Computer System Networking and Telecommunications, 计算机系统网络和电信 (2)2020,2
ISSN:2661-3719(Print)

基于 ARIMA 模型和三次指数平滑法的 PM2.5 浓度分析预测

蒲　腾
西南科技大学信息与计算科学　四川　绵阳　621000

【摘　要】21 世纪以来，随着工业化水平的不断提高，环境污染也日益严重。近年，我国空气污染问题尤其突出，大气污染不仅危害人们的健康还对产业的发展起到不利的作用。所以控制空气污染，加强对空气的治理迫在眉睫。PM2.5 作为目前空气污染的污染物，PM2.5 的控制是大气污染和防止的关键性工作，由于 PM2.5 只能在当天检测出来，无法为当日提供实际性帮助。本文针对绵阳市近 6 年每月的 pm2.5 浓度的数据和特征，进行 ARIMA 和指数平滑建模，旨在准确预测出短期内 PM2.5 浓度，为行人的出行和政府决策提供较准确的出行建议和数据支撑。

【关键词】空气污染；PM2.5；ARIMA；指数平滑；提供出行建议和数据支撑

省环境保护厅的报告显示, PM2.5 的浓度受机动车尾气，建筑灰尘和火力发电厂的废气影响最大，而且风力的扩散也影响浓度监测。因此，预测 PM2.5 会遇到一些障碍，近年来由于我国的空气污染的重视程度的增加，有越来越多的学者开始投入到关于空气污染的研究中来，并且提出了其存在的问题和改进方法。如：王晓飞等人提出的 Prophet 和长短期记忆（LSTM）相结合的组合预测方法 [1]；中国科学院生态环境研究中心的薛同来等人采用机器学习的方式，建立的 BP 神经网络的非线性回归预测对 PM2.5 的综合评价 [2]。国外的研究者也对空气污染的的预测做了广泛研究，主要是多元分析，BP 神经网络等模型，灰色 GM 预测等模型。其中运用最多的神经网络模型也只是在对高浓度 PM2.5 预测上有较好的准确率。

1. 原理和步骤

1.1 模型的结构

ARIMA 模型由 Box-Jenkins 在 1970 年提出，又称为自回归移动平均模型，简记 ARIMA（p，d，q），通过对历史值，当前值和误差值的综合考虑，达到提高模型预测精度的目的。

$$\begin{cases}\phi(L)\nabla^d x_t=\Theta(L)\varepsilon_t+c\\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_t^2, E(\varepsilon_t,\varepsilon_s)=0, s\neq t\\ E(x_t,\varepsilon_t)=0,\forall s<t\end{cases}$$

其中：$\phi(L)=1-\varphi_1 L-\cdots-\varphi_p L^p$ 为自回归系数多项式，L 为滞后算子，d 为差分阶数。p 为自回归多项式阶数，d 为差分阶数，q 为移动平均多项式阶数，$\nabla^d=(1-L)^d$，xt 为时间序列，c 为常数项，$\varepsilon_t(t=1,2,\cdots)$ 为白噪声序列；s 和 t 代表时间序列的不同时刻。$E(\varepsilon_t)$ 为 t 时刻白噪声序列的均值；$Var(\varepsilon_t)$ 为 t 时刻白噪声序列的方差；$E(\varepsilon_t,\varepsilon_s)$ 为 t 与 s 时刻噪声序列的协方差；$E(y_t,\varepsilon_s)$ 为 $\{y_t\}$ 序列 t 时刻与白噪声序列 s 时刻的协方差；$\phi(L)=1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p$ 为 ARMA（p，q）模型自回归系数多项式，其中 $\phi(i)\ (i=1,2,\cdots,p)$ 为自回归多项式的待估系数 [3]；指数平滑法是布朗所提出，在计算指数平滑值时，将更多的权重放在最近的数据上，原理是任何周期的指数平滑值是当前周期的实际观测值和前一周期的指数平滑值的加权平均。

