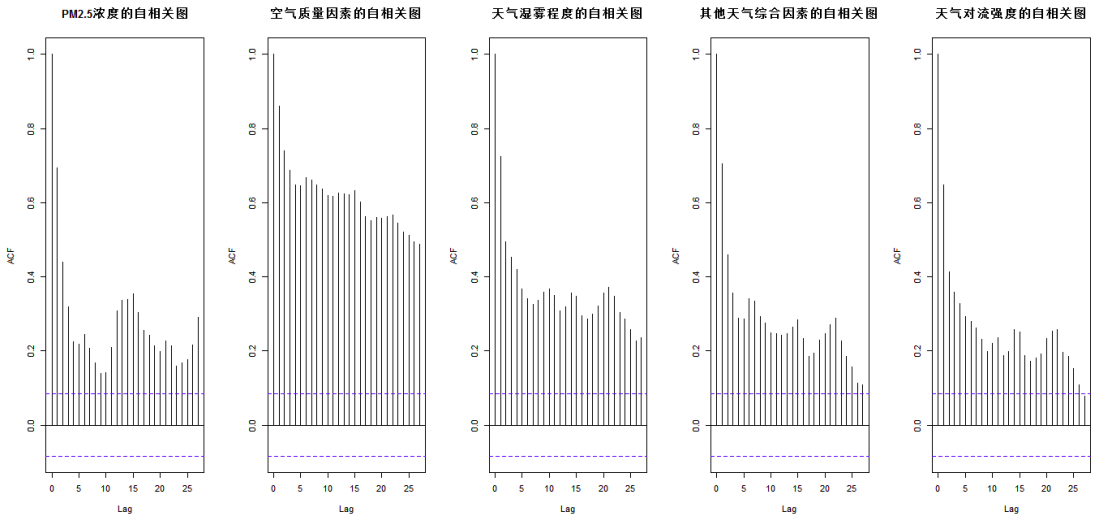

图 5-2-2 PM2.5 浓度与新的解释变量的自相关图

观察图 5-2-2，可以看出在 25 个滞后项各个变量的自相关图仍没有很快地衰减为 0，直观上可以判断这些变量是不平稳的。为了更精确地说明这些时间序列不是平稳的，我们对此作了 ADF 平稳性检验，结果如表所示 5-2-1 所示。

表 5-2-1 平稳性检验的结果

变量	统计量	P 值
PM2.5	94.082	6.104e-10
Y1	92.733	1.019e-09

Y2	101.04	4.195e-11
Y3	86.448	1.078e-08
Y4	90.441	2.423e-09

观察表 5-2-1，我们可以看出 p 值是显著小于 0.01 的，也就是说，在 1%的显著水平下，有理由拒绝原假设（时间序列是平稳的），即 PM2.5 和四个新的解释变量是不平稳的。为了模型的需要，对这些时间序列进行差分处理，然后进行 ADF 单位根检验，并把检验的结果与差分前的结果进行比较，结果如表 5-2-2 所示。

表 5-2-2 ADF 单位根检验结果

	Y1	Y2	Y3	Y4
差分前的检验统计量	-0.3522	-0.0031	-0.1512	-0.1666
差分后的检验统计量	-14.1681	-13.533	-14.6512	-14.1213

在 1%显著水平下的单位根检验统计量的临界值为-1.62，观察表 5-2-2 我们可以看出，差分前的统计量均大于-1.62，说明没有足够的理由拒绝原假设(差分前的时间序列有单位根)，这意味着差分前的时间序列是非平稳的；而差分后的检验统计量都小于-1.62，拒绝原假设，说明差分后的时间序列满足非平稳性。

表 5-2-3　格兰杰协整检验

变量	统计量	P 值
LN(PM2.5)对 ∇Y_1	2.179	0.9819
LN(PM2.5)对 ∇Y_2	2.1612	0.9704
LN(PM2.5)对 ∇Y_3	2.2072	0.9924
LN(PM2.5)对 ∇Y_4	2.1767	0.9809

观察表 5-2-3，可知变量之间没有协整关系，所以没有必要选取误差纠正模型.

