北京市 2016～2018 年空气污染指标 AQI 值预测

应用统计学 17 杨谨行 2017312292

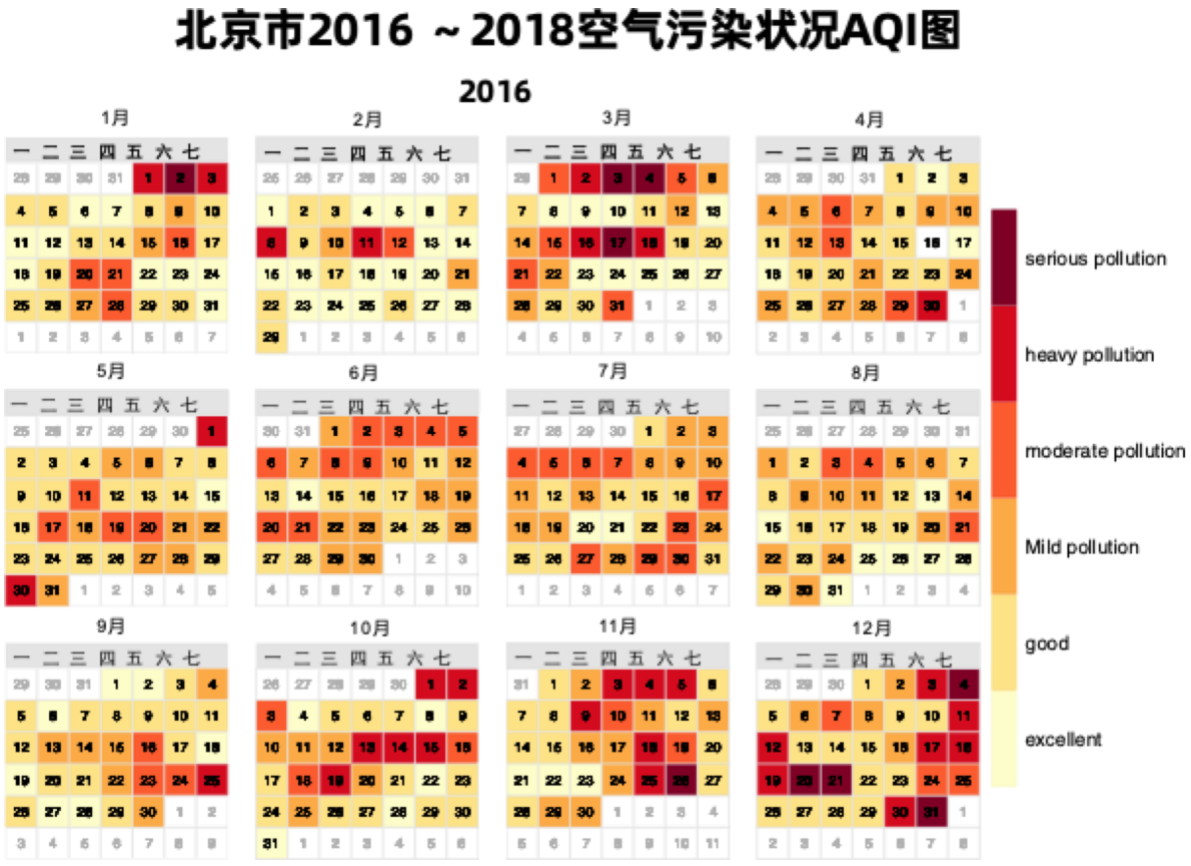
2018 年 9 月北京公布《北京市打赢蓝天保卫战三年行动计划》提出了“十三五”规划目标，空气污染治理就此拉开新一轮的帷幕。目前上次行动计划已近尾声，北京交上了一份漂亮的答卷，下面我们就 2016 年～2018 年北京空气污染的数据进行建模分析，对空气污染指数 AQI 进行预测。