三次指数平滑法：

$$\begin{cases}\hat{y}_{t+T}=a_t+b_t T+c_t T^2\\ a_t=3S_t^{(1)}-3S_t^{(2)}+S_t^{(3)}\\ b_t=\dfrac{\alpha}{2(1-\alpha)^2}[(6-5\alpha)S_t^{(1)}-2(5-4\alpha)S_t^{(2)}+(4-3\alpha)S_t^{(3)}]\\ c_t=\dfrac{\alpha^2}{2(1-\alpha)^2}[S_t^{(1)}-2S_t^{(2)}+S_t^{(3)}]\end{cases}$$

1.2 建模步骤

1.2.1 数据预处理

平稳性检验一般有三种方法：①时序图检验；②自相关图检验；③统计量检验。若该序列为非平稳序列，还需进行差分操作转换为平稳序列，选取非白噪声序列进行研究。本文使用时序图和单位根（ADF）检验，并对序列的随机性进行判别。

1.2.2 对平稳序列建模

（1）观测序列的自相关系数和偏自相关系数样本值。

（2）根据模型判别准则确定 ARIMA（p，q）模型的适当阶数。

（3）估计模型中未知参数的值。

（4）检验模型的显著性，舍弃拟合模型未通过检验的拟合模型，本文对模型进行了较为精确的假设检验和显著性检验。可以使得模型的可行性和适用性更符合应用标准。

（5）模型优化。建立多个拟合模型，通过赤池信息量和贝叶斯信息规则，在所有通过检验的拟合模型中挑选出最优模型。

（6）模型评价。通过对 ARIMA 模型和指数平滑法模

型预测结果的分析，对模型的效果进行评价

（7）应用分析。使用模型进行预测可视化，分析模型的应用效果[4]。

2 实证分析

作为中共中央和国务院批准建设的中国唯一一座科技城，绵阳市坐落于四川盆地的西北部，培江中上游。是我国重要的国防设施和电子工业生产制造基地，成渝经济圈区域中心和四川第二大城市。对绵阳市 PM2.5 浓度分析预测具有很好的实际意义。本文通过对绵阳市天气网提供的准确数据，对 2014~2019 年每月 PM2.5 浓度进行建模，对绵阳市短期的 PM2.5 浓度和空气质量进行预测，可以很好的反应绵阳市的空气质量水平，并为绵阳市的空气治理提供帮助。

2.1 时序图检验

2.1.1 绘制 2014~2019 年每年平均温度序列时序图

图 1 显示 2014~2019 年绵阳市每月 PM2.5 浓度始终围绕在 50 附近随机波动，没有明显的趋势或周期，初步视为平稳序列。为了更加稳妥，再利用 Box-Piercetest 对延迟 6 阶和延迟 12 阶进行平稳性检验。

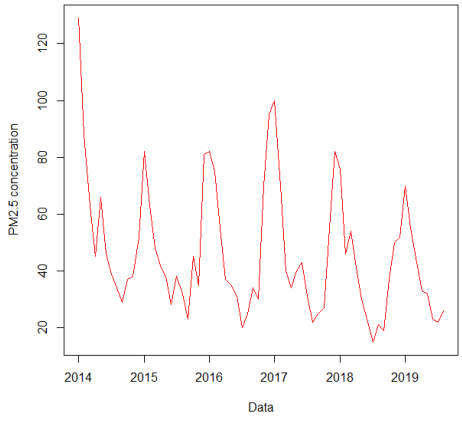

图 1 绵阳市 2014~2019 年每月 PM2.5 浓度时序图

2.1.2 差分运算

在平稳性检验之后，当 p 的值小于 α 时，可以将序列视为平稳的。并且在差分或不差分的情况下，尽量选择不差分的方法。尽管差分可以消除长期趋势的影响并达到规律性，但是也会有信息的丢失。因此，本文使用原始数据进行后续试验。

2.2 白噪声检验

使用 Box.test 函数对延迟 6 期，延迟 12 期的 Qlb 统计量进行计算可知延迟六阶 p 值为 $3.675e^{-14}$，延迟十二阶 p 值为 $2.2e^{-16}$。由于 p 值都显著大于显著性水平 α，所以该序列可以拒绝纯随机的原假设，进一步可以断定 2014~2019 年绵阳市每月 PM2.5 浓度序列为非白噪声序列[5]。