结合取对数后的 PM2.5 浓度，本文选取二阶后移滞后算子，利用最小二乘法拟合 VARMAX 模型，结合 R 软件，输出的拟合矩阵结果如表 5-2-4，同时给出了该拟合模型的残差自相关图如图 5-2-3 所示，由图可知残差序列是满足独立性的假设条件的。

表 5-2-4 VARMAX 模型的拟合矩阵

算子	表达式
$A(L)$	$1+0.152L+0.092L^2$
$B(L)$	1
$C_1(L)$	$2.576+0.395L$
$C_2(L)$	$1.298+0.851L$
$C_3(L)$	$-0.826-0.209L$
$C_4(L)$	$-3.215-1.958L$

把表 5-2-4 中的结果代入 VARMAX 模型中，整理后可以的出杭州市的 PM2.5 浓度分析预测模型，可以写为：

$$\begin{aligned} P_t = {} & 0.848P_{t-1}+0.060P_{t-2}-0.092P_{t-3} \\ & +2.577Y_{1,t}-2.182Y_{1,t-1}-0.395Y_{1,t-2} \\ & +1.298Y_{2,t}-0.447Y_{2,t-1}-0.850Y_{2,t-2} \\ & -0.826Y_{3,t}-1.035Y_{3,t-1}+0.209Y_{3,t-2} \\ & -3.215Y_{4,t}-5.173Y_{4,t-1}+1.958Y_{4,t-2}+\varepsilon_t \end{aligned}$$

结论：从这个式子里我们可以分析得到杭州当天的 PM2.5 浓度与前一期的 PM2.5 浓度有着很高的相关性，同时与当天的空气质量因子、天气湿雾程度有着较强的正相关。与此同时它与当天的天气对流强度有着较强的负相关，这与其他的一些研究中现实，对流强度强有利于空气中的污染成分的消散。

而且模型的参数数估计都较为显著，接下来我们看看残差序列是否满足独立性的假设。画出了用 VARMAX 模型拟合杭州市 PM2.5 所得的残差序列的自相关图，

如图 5-2-3 所示。

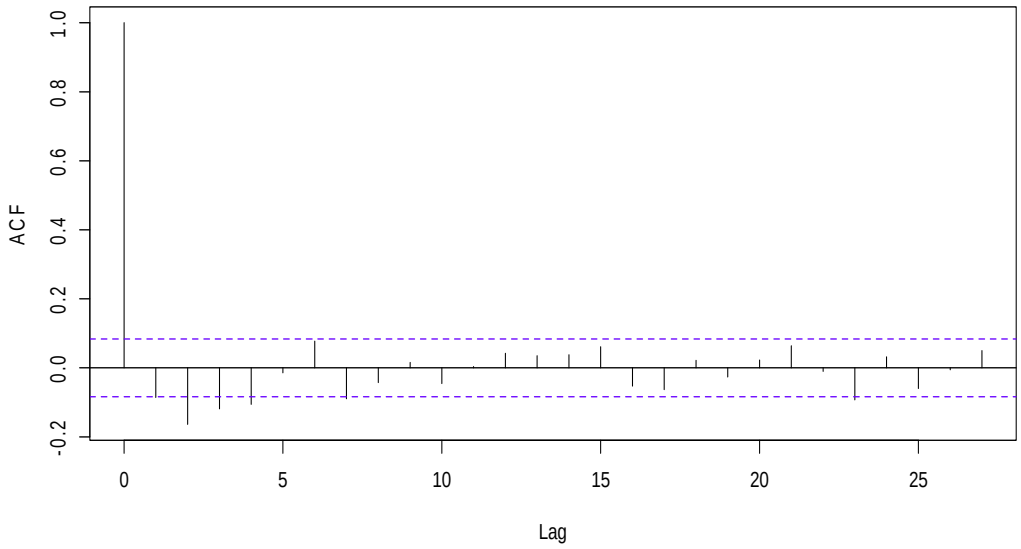

图 5-2-3 残差自相关图

观察选取后移滞后算子的阶数为 2 时的残差自相关图,发现残差自相关系数在滞后项为 4 后基本上都落在 2 倍的标准差之内，而且很快衰减为 0,说明残差项不存在时间序列相关。也说明了 VARMAX 模型的估计结果是十分稳健的。

