上海市空气质量状况分析与预测

周天人 聂蕴哲 戴彧

指导老师：张青

江苏省苏州中学

二〇一五年八月

摘要

空气质量指数（Air Quality Index，简称 AQI）是定量描述空气质量状况的指标，其大小与人们日常生活息息相关。本文基于时间序列的方法，研究数据是上海市 2014 年 8 月至 2015 年 3 月的空气质量指数的时刻数据，在此基础上对数据进行数据平稳化，拟合模型，估计参数，对模型进行定阶，从而建立时间序列模型并对模型进行检验，最终确定了上海市空气质量指数的模型为 ARIMA(3,1,7)模型。在此模型的基础上，我们对数据进行了样本内和样本外预测，预测结果的绝对误差均控制在 6 以内，说明此模型具有有效性。之后，我们通过相关性分析研究风力和大气温度等因素对上海市空气质量指数的影响和相关性。此外，通过 Granger 因果检验进一步得出结论，大气温度与空气质量指数是相互影响的，风力会影响空气质量指数的大小，而空气质量并不会直接影响风力的大小。

关键词：时间序列，空气质量指数，ARIMA，协整，Granger

ABSTRACT

Air Quality Index is a measure of the quality of air, which is connected with the quality of people's living standards. Based on time series theory and Shanghai from Aug.2014 to March.2015 the AQI, we establish a time series model, and model testing, to determine a suitable model for autoregressive moving average model ARIMA（3,1,7）,which based on the theory include the smooth of the data processing, the identification of the model and the estimation of parameter. Model of Shanghai's AQI forecast in the sample and five step forward forecasts and compare with the actual values. Results showed that the absolute error is below 6, which shows that the model is suitable. Then, we analyze the effect of wind strength and atmospheric temperature to the quality of air by cointegration relation and correlation analyze. By Granger test, we find that atmospheric temperature and air quality is affected with each other. However, wind strength influences the value of AQI, while air quality is not affected with the force of wind.

Key Word: Time series, AQI, ARIMA,Cointegration,Granger

目录

一、研究背景介绍.....1
二、时间序列的定义和基本方法.....1
三、数据来源.....4
四、建立模型.....5
五、样本内预测与样本外预测.....11
六、相关性分析.....14
七、结论.....20

一、研究背景介绍

近年来，随着我国经济科技发展速度的不断提高，许多环境问题不断涌现。城市大气污染问题伴随城市化进程的不断深入也越来越被关注。城市的空气质量水平不仅关系到城市环境质量水平，更重要的是良好的空气质量水平是城市居民健康生活工作学习的保证。因此探求城市环境质量变化趋势以及空气质量的预测是很有必要的，它可以为环保等部门制定相关政策和措施时提供一定前瞻性的依据。

上海作为中国发展最前沿的城市，在经济实力不断攀升的同时，其空气质量也面临着日益严峻的挑战。上海作为国家重点环保城市，其空气污染特征在全国城市范围内具有一定代表性，对其的研究结果也能在全国其他城市起到借鉴和示范作用。

二、时间序列的定义和基本方法

（一） 时间序列的定义

在统计研究中，常用按时间顺序排列的一组随机变量

$$X_1，\ X_2，\ \cdots X_t，\ \cdots$$

来表示一个随机事件的时间序列，简记为$\{X_t, t \in T\}$或$\{X_t\}$。

用

$$x_1，\ x_2，\ \cdots，\ x_n$$

来表示该随机序列的 n 个有序观测值，称为序列为 n 的观察值序列。

（二） 时间序列的预处理

1. 差分运算

一阶差分 $\nabla y_t = y_t - y_{t-1}$

p 阶差分　　$\nabla^p x_t = \nabla^{p-1} x_t - \nabla^{p-1} x_{t-1}$

k 步差分　　$\nabla_k y_t = y_t - y_{t-k}$

一般来说，对于含有明显的线性趋势的数据，一阶差分就可以使数据平稳，而对于含有明显的曲线趋势的数据，通常利用二阶差分或三阶差分就可以使数平稳。对于含有周期性或者季节性的数据，可以按照周期的大小或季节进行 k 步差分，使数据平稳便于处理。

2. 平稳性检验

宽平稳（weak stationary）是使用序列的特征统计量来定义的一种平稳性。如果$\{X_t\}$满足如下三个条件：

（1）任取$t \in T$，有$EX_1^2 < \infty$；

（2）任取$t \in T$，有$EX_t = \mu$，μ为常数；

（3）任取t，s，$k \in T$，且$k+s-t \in T$，有$\gamma(t,s) = \gamma(k,k+s-t)$

则称为$\{X_t\}$为宽平稳时间序列。

平稳性是某些数据所具有的特征，具有平稳性的数据可以极大地减少随机变量的个数，并增加了待估变量的样本容量。

平稳性检验主要有三种，时序图检验、自相关图检验和假设检验。

3. 纯随机性检验

如果时间序列$\{X_t\}$满足如下性质：

（1）任取$t \in T$，有$EX_t = \mu$；

（2）任取t，$s \in T$，有

$$\gamma(t.s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s \end{cases}$$

称序列$\{X_t\}$为纯随机序列，也成为白噪声序列。

纯随机序列各项之间没有任何关联，序列在进行完全无序的随机波动。一旦

某个随机事件呈现出纯随机波动的特征，就认为随机事件没有包含任何值得提取的有用信息，我们就应该终止分析了。

纯随机性检验也称为白噪声检验，是专门用来检查序列是否为纯随机序列的一种方法。即检验

$$\gamma(k)=0,\quad \forall k\neq 0$$

是否成立。实际上，由于观察值序列的有限性，导致纯随机序列的样本自相关函数不会绝对为零。

（三） 时间序列的模型

1. AR 模型

AR 模型称为(Auto regressive model），可以表达为

$$x_t=\phi_0+\phi_1 x_{t-1}+\phi_2 x_{t-2}+\cdots+\phi_p x_{t-p}+\varepsilon_t,$$