表 1 模型判断表

$\hat{\rho}_k$	$\hat{\phi}_{kk}$	模型选择
拖尾	p 阶截尾	AR（p）
q 阶截尾	拖尾	MA（q）
拖尾	拖尾	ARMA（p，q）

2.3 模型的识别

通过观察图 2，可以看到明显的正弦波路径，这表明自相关系数减小到零是一个连续的渐进过程，而不是突然的过程，这是 acf 拖尾性的典型特征。

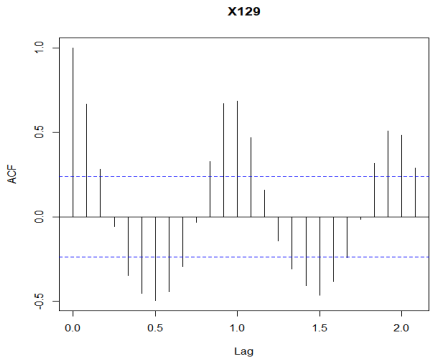

图 2 自相关图

图 3，偏自相关图也大致显示出不截尾的性质或 1 阶截尾性所以可以人工选择 ARIMA（1，0，0），ARIMA（1，0，1），ARIMA（2，0，0）ARIMA（2，0，1），ARIMA（3，0，0），ARIMA（3，0，1）6 个模型模型和 auto.arima 函数自动定阶的 ARIMA（1，0，0），ARIMA（1，1，0）进行拟合。

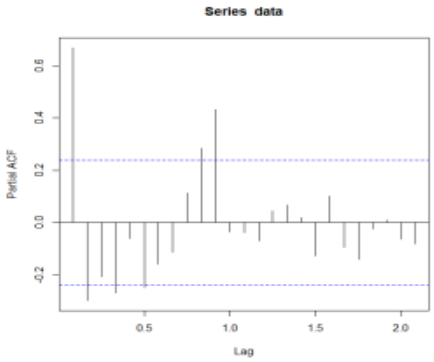

图 3 偏自相关图

运用 R 语言中的 arima（）估计模型参数[6]，以上 7 种模型的表达式如下：

$$\begin{cases} y_t = 42.0767 + 0.6785\, y_{t-1} + \varepsilon_t, Var(\varepsilon_t) = 201.9 \\ y_t = 45.3263 + 0.5672\, y_{t-1} + \varepsilon_t + 0.3379\, \varepsilon_{t-1}, Var(\varepsilon_t) = 196 \\ y_t = 45.2388 + 0.9652\, y_{t-1} + \varepsilon_t - 0.3607\, \varepsilon_{t-1}, Var(\varepsilon_t) = 186.7 \\ y_t = 44.8846 + 1.4491\, y_{t-1} - 0.7127\, y_{t-2} + \varepsilon_t - 0.6736\, \varepsilon_{t-1}, Var(\varepsilon_t) = 161.7 \\ y_t = 45.2420 + 0.8895\, y_{t-1} - 0.1594\, y_{t-2} - 0.2146\, y_{t-3} + \varepsilon_t, Var(\varepsilon_t) = 178.2 \\ y_t = 45.0982 + 1.4687\, y_{t-1} - 0.6922\, y_{t-2} - 0.0552\, y_{t-3} + \varepsilon_t - 0.7099\, \varepsilon_{t-1}, Var(\varepsilon_t) = 162.8 \\ y_t = 0.1652\, y_{t-1} + \varepsilon_t, Var(\varepsilon_t) = 239 \end{cases}$$

2.4 模型检验

2.4.1 模型的显著性检验

良好的拟合模型应该可以提取出观察序列中的几乎所有样本信息，这种模型称为显著有效模型。相反，如果残差序列是非白噪声序列，则意味着残差序列中任存在尚未提取的相关信息，就会照成信息的损失，所以对模型的残差进行白噪声检验[7]。

原假设和备择假设分别为：

H0：ρ1=ρ2=……ρm=0，∀m>=1

H1：至少存在某个 ρi ≠ 0，∀m>=1，i<=m.