总而言之,本文使用 VARMAX 模型来拟合杭州市的 PM2.5 浓度与其影响因素之间的关系是合理的。

5.3 状态空间模型

5.3.1 状态空间模型简介

状态空间模型是一种结构模型，基于状态空间分解模型的时间序列预测。建立一个独立于时间的线性状态空间的模型为

$$Z_t = FZ_{t-1} + GY_t + Qn_t$$
$$P_t = HZ_t + Re_t$$

其中，Z_t 为不可预测的 n 维状态变量，F 为状态转移矩阵，G 为输入矩阵，H 为输出矩阵，Q 系统噪声矩阵，n_t 为系统噪声，R 为输出（测量）噪声矩阵，Y_t 为 3 维的解释变量，由 3 个解释变量的时间序列构成，P_t 为 PM2.5 浓度的时间序列。

5.3.2 模型的求解

通过 R 软件里的 dse 包求解得到系数矩阵为

$$F=\begin{bmatrix}0, & -0.092\\ 1, & -0.1517\end{bmatrix}\ G=\begin{bmatrix}0.395, & 0.851, & -0.208, & -1.958\\ 2.577, & 1.297, & -0.826, & -3.214\end{bmatrix}$$

$$H=[0,1]\ K=\begin{bmatrix}-0.092\\ -0.152\end{bmatrix}$$

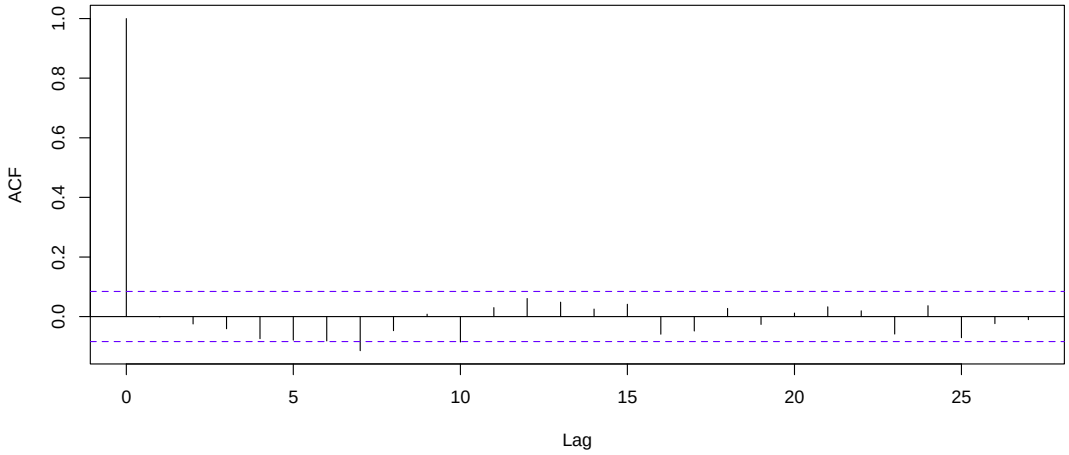

图 5-3-1 残差序列的自相关函数图

由 5-3-1 可知, 采用状态空间模型拟合拟合 PM2.5 浓度与解释变量之间的关系是恰当的。

5.4 模型预测性能评价

预测是对未来的估算, 所以它势必会与客观事实之间存在着差距, 而预测误差这一标准常常被用来衡量此误差。本文采用的三种误差标准对模型进行评价, 其定义如下：

⑴平均绝对误差（Mean Absolute Error, MAE）

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_t-\hat{y}_t|$$

⑵均方误差（Mean Square Error, MSE）

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(y_t-\hat{y}_t\right)^2$$

⑶平均百分比误差(Mean Absolutely Percentage Error, MAPE)