其中ϕ_i，i=1,2,...p 是自回归系数，p 是自回归模型的阶数，ε_t是白噪声序列。x_t是 p 阶自回归序列，用 AR(p)表示。

2. MA 模型

MA 模型称为移动平均模型(Moving average model），可以表达为

$$x_t=\mu+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q},$$

其中θ_i，i=1,2,...q 是移动平均系数，q 是移动平均模型的阶数，ε_t是白噪声序列。x_t是 q 阶移动平均序列，用 MA(q)表示。

3. ARMA 模型

ARMA 模型称为自回归移动平均模型（Auto regressive moving average model），由自回归模型和移动平均模型两部分共同构成，模型可以表达为

$$x_t=\phi_0+\phi_1 x_{t-1}+\phi_2 x_{t-2}+\cdots+\phi_p x_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q},$$

记为 ARMA(p，q)，其中ϕ_i，i=1,2,...p 是自回归系数，θ_i，i=1,2,...q 是移动平均系数。

4. ARIMA 模型

ARIMA 模型称为求和自回归移动平均模型(Auto regressive integrated moving average model)，模型可以表达为

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$$

其中 B 是延迟算子，满足 $x_{t-1} = Bx_t$，$\nabla^d = (1-B)^d$，$\Phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p$ 是自回归系数多项式，$\Theta(B) = 1-\theta_1 B-\cdots-\theta_q B^q$ 是移动平均多项式。对于模型 ARIMA（p，d，q），其中 AR 是自回归函数，p 为自回归项数；MA 为移动平均函数，q 为移动平均项数，d 为时间序列成为平稳时所做的差分次数。

模型 ARIMA 模型是将非平稳时间序列转化为平稳时间序列，然后在平稳的基础上建立自回归移动平均模型，因此求和自回归移动平均模型的实质就是差分运算与 ARMA 模型的组合。

三、数据来源

为了研究该题目，我们在上海市环境监测中心（www.semc.gov.cn）网站上获取了从 2014 年 8 月 1 日 0：00 至 2015 年 3 月 19 日 23：00 的以小时划分的上海市空气质量指数（AQI）数据，共计 5518 个数据。由于网站数据的关系，我们收集到的数据中存在缺失值。为了处理这些缺失值，我们采用移动平均的方式对缺失值进行估计和补全。由于数据量很大，且观测值精确到小时，具有很大的研究意义。

四、建立模型

（一）ARIMA 模型

1.建模流程图

尝试使用 ARIMA 模型对观察序列建模是一件比较简单的事情。它遵循如下操作流程：

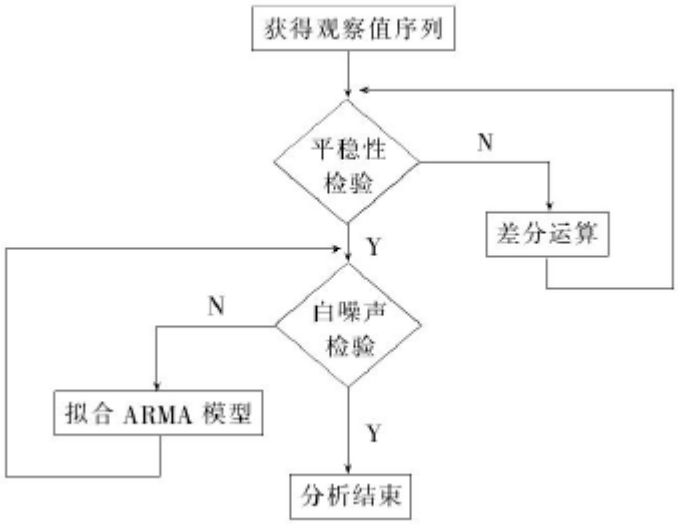

图 1.　ARIMA 建模流程图

这就是一个完整的 ARIMA 建模思路。

2.检验平稳性和随机性

我们首先利用 R 软件绘出空气质量指数的时序图，如图 2.

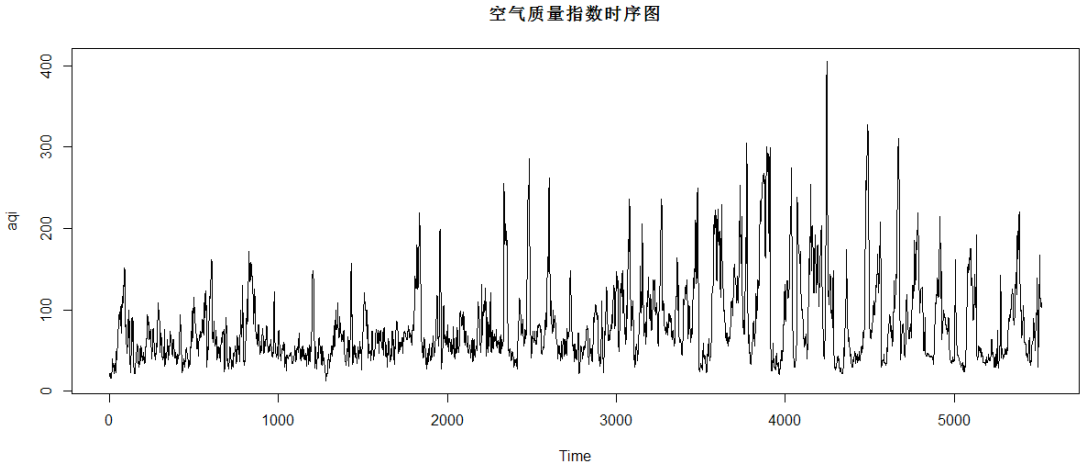

图 2. 空气质量指数时序图

从时序图中可以看出，原数据没有明显的上升或下降的趋势，因此可以排除趋势因素。又时序图中也并没有明显的周期性，我们也可以排除周期性的影响。另外，由于原数据选取的时间跨度不足一年，可以排除季节性的影响。由于数据的波动较大，为了稳妥起见，我们对数据进一步通过统计量进行平稳性检验，结

果如表 1.

表 1.　AQI 时间序列平稳性检验

AQI	方法	P-value	结论
平稳性	KPSS	<0.01	不平稳

其中，KPSS 检验的原假设是 H_0:原序列为平稳序列，当 P 值小于 0.01 时，我们则拒绝原假设，当 P 值大于 0.01 时，我们则接受原假设。从检验结果看出，检验的 P 值是小于 0.01 的，因此我们有理由拒绝原假设，认为序列是不平稳。

接下来我们对数据进行了随机性检验，采用的检验是 Box.test，检验结果如表 2.