使用 Ljung-Box 检验统计量：

$$Q(m)=T(T+2)\sum_{i=1}^{m}\frac{\hat{\rho}_i}{T-i}$$

通过观察图 4、表 2，6 阶或 12 阶延迟下的 P 值 <0.05，并且 QQ 图上前面点与线的拟合效果不好。拟合模型的残差不视为白噪声序列，所以 ARIMA（1，0，0），ARIMA（1，0，1），ARIMA（2，0，0），ARIMA（1，1，0）模型不显著。由于 6 阶或 12 阶延迟下的 P 值都显著大于 0.05。并且结合 QQ 图，可知残差符号正态性假设且不相关。认为 ARIMA（2，0，1），ARIMA（3，0，0），ARIMA（3，0，1）拟合模型显著。

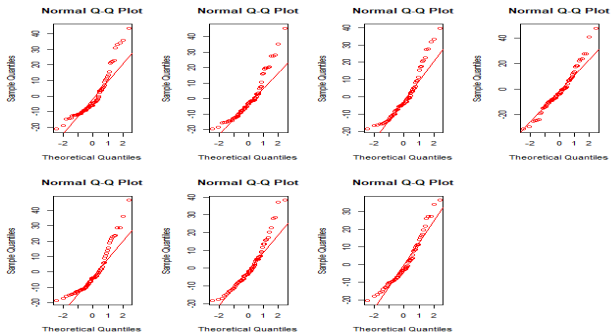

图 4 模型的 QQ 示例图

表 2 模型的延迟 6 · 12 阶 P 值统计表

	ARIMA（1,0,0）	ARIMA（1,0,1）	ARIMA（2,0,0）	ARIMA（2,0,1）	ARIMA（3,0,0）	ARIMA（3,0,1）	ARIMA（1,1,0）
延迟 6 阶 p 值	0.013	0.2307	0.3429	0.9569	0.7145	0.7005	0.5465
延迟 12 阶 p 值	$1.052e^{-7}$	0.001602	0.0252	0.2452	0.06988	0.09964	0.004587

2.4.2 参数的显著性检验

进行参数检验可以精简所使用模型。并且，如果某个参数不显著，即意味着与参数相对应的自变量对因变量没有明显的影响，可以从拟合模型中消除该自变量。由于 R 语言无法提供为参数的显著性结果，所以自己计算参数的 t 统计量的值及统计量的 p 值，调用 t 分布 p 值函数，pt 即可获得该统计量的 p 值。采用 t 统计量进行比较，显著性水平取 0.05，那就相当于系数和 1.96 比较，通过表 3 可以看出，ARIMA（2，0，1），参数都十分显著，ARIMA（3，0，0），显著性不好，ARIMA（3，0，1）除了 AR3 系数，其余参数都十分显著。所以选择 ARIMA（2，0，1），ARIMA（3，0，1）模型显著性较好，并进化后续模型优化[8]。

表 3 模型系数表

	AR1	AR2	AR3	MA1
ARIMA（2，0，1）	10.073576	-6.491124	5.324864	-2.630915
ARIMA（3，0，0）	7.711469	-1.095534	-1.733257	-2.135684
ARIMA（3，0，1）	7.4328064	-2.8346495	0.1411426	-3.8903843

2.5 模型优化

运用 AIC 和 SBC（BIC）信息准则对模型进行优化。最小信息量显示，ARIMA（3，0，0，ARIMA（3，0，1）AIC，BIC 值分别为 556.5591，567.6567；552.8623，566.1793.AIC 和 BIC 准则判断，ARIMA（3，0，1）模型都要优于 ARIMA（2，0，1），ARIMA（3，0，0）所以模型 ARIMA（0，1，2）是最优拟合模型。

2.6 建立霍尔特指数平滑预测模型[9]