$$MSPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|\times 100\%$$

其中y_t和$\hat{y}_t$分别为预测期内t时刻的观察值和预测值，n为预测期的长度。

下面我们就上述建立的VARMAX模型和状态空间模型，对杭州2015年4月最后3天的PM2.5浓度进行预测，得到的预测结果表5-4-1。

表5-4-1 VARMAX模型预测值与实际值的对比

预测时间	实际值	预测值	MAE	MPE	MSPE
2015/4/30	56	53.34	2.66	7.07	0.047
2015/4/29	42	45.23			
			5.21	27.19	0.103
2015/4/30	56	48.8			
2015/4/28	51	55.92			
2015/4/29	42	48.45	5.18	27.7653	0.106
2015/4/30	56	60.12			

表5-4-2 状态模型预测值与实际值的对比

预测时间	实际值	预测值	MAE	MPE	MSPE
2015/4/30	56	50.02	5.98	35.76	0.016
2015/4/29	42	38.76			
			7.05	64.22	0.135
2015/4/30	56	45.14			
2015/4/28	51	56.32			
2015/4/29	42	50.98	5.84	39.81	0.124
2015/4/30	56	59.26			

对比表5-4-1与5-4-2的MAE,MPE及MSPE的值,可以得知采用VARMAX模型预测PM2.5浓度的效果会更好。而后，我们把一部分数据作为训练集，另一部分数据作为预测集，进行整体预测得到了如图5-4-1。图中黑色线表示实际值，蓝色虚线表示VARMAX模型的预测值，红色虚线表示状态空间模型的预测值。观察图5-4-4,发现上述两种模型的对PM2.5浓度预测趋势与实际值的趋势基本一致。结合表5-4-1以及5-4-2的结果，我们认为VARMAX模型的预测性能更好一些。

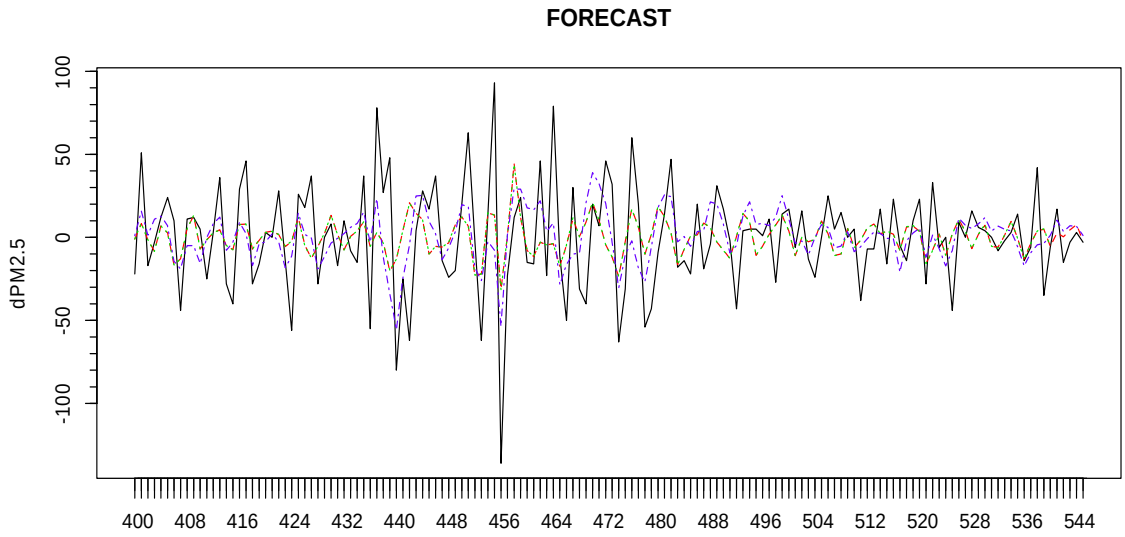

图 5-4-1 预测曲线

六、结论和展望

6.1 结论及建议

本文的主要内容是对影响 PM2.5 浓度的影响因素进行分析，也考虑了时间上的滞后关系，同时本文利用不同的预测模型对杭州市 PM2.5 浓度进行短期预测，并选出更好的预测模型，以便提供有效，精确的 PM2.5 浓度的预测分析，从而保障杭州市民的身体健康与正常的生活。