表 2.　AQI 时间序列随机性检验

AQI	方法	P-value	结论
随机性	Box.test	< 2.2e-16	具有相关性

Box.test 是对序列的随机性进行检验，原假设：残差序列为白噪声序列，即

$$H_0:\rho_1=\rho_2=\cdots=\rho_m=0,\forall m\geq 1$$

备择假设：残差序列为非白噪声序列，即

$$H_1:\text{至少存在某个}\rho_k\neq 0,\forall m\geq 1,\ k\leq m\text{。}$$

Box 检验的统计量为

$$LB=T(T+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{T-k}$$

式中 T 为序列观察期数，m 为指定延迟期数。经 R 软件计算得到检验的 P 值为 2.2e-16，远小于临界值 0.01，因此我们有充分的理由拒绝原假设，认为原序列不具有随机性。

3.非平稳序列进行差分

为了使数据能够变成平稳序列，我们对数据做一阶差分处理，并用 R 软件画出了一阶差分后的数据时序图，如图 3 所示。

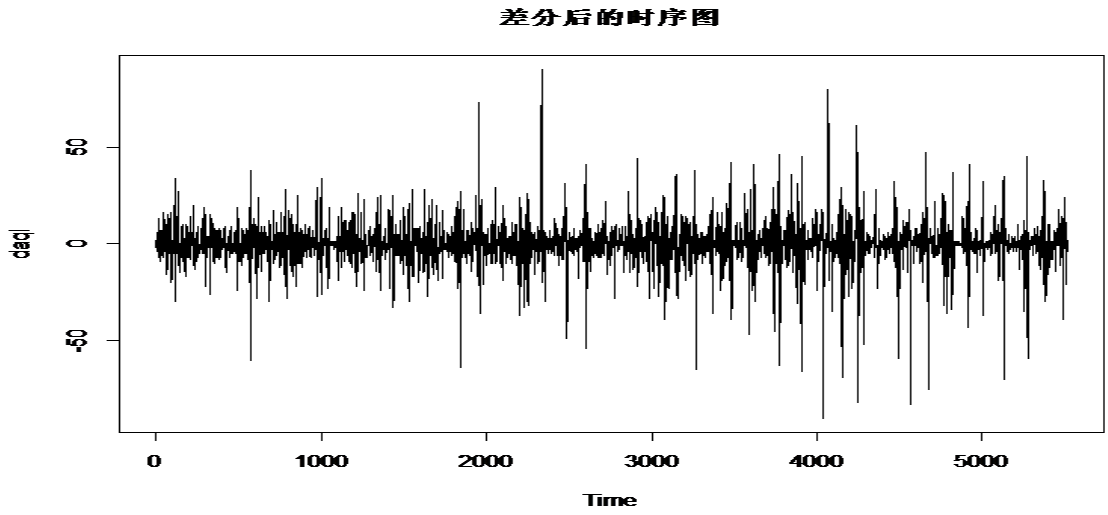

图 3.　AQI 差分后的时序图

从时序图来看，差分后的数据波动变小，且围绕一个中心上下波动，因此可以认为差分后的序列是平稳的。为了进一步确定，我们对差分后的数据又进行平稳性检验，并且做了随机性检验，检验结果如表 3.

表 3.　AQI 差分后的平稳性和随机性检验

AQI 差分后	方法	P-value	结论
平稳性	ADF	<0.01	无单位根
	PP	<0.01	无单位根
	KPSS	>0.1	平稳
随机性	Box.test	< 2.2e-16	具有相关性

根据表 3 的检验结果可知，差分后的数据是平稳的且具有相关性。其中，ADF 与 PP 单位根检验的原理是：通过检验序列是否存在单位根来判断序列是否平稳，两者的原假设均是存在单位根，从表中可以看出检验的 P 值小于 0.01，拒绝原假设，KPSS 检验的 P 值大于 0.1，进一步验证了差分后的序列是平稳的。

4.差分后对模型进行定阶

由以上验证结果，我们尝试用 ARIMA 模型对数据进行拟合。根据时间序列相关原理，我们利用数据的自相关图（ACF）和偏自相关图（PACF）对 ARIMA 进行定阶。我们用 R 画出了相应的 ACF 图和 PACF 图，如图 4 和图 5 所示。