以 2014 年到 2018 年一月数据作为原始数据，预测后 5 个月与真实数据比较如图 5 所示。

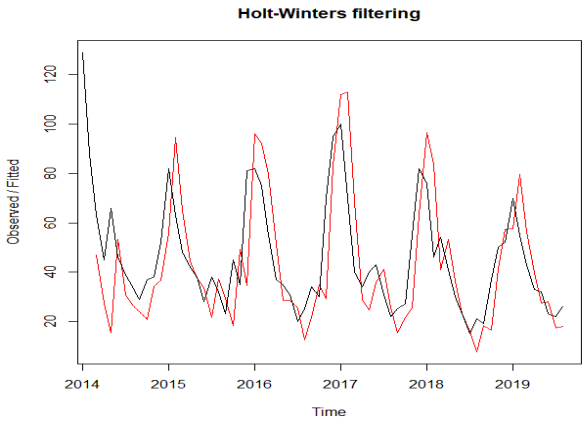

图 5 绵阳市 2014~2019 年每月 PM2.5 浓度指数平滑预测模型图

黑色曲线为原始序列，红色曲线为预测值，由图看预测效果还行，接下来由实验相关预测值显示，alpha 值为 1；beta 预测值为 0.34，这些值相对较高，表明在水平上和趋势的斜率上，当前期间的实际值在很大程度上取决于最近的观察结果，这个的结果也与预期吻合。

2.7 预测效果分析

Ljung-Box 检验时，p=0.01822，意味着置信度有 99%，如此低的值不足以拒绝“一到二十阶上误差为非零且自相关的”，所以认为预测误差在 1~20 阶是非零且自相关的。同样我们进一步验证测试实验误差为偶然误差，不收到某种控

制以及是否符合零均值正态分布，画出时间预测误差图如图 6 所示。

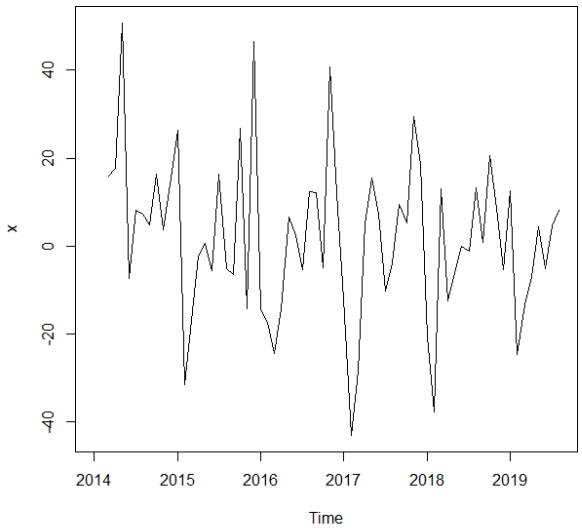

图 6 绵阳市 2014~2019 年每月 PM2.5 浓度指数平滑误差模型图

可见预测误差的方差是大致不变的，验证了测试误差符合零均值正态分布，在预测 PM2.5 浓度上指数平滑模型拟合较好。

通过对图 7 中观察值和拟合值序列的观察，可以看出 ARIMA 模拟拟合程度比较好。

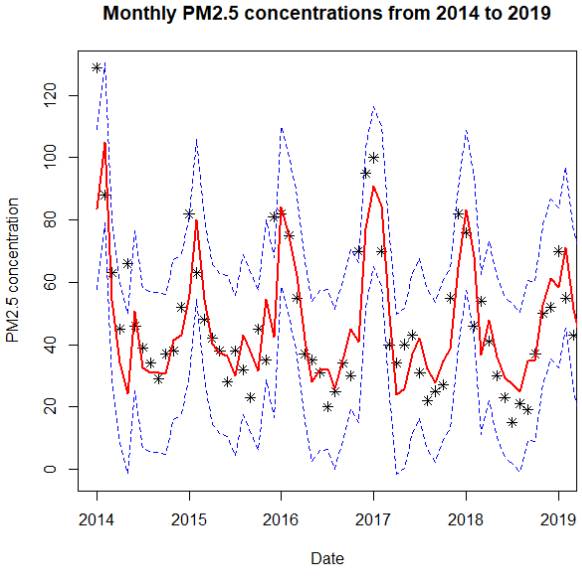

图 7 ARIMA 模型绵阳市 2019 年 9 月 -2020 年 1 月预测序列拟合范围图

注：黑色 * 号为原始观察值序列，红色实线为拟合值，蓝色虚线为 95% 置信水平为的置信线。

利用 ARIMA（3，0，1）模型以及三次指数平滑法预测 2019 年 9 月 -2020 年 1 月与实际数据比较，从表 4 中可以看出没有其他因素的影响绵阳市 PM2.5 浓度初和年中会上升，8、9 月达到最低值，年后会逐渐上升。这与实际调查所得情况符合且满足历史数据的规律。因为实际 PM2.5 浓度受到多方面因素的影响，因此也只能做短期的预测。