1、从静态关系来看，杭州的 PM2.5 浓度与 CO、NO2、SO2 即我们定义的空气质量因子有着很强的相关度，说明杭州的雾霾影响因素很大情况下由于尾气排放以及工业废弃。

2、同时杭州的 PM2.5 还与空气湿度情况有关，天气湿度大，天气对流差的天气，PM2.5 浓度会比较高。由于杭州在冬春常出现这种天气，所以建议在这个阶段，因加大对杭州的车流量的控制、以及周边工厂的废弃排量。

6.2 展望

本文使用的向量自回归模型（1）不需要假设 PM2.5 浓度与影响其变化的因素之间背后的正确结构以及各变量之间的潜在关系；(2) 向量自回归模型重在关注时间序列的潜在相关关系以及动态结构；(3) 向量自回归模型将每个自变量回归到其自身及其他变量的过去值的方程系统。但是该模型也存在一些缺点：(1) 向量自回归模型最大的缺点就是过度参数化，即便是较小的向量自回归模型，也包含了数量很大的回归参数。一些文献中也指出，大量的参数导致他们在向量自

回归模型中被无效的估计。(2) 本文仅考虑杭州市的 PM2.5 浓度及其影响因素之间的关系，未考虑其他地区的 PM2.5 浓度对杭州市的影响，即忽略了空间地域对杭州市 PM2.5 浓度的影响。

参考文献

[1] 姜启源，叶俊.时间序列(第三版) [M]. 北京：高等教育出版社，2003.

[2] Patrick T. Brandt ,John T. Williams 著.辛济云译. 多元时间序列模型[M].上海：上海人民出版社 2012.

[3] Johnathan D。Cryer, Kung-Sik Chan 著，潘宏宇等译.时间序列分析及应用[M].北京：机械出版社,2011.

[4] Norval D.Glenn 等著，吴晓刚主编.纵贯数据分析[M].上海：上海人民出版社和格致出版社.2011.

[5] Chan YC,Simpson R.W.,Mctainsh G.H.,et al .Source apportionment of visibility degradation problems in Brisbane(Australia) using the multiple linear regression techniques [J].Atmospheric Environment ,1999,33:3237-3250

[6] Sofuglu S.C.,Sofuglu. S A.,Birgili S., Tayfur G.. Forecasting ambient air SO2 concentrations using artificial neural networks [J]. Energy Sources Part B-Economics Planning and Policy, 2006,1:127-136.

[7] Paschalidou K.,Kassomenos P.,Bartzokas A..A comparative study on various statistical techniques predicting ozone concentrations: implications to environmental management [J].Environ Monit Assess 2009;148:277-289.

[8] Kasparisan J.,Frejafon E.,Rambaldi P.,et al Characterization of urban aerosols using SEM-microscopy,X-Ray analysis and Lidar measurements [J]. Atmospheric Environment,1998,30:2957-2967.

[9] 李赋莲，管峰，蒋建华．浅析中国 PM2.5 现状及防控措施[J].能源与节能.2012(06)

[10] 吴喜之.复杂数据统计方法--基于 R 的应用[M].北京：中国人民大学出版社.2012(10)