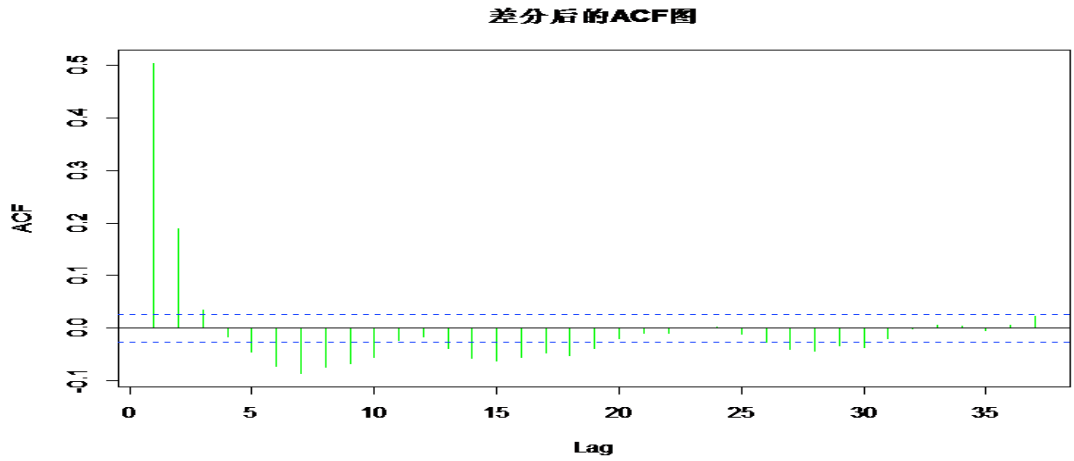

图 4.　AQI 差分后的 ACF 图

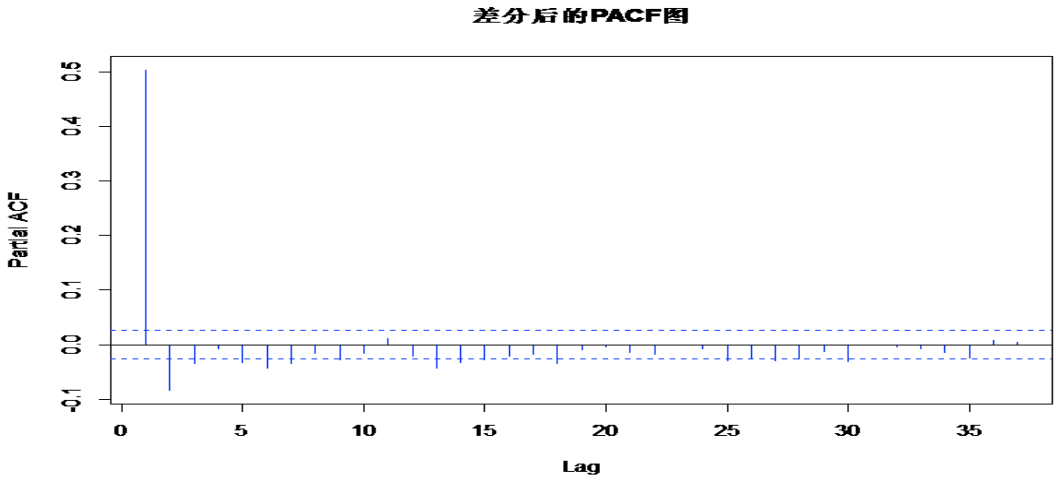

图 5.　AQI 差分后的 PACF 图

从图 4 和图 5 可以看到 ACF 和 PACF 都出现拖尾现象，这样就验证了我们用 ARIMA 的模型去构建模型的思路。但是，仅凭借上面的 ACF 图与 PACF 图不能有效确定出 ARIMA 中的阶数，所以我们引用了拓展的自相关系数图 EACF 来确定阶数。在 R 语言 TSA 包中的 eacf（）函数能够绘制拓展自相关函数 EACF 图，如图 6 所示。利用 EACF 图定阶的原理是，找到一个下三角，使得该三角形内部以及边界都是圆圈，通过三角形的顶点位置来确定模型阶数。根据图 6，例如，顶点位置在第三行和第七列的三角形，对应可得出模型 AR 阶数是 3，MA

阶数是 7，又如，顶点位置在第一行和第十列的三角形，对应可得出模型 AR 阶数是 1，MA 阶数是 10，如此下去可得到很多符合要求的阶数。因此需对模型优化，模型优化的方法是通过模型的 AIC 信息量达到最小的模型就是相对最优的模型，根据比较结果（见表 4），相对最优的模型为 ARIMA(3,1,7)。

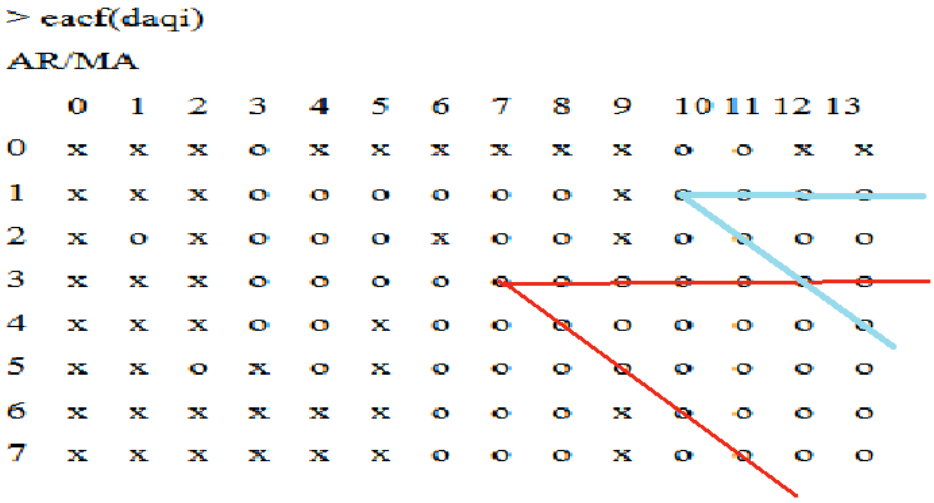

图 6. AQI 差分后的拓展自相关系数图

表 4. 模型优化

模型	AIC 值
ARIMA(3,1,7)	39450.67
ARIMA(1,1,10)	39452.73
ARIMA(2,1,10)	39454.96
ARIMA(1,1,11)	39451.09
ARIMA(2,1,11)	39453.03

由 EACF 的定阶，可以看出模型的阶数较大，对此的解释是：由于原数据是以小时为计数间隔的，某一时间点的空气污染程度将会受到前期时段的空气污染程度很大的影响，同理该时间段的空气污染程度也会对未来时刻的空气质量指数产生很大的影响。结合常识来考虑，模型的阶数较大也具有合理性。因此，我们拟用 ARIMA(3,1,7)模型对数据进行拟合。利用 R 程序的拟合，得到拟合结果如

图 7。

```
> arima
Call:
arima(x = aqi, order = c(3, 1, 7))
Coefficients:
          ar1      ar2      ar3      ma1      ma2      ma3      ma4      ma5
      0.3824  -0.0307   0.5071   0.1326   0.0197  -0.5601  -0.3092  -0.1509
s.e.  0.4308   0.2337   0.2962   0.4305   0.1958   0.2529   0.1666   0.0860
          ma6      ma7
      -0.0703  -0.0533
s.e.   0.0271   0.0165
```

图 7.　ARIMA(3,1,7)模型拟合结果

模型拟合完后，我们对拟合后的残差再进行平稳性和白噪声检验。残差的时序图如图 8，不难发现，残差是平稳的。再通过单位根检验，进一步验证，不难发现 ADF、PP 的 P 值都小于临界值 0.01，KPSS 的 P 值大于 0.1，因此我们可以拒绝原假设，认为残差序列是平稳的。