表 4 ARIMA 模型预测与实际数据比较

时间	ARIMA 预测值	指数平滑预测值	真实值
2019-09	39	24	24
2019-10	52	23	23
2019-11	61	22	39
2019-12	63	21	63
2019-01	61	19	70
2019-02	57	18	55
2019-03	51	16	43
2019-04	44	15	33

3 预测结果分析和建议方案

以上预测过程将实验结果与 2019 年 9-12 月和 2018 年 1~4 月数据比较显示，指数平滑法对年中的 PM2.5 浓度预测较好，ARIMA 模型对年初和年末的预测较好，2019 年 12 月预测的 PM2.5 浓度都与实际值十分接近，模型拟合效果较为理想，短期预测精度相对较高。由以上结果可知：PM2.5 浓度总体在开年的上半年浓度逐月降低，后半年会逐渐增加。PM2.5 月均浓度高值主要集中在 1~4 月、10~12 月，而在 5~9 月份其浓度值较低，绵阳市 2019 年污染防治攻坚战 6 月战报显示，6、7 月绵阳市对环境和空气的治理力度的加大，明显使得 91011 月的空气质量得到明显改善，所以实际 PM2.5 浓度较预测小，但对照历年数据可知与预测结果吻合效果符合正常预期。

该文针对绵阳市历年的 PM2.5 数据对 PM2.5 浓度进行的短期预测，效果十分理想，模型与实际结合情况较好。建立 ARIMA 和指数平滑的对比模型是一种较好的拟合 PM2.5 浓度的方法，可以很好的推广到绵阳市的其他空气质量数据的观测中去。但由于 PM2.5 浓度受气象条件和节假日等多方面影响，以及模型自身对数据的要求，本质上只能处理线性关系。所以只适用于进行短期预测，在未来可以结合机器学习中支持向量机和灰色模型为环境治理和行人出现提供出行预警和治理方法 [10]。

【参考文献】

[1] 王晓飞，王波，陆玉玉 . 基 Prophet-LSTM 模型的 PM2.5 浓度预测研究 [J/OL]. 软件导刊：1-4[2020-03-31].

[2] 薛同来，赵冬晖，韩菲 . 基于 BP 神经网络的北京市 PM(2.5) 浓度预测 [J]. 新型工业化，2019，9(08)：87-91.

[3] 汪伟舵，吴涛涛，张子振 . 基于 ARIMA 模型的杭州市 PM2.5 预测 [J]. 哈尔滨师范大学自然科学学报，2018，34(03)：49-55.

[5] 王燕 . 时间序列预测分析 - 基于 R[M]. 北京：中国人民出版社，2015：8-40.

[6] 王燕 . 应用时间序列分析 (第四版)[M]. 北京：中国人民出版社，2015：12-32.

[7] 黄芸，姜国，徐治欠 .ARIMA 模型在黄石市 PM2.5 浓度预测中的应用 [J]. 湖北师范大学学报，2017，37(02)：38-42.

[8] 李杰，彭晓明 . 基于 ARIMA 模型的军事训练数据分析和预测 [J]. 舰船电子对抗，2019，42(05)：98-101，120.

[9] 严宙宁，牟敬锋，赵星，严燕，罗文亮，胡满达 . 基于 ARIMA 模型的深圳市大气 PM2.5 浓度时间序列预测分析 [J]. 现代预防医学，2018，45(02)：220-223.

[10] 孔德川，潘浩，郑雅旭，姜晨彦，韩若冰，吴寰宇，陈健 . 指数平滑模型在上海市猩红热发病率预测中的应用 [J]. 疾病监测，2019，34(10)：932-936.