[11] 洪永森.高级计量经济学[M].北京：高等教育出版社.2011.07

附录

程序：

```
㈠ data=read.csv("new.csv",head=TRUE)#原始数据
library(urca)
lc.df <- ur.df(data$tem, lags=7, type='none')
summary(lc.df)
kappa(pm)
pm=data[,-c(1,7)]
head(pm)
attach(pm)
options(digits=2)
var=var(pm)
pm.ap=cor(pm)
pricomp=princomp(pm.ap)
pricomp$loadings
summary(pricomp)
pm1=read.csv("zhu.csv",head=TRUE)#主成分的数据
head(pm1)
attach(pm1)
d1=pm1$comp1
d2=pm1$comp2
d3=pm1$comp3
d4=pm1$comp4
par(mfrow=c(1,5))
acf(t,main="PM2.5浓度的自相关图")
acf(d1,main="空气质量因素的自相关图")
acf(d2,main="天气湿雾程度的自相关图")
acf(d3,main="其他天气综合因素的自相关图")
acf(d4,main="天气对流强度的自相关图")
t11=diff(t)
d11=diff(d1)
d12=diff(d2)
d13=diff(d3)
d14=diff(d4)
par(mfrow=c(1,5))
acf(t11,main="差分后的PM2.5浓度的自相关图",lag.max=25)
acf(d11,main="差分后的空气质量因素的自相关图",lag.max=25)
acf(d12,main="差分后的天气湿雾程度的自相关图",lag.max=25)
```

```
acf(d13,main="差分后的其他天气综合因素的自相关图",lag.max=25)
acf(d14,main="差分后的天气对流强度的自相关图",lag.max=25)
library(urca)#检验主成分序列的的稳定性
lc4.df <- ur.df(pm1$comp4, lags=7, type='none')
summary(lc4.df)
lc3.df <- ur.df(pm1$comp3, lags=7, type='none')
summary(lc3.df)
lc2.df <- ur.df(pm1$comp2, lags=7, type='none')
summary(lc2.df)
lc1.df <- ur.df(pm1$comp1, lags=7, type='none')
summary(lc1.df)
#检验差分后的主成分序列的的稳定性
lc.df1<- ur.df(diff(pm1$comp1), lags=5, type='none')
summary(lc.df1)
adf.test(diff(pm1$comp1))
lc.df2<- ur.df(diff(pm1$comp2), lags=5, type='none')
summary(lc.df2)
adf.test(diff(pm1$comp2))
lc.df3 <- ur.df(diff(pm1$comp3), lags=5, type='none')
summary(lc.df3)
adf.test(diff(pm1$comp3))
lc.df4<- ur.df(diff(pm1$comp4), lags=5, type='none')
summary(lc.df4)
t=data[,7]
lc.df<- ur.df(y1, lags=5, type='none')
summary(lc.df)
lc<- ur.df(y, lags=5, type='none')
summary(lc)
lt.df <- ur.df(t, lags=7, type='none')
summary(lt.df)
t=t[-1]
reg1=lm(t~d11)
library(lmtest)
dw1=dwtest(reg1)
```

（二）

```
PM2.5<-read.table(file.choose(),head=T,fill=TRUE)
PM2.5a<-PM2.5[,-c(1,7,8,9)]
Sigma.pm<-var(PM2.5a)
p.pm<-cor(PM2.5a)
PM2.5b=PM2.5[,c(8,13,14,15)]
```

```
library(dse)
attach(PM2.5b)
fld<-TSdata(input=PM2.5b[,2:4],output=PM2.5b[,1])
fld<-tframed(fld,list(start=c(2013,304),frequency=365))
seriesNamesInput(fld)<-c("COM1","COM2","COM3","COM4")
seriesNamesOutput(fld)<-"PM2.5"
####
fld.ls<-estVARXls(fld,max.lag=2)
print(fld.ls)
stability(fld.ls)
rr=checkResiduals(fld.ls)
par(mfrow=c(1,2))
acf(rr$re)
pacf(rr$re)
fld.ls2<-estVARXls(window(fld,end=c(2015,90)),max.lag=3)
f.p=forecast(fld.ls2,conditioning。 inputs=fld$input)
tfplot(f.p,start=c(2013,11))
par(new=T)
plot(PM2.5b[,1],type='l')
##
fld.ss<-estSSfromVARX(fld,max.lag=3)
print(fld.ss)
stability(fld.ss)
rs=checkResiduals(fld.ss)
par(mfrow=c(1,2))
acf(rr$re)
pacf(rr$re)
fld.ss<-estSSfromVARX(window(fld,end=c(2015,90)),fld,max.lag=3)
f.s.p=forecast(fld.ss,conditioning.inputs=fld$input)
tfplot(f.s.p)
par(new=T)
plot(PM2.5b[,1],type='l')
informationTests(fld.ls,fld.ss)
tfplot(fld.ls,fld.ss ,start=c(2013,304))
```