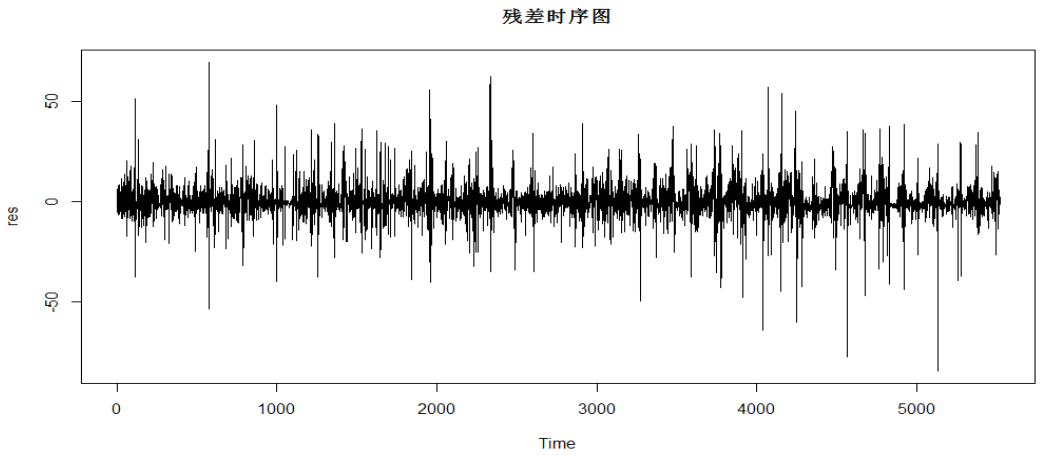

图 8.　ARIMA(3,1,7)模型拟合后残差时序图

表 5.　ARIMA（3,1,7）模型拟合后残差的平稳性检验

方法	P-value	结论
ADF	<0.01	无单位根
PP	<0.01	无单位根
KPSS	>0.1	平稳

接着我们对拟合后的残差模型进行白噪声检验，分别应用了 ACF 图和 BOX 检验。ACF 图如图 9 所示，残差自相关函数一致落在二倍误差内的，

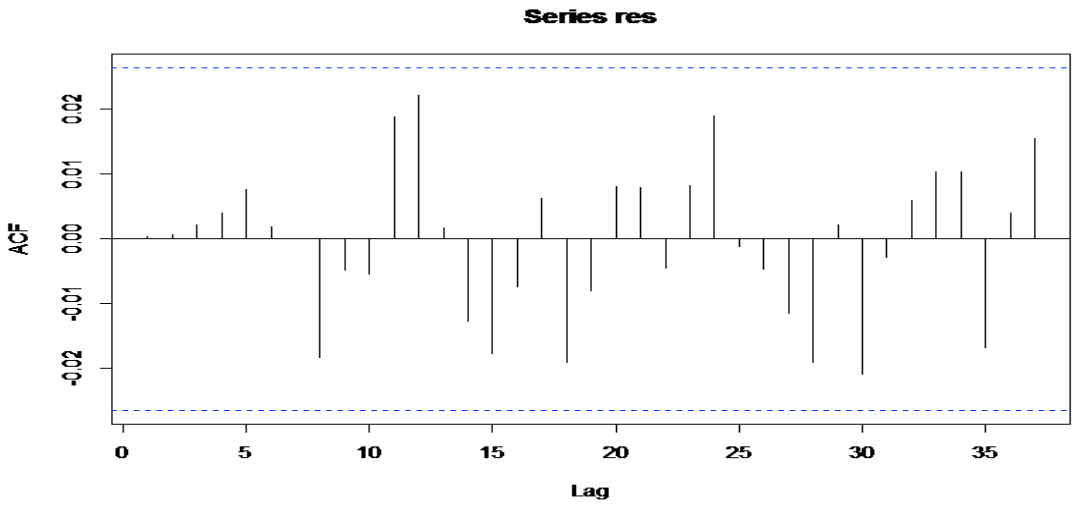

图 9.　ARIMA（3,1,7）模型拟合后残差的 ACF 图

因此我们认为拟合后的残差具有随机性的。为了保险起见，我们又应用了 BOX 统计量进行检验，BOX 检验的 P 值为 0.9658，远大于临界值 0.01，因此我们认为不能拒绝原假设，即接受拟合后的残差序列具有随机性这一假设。综合 R 软件拟合和检验的结果来看，该模型是显著的，确定模型为 ARIMA(3,1,7)模型。

五、样本内预测与样本外预测

为了验证模型的合理性，我们利用建立的模型分别进行了样本内预测和样本外预测，来展示模型拟合的效果。

考虑到空气质量指数这一指标的取值范围跨度较大，最低可低至 20 以下，最高可达到 400 以上。且考虑到基数会强烈影响到相对误差，例如在基数为 20 的情况下，即使绝对误差只有 1，相对误差也会达到 5%，而基数为 400 时，绝

对误差到达 20 时，相对误差也只有 5%。因此用相对误差来衡量预测结果的好坏程度在本文中并不科学。所以，我们用绝对误差来衡量预测的效果，只要绝对误差控制 6 或 6 以下，我们就可以认为模型的预测结果良好。

1．样本内预测

为了验证模型的合理性，首先我们进行样本内预测。我们在 ARIMA(3,1,7)模型的基础上，取出样本中第 5514 个至 5518 个数据进行样本内预测。考虑到 ARIMA 对具有短期预测有效性，因此我们预测时采用迭代数据预测，即把第 n 次预测的数据加到原始数据里，用来预测第 n+1 个数据。预测结果与原数据的对比如表 6：

表 6. 样本内预测值与实际值的比较

时刻	实际值	预测值	绝对误差
5514	103	102	1
5515	105	102	3
5516	101	100	1
5517	102	99	3
5518	101	98	3

在此基础上，我们可以从表 6 中看出，绝对误差最大为 3，可以认为预测结果合理。此外我们把样本内预测通过绘图放大予以展现出来，如图 10 所示（用迭代法），并附上 95%的置信带。

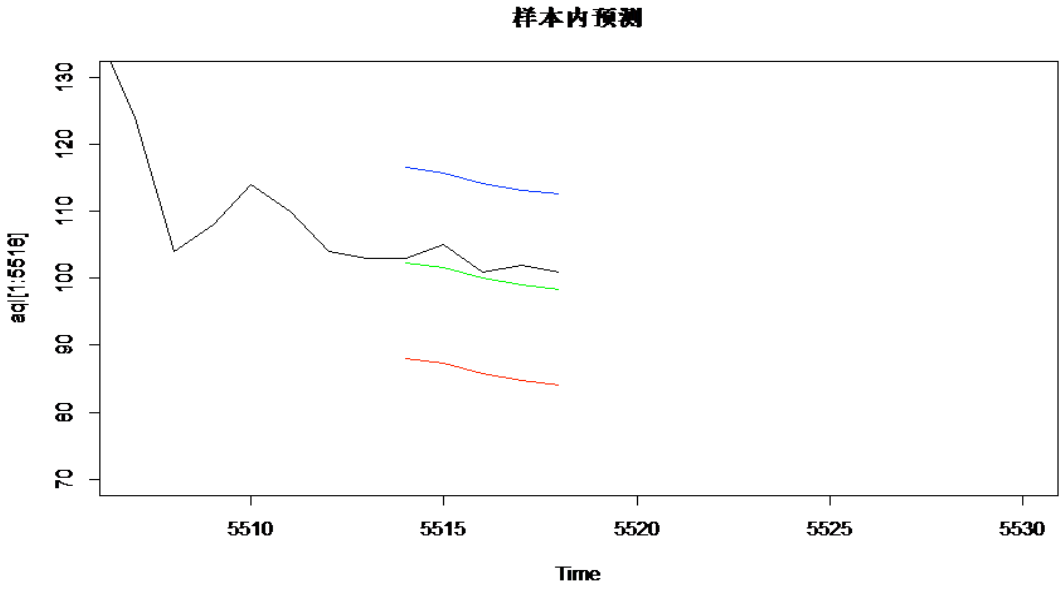

图 10. 模型 ARIMA（3,1,7）预测样本内第 5514 至 5518 个值的时序图和 95%置信带

2. 样本外预测

在样本内预测合理的基础上，我们对数据做出了向前 5 步的样本外预测，得出未来五个小时的空气质量指数，预测结果与原数据的对比如表 7：

表 7. 样本外预测值与实际值的比较

时刻	实际值	预测值	绝对误差
5519	101	100	1
5520	100	98	2
5521	101	97	3
5522	102	96	6
5523	94	96	-2

样本外预测的图形经过放大绘制如图 11，并且附上 95%的置信带。

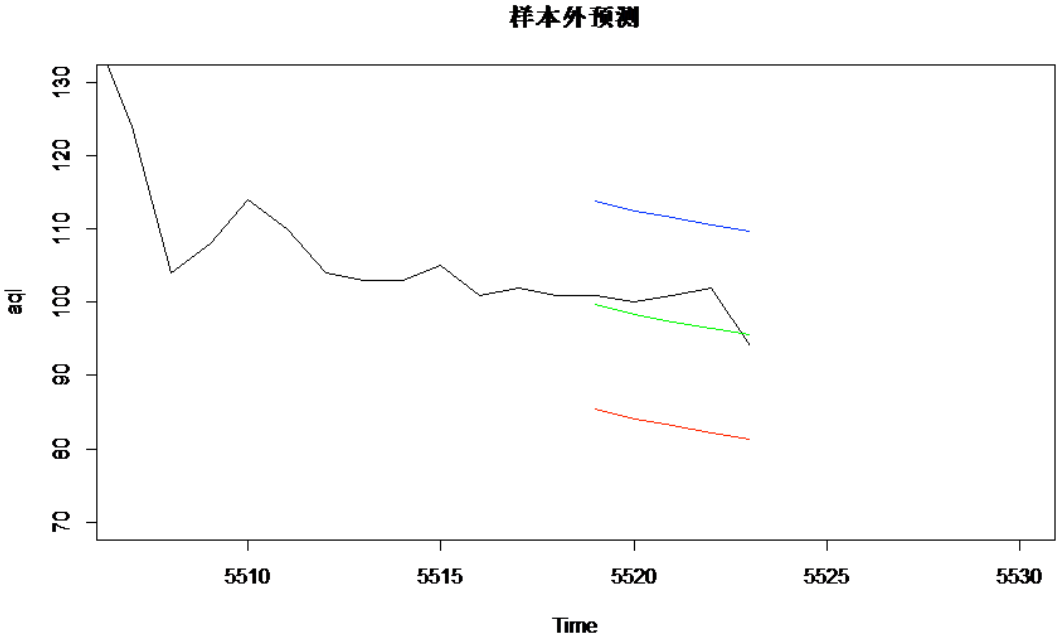

图 11. 模型 ARIMA（3,1,7）预测样本外第 5519 至 5523 个值的时序图和 95%置信带

由图 11 可以看出，预测结果与实际结果的误差在合理范围内，并且可以看出预测结果显示未来几个小时内的空气质量指数有下降的趋势，而实际结果的趋势也是下降的，与预测结果相吻合，因此我们可以认为该模型具有合理性。

六、 相关性分析

为了进一步研究分析影响空气质量指数的因素，我们收集了上海市每日空气质量指数的数据，以及每日温度、每日风力等数据，通过相关性分析的方式研究空气质量指数的其他影响因素。本文所用到的相关性分析的方法主要有：卡方检验，协整和 Granger 因果检验。

卡方检验，又称为相关性检验，是用来检验两个变量之间是否独立（即是否存在相关关系），其原假设是两个变量是相互独立的。

如果两个（或两个以上）的时间序列变量是同阶单整的，它们的某种线性组合是平稳性，则这些变量之间存在长期稳定关系，则称这些时间序列是协整的。

Granger 因果检验是对时间序列之间的领先与滞后关系的检验，检验时间序列在时间上的因果关系，重在影响方向的确认，而非完全的因果关系。Granger

因果检验的基本思想是：对于时间序列 X 和 Y，若 X 的变化引起了 Y 的变化，X 的变化应当在 Y 的变化之前，即 X 是 Y 的 Granger 因（X Granger-(G)-causesY）。反之，若 Y 的变化同样能引起 X 的变化，则 Y 也是 X 的 Granger 因（X Granger-(G)-causesY）。Granger 因果检验原假设是 X 不是引起 Y 变化的原因。通过 Granger 因果检验，我们可以进一步得出两个时间序列之间的相关关系。

（一） 空气质量指数与温度的相关性

首先，对数据进行处理，为了计算出当日的大气温度，我们计算出当日最高温度与最低温度的平均值。由于温度的取值与 AQI 的取值范围相差较大，为了更直观地显示两者的相关性，将平均温度乘上 7 纵向拉升，然后绘制出上海市空气质量指数与当天平均温度时序图，如图 12 所示：

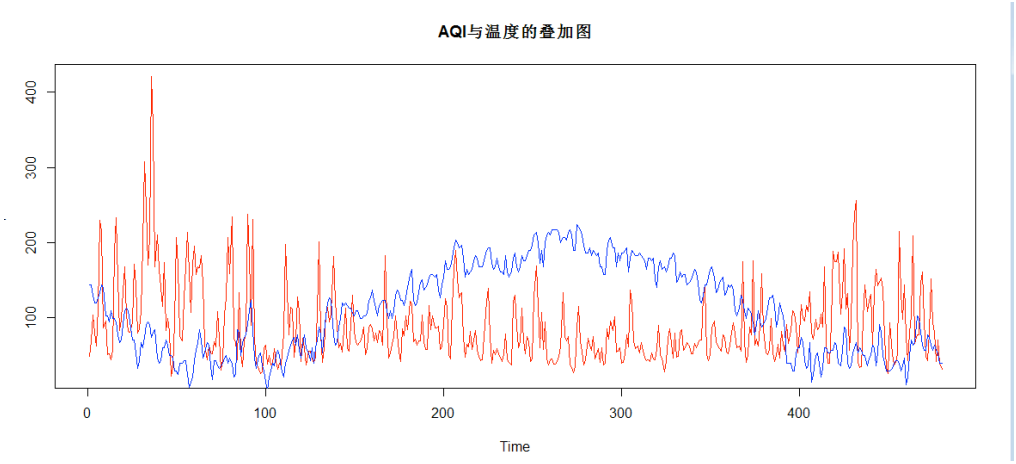

图 12. AQI 与平均温度的叠加时序图

可以直观地看出，当平均温度上升到峰值时（即夏天），AQI 指数会降低到低谷。同样，当平均温度降低到低谷时（即冬天），AQI 指数会上升，形成峰值。因此，我们初步判断，AQI 与平均温度是反向相关的。

在时序图的基础上，我们初步假设空气质量指数与温度之间存在相关关系。接下来将通过卡方检验的方式进一步验证这一假设。通过 $Pearson-\chi^2$ 独立性检验，我们可以计得出下列检验结果（图 13）：

Pearson's Chi-squared test

data: y

X-squared = 4302.1, df = 479, p-value < 2.2e-16

图 13. AQI 与平均温度的 Pearson-χ^2 独立性检验结果

卡方检验的原假设是检验的两者不存在相关关系，由于 P 值很小，所以我们可以认为两者之间存在相关关系。

我们通过协整的方式来验证两者的相关性，为了进行协整检验并构建协整模型，首先检验这两个数据及其一阶差分的单整情况。我们使用了 PP，ADF 及 KPSS 三种检验方法。结果见表 8：

表 8. AQI 与平均温度序列的单整检验

序列	0 阶差分平稳性	一阶差分平稳性
空气质量指数 AQI	不平稳	平稳
平均温度	不平稳	平稳

从检验结果可以看出，空气质量指数与平均温度都是一阶单整的，可以进行协整分析。为了拟合 AQI 与平均温度，我们用 R 语言的线性回归做出线性拟合并得到拟合的残差，并对拟合后的残差进行 ADF 检验，得到的 P 值小于 0.01，可以认为残差是平稳的，即两者之间存在着协整关系。

为了进一步验证上海市空气质量指数与平均温度的关系，我们对两者进行格朗杰因果检验。利用 R 中的 lmtest 包中的 grangertest()函数得到的结果如图 16 所示：

```
> grangertest(aq,tem)
Granger causality test

Model 1: tem ~ Lags(tem, 1:1) + Lags(aq, 1:1)
Model 2: tem ~ Lags(tem, 1:1)
  Res.Df Df      F    Pr(>F)
1    476
2    477 -1 9.2233 0.002521 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> grangertest(tem,aq)
Granger causality test

Model 1: aq ~ Lags(aq, 1:1) + Lags(tem, 1:1)
Model 2: aq ~ Lags(aq, 1:1)
  Res.Df Df      F    Pr(>F)
1    476
2    477 -1 8.2856 0.004176 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

图 14.　AQI 与平均温度 Granger 因果检验结果

图 14 中第一个部分的原假设为“AQI 不是引起温度变化的原因”，从 P 值为 0.0025 远小于临界值，因此可以拒绝原假设，认为 AQI 是引起温度变化的原因。同理，图中第二部分的原假设为“温度不是引起 AQI 变化的原因”，从 P 值为 0.0042 远小于临界值，因此拒绝原假设，认为温度是引起 AQI 变化的原因。由因果关系检验的理论指导，温度和 AQI 有 Granger 因果关系。

（二）　空气质量指数与风力的相关性

首先对数据进行处理，为了对风进行量化分析，我们通过引进了虚拟变量的方法，得到当日的平均风力等级。我们利用每日的风力等级通过如下方法计算平均风力：

如果当天天气预报显示上海为 i 级风，那么记

$$W = i\ ,\ i \in [0,12];$$

如果当天天气预报显示上海的风力为 i~i+1 级风，那么记

$$W = \frac{i + (i+1)}{2}, i \in [0,12];$$

如果当天天气预报显示上海的风力为 i~i+1 级转 j 级风，则记

$$W = [\frac{i + (i+1)}{2} + j]/2 \quad i, j \in [0,12]$$

如果当天天气预报显示上海的风力为 i~i+1 级转 j~j+1 级，那么记

$$W=[\frac{i+(i+1)}{2}+\frac{j+(j+1)}{2}]/2=\frac{(i+j+1)}{2},\quad i,j\in[0,12]$$

通过这样的方法，我们就能够很好的来研究风对空气污染的影响了。由于风力的取值与 AQI 的取值范围相差较大，为了更加直观地观察两者之间的相关关系，将平均风力乘上 35 纵向拉升，绘制了上海市空气质量指数与当天平均风力时序图，如图 15。

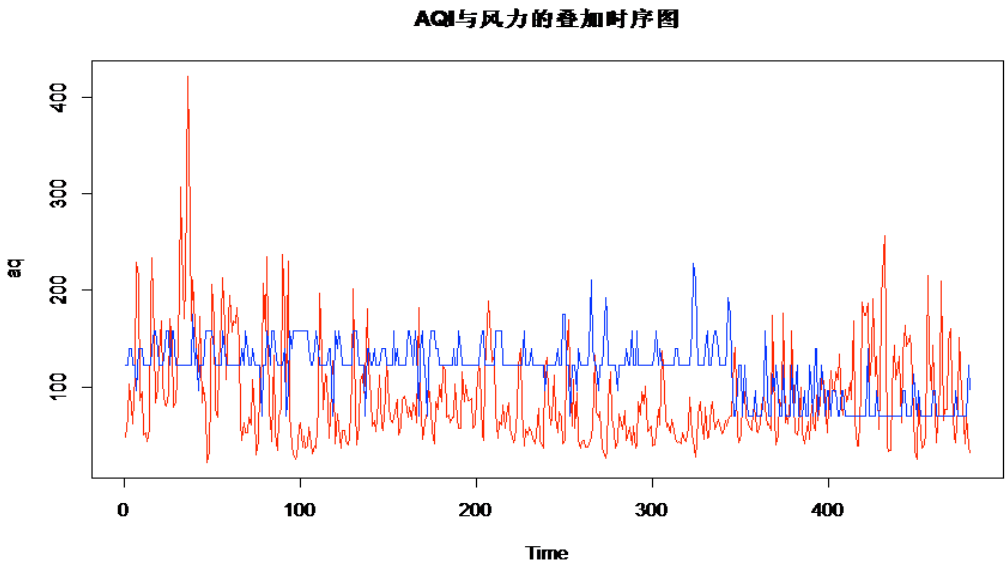

图 15.　AQI 与风力的叠加时序图

可以直观地看出，当平均风力达到一个峰值时，AQI 指数会降低到低谷。同样，当平均风力达到降到一个低谷时，AQI 指数会上升，形成一个峰值。因此，我们初步判断，AQI 与风力也是反向相关的。

同样地，在此基础上，我们初步假设空气质量指数与风力之间存在相关关系，接下来将通过卡方检验的方式进一步验证这一假设。通过 $Pearson-\chi^2$ 独立性检验，我们可以计得出下列检验结果（图 16）：

```
        Pearson's Chi-squared test

data:  y
X-squared = 688.5598, df = 479, p-value = 9.97e-10
```

图 16.　AQI 与风力的 Pearson-χ^2 独立性检验结果

由于 P 值很小，我们可以认为空气质量指数与风力之间存在相关关系。

同样，我们也用协整的方式来验证两者的相关性，首先检验空气质量指数与风力及其一阶差分的单整情况。我们使用了 PP，ADF 及 KPSS 三种检验方法。结果见表 9：

表 9. AQI 与风力序列的单整检验

序列	0 阶差分平稳性	一阶差分平稳性
空气污染指数 AQI	不平稳	平稳
平均风力	不平稳	平稳

从检验结果看，空气质量指数与平均风力都是一阶单整的，可以进行协整分析。为了拟合 AQI 与风力，我们用 R 语言的线性回归做出线性拟合并得到拟合的残差，并对拟合后的残差进行 ADF 检验，得到的 P 值小于 0.01，可以认为残差是平稳的。

两者格朗杰因果检验的结果如图 17。

```
> grangertest(aq,wind)
Granger causality test

Model 1: wind ~ Lags(wind, 1:1) + Lags(aq, 1:1)
Model 2: wind ~ Lags(wind, 1:1)
  Res.Df Df      F Pr(>F)
1    476
2    477 -1 0.0495  0.824
> grangertest(wind,aq)
Granger causality test

Model 1: aq ~ Lags(aq, 1:1) + Lags(wind, 1:1)
Model 2: aq ~ Lags(aq, 1:1)
  Res.Df Df      F Pr(>F)
1    476
2    477 -1 6.3296 0.0122 *
```

图 17.　AQI 与风力 Granger 因果检验结果

通过检验结果可知，第一部分可以看出空气质量指数的下降却不是风力增强的 Granger 因，第二部分可以看出风力的增强是空气质量指数下降的 Granger 因。结合理论知识可知，风对污染物在空气中的稀释扩散起着重要作用，风力的增加导致风速的加大，会使得空气中的污染物向周围流动，从而导致在给定的立体空间区域内污染物的浓度下降，这样就实现了通过风的作用达到了对污染物的稀

释。而风力的大小受到自然界的因素影响很大，很少或者几乎不受到空气质量的影响。由以上检验结果可知，风力也是影响空气质量指数的一个因素。

七、结论

综上所述，上海市空气质量指数的模型为 ARIMA(3,1,7)模型，样本内预测与样本外预测的误差均控制在合理范围内，可以认为该模型有效。在此模型的基础上，我们通过协整和相关性分析研究风力和大气温度等因素对空气质量指数的影响，得出结论：上海市空气质量指数分别与风力和大气温度都存在协整关系，且均为负相关关系。此外，通过 Granger 因果检验我们可以进一步得出结论，大气温度与空气质量指数是协整的，他们是相互影响的。而风力会影响空气质量指数的大小，而空气质量指数并不会直接影响风力的大小。

参考文献

[1] Alireza Rashki&C.J.deW Rautenbach.Temporal changes of particulate concentration in the ambient air over the city of Zahedan, Iran,2013

[2] Anikender Kumar&P. Goyal.Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis,2013

[3] To, Teresa&Shen, Shixin.The Air Quality Health Index and Asthma Morbidity: A Population-Based Study,2013

[4] 齐国辉.城市空气污染治理的有效途径与探讨 2013.12

[5] 任毅斌.基于伦敦治污经验的中国城市空气污染治理探讨 2013.9

[6] 约翰·福克斯著,於嘉译.回归诊断简介. 上海人民出版社,2012

[7] 王惠文,孟洁.变量筛选、模型分类及自动化建模方法.北京:科学出版社,2013

[8] 吴喜之,刘苗编著.应用时间序列分析:R 软件陪同.北京:机械工业出版社,2014