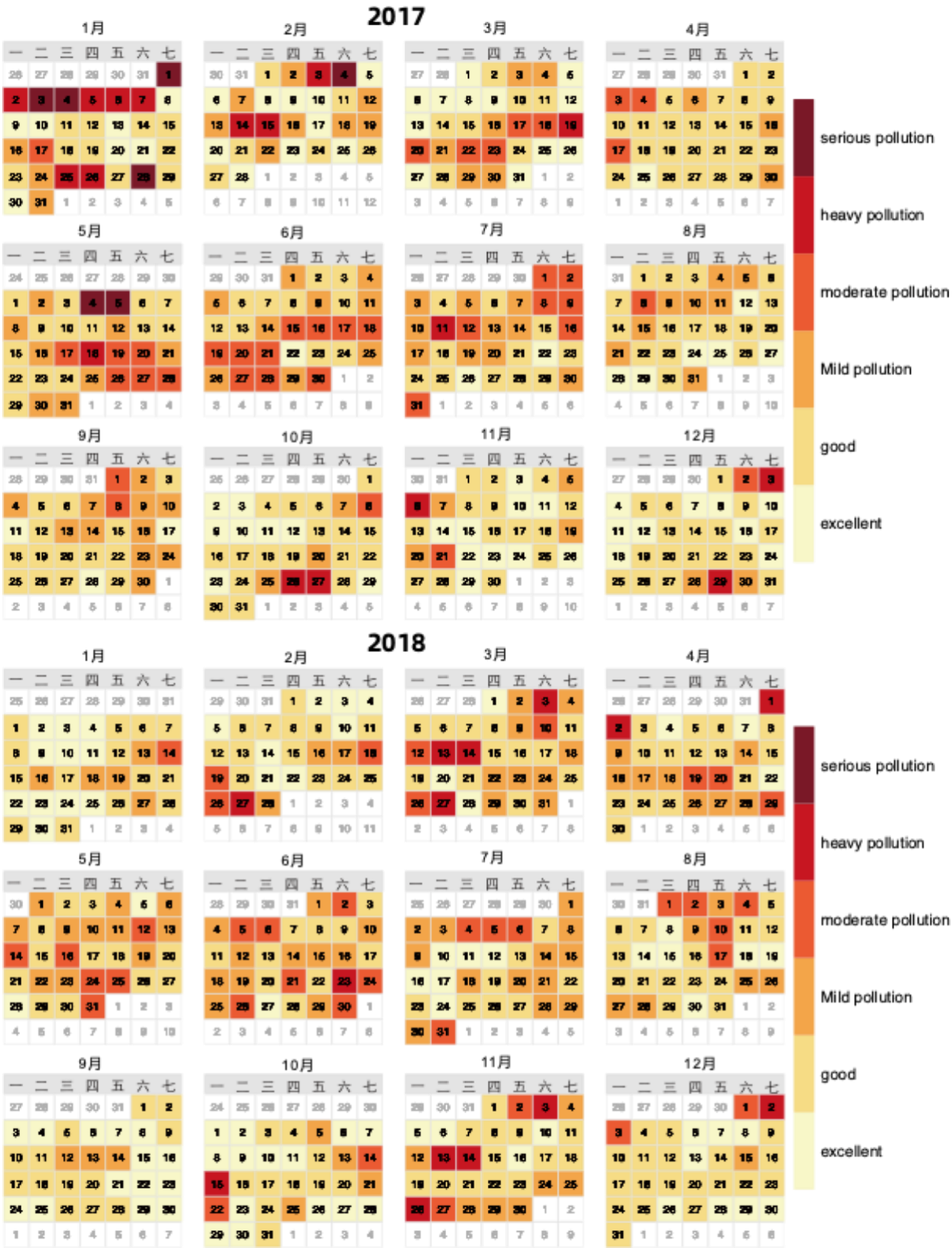
数据来自中国空气质量在线监测平台（https://www.aqistudy.cn/），数据包含每日的空气污染指标 AQI 值，以及 AQI、PM2.5、PM10、SO2、CO、NO2、O3 等空气污染物的指数，是较为典型的时间序列。

一、数据展示与描述性统计

1、北京市 2016~2018 三年度空气质量数据的年度日历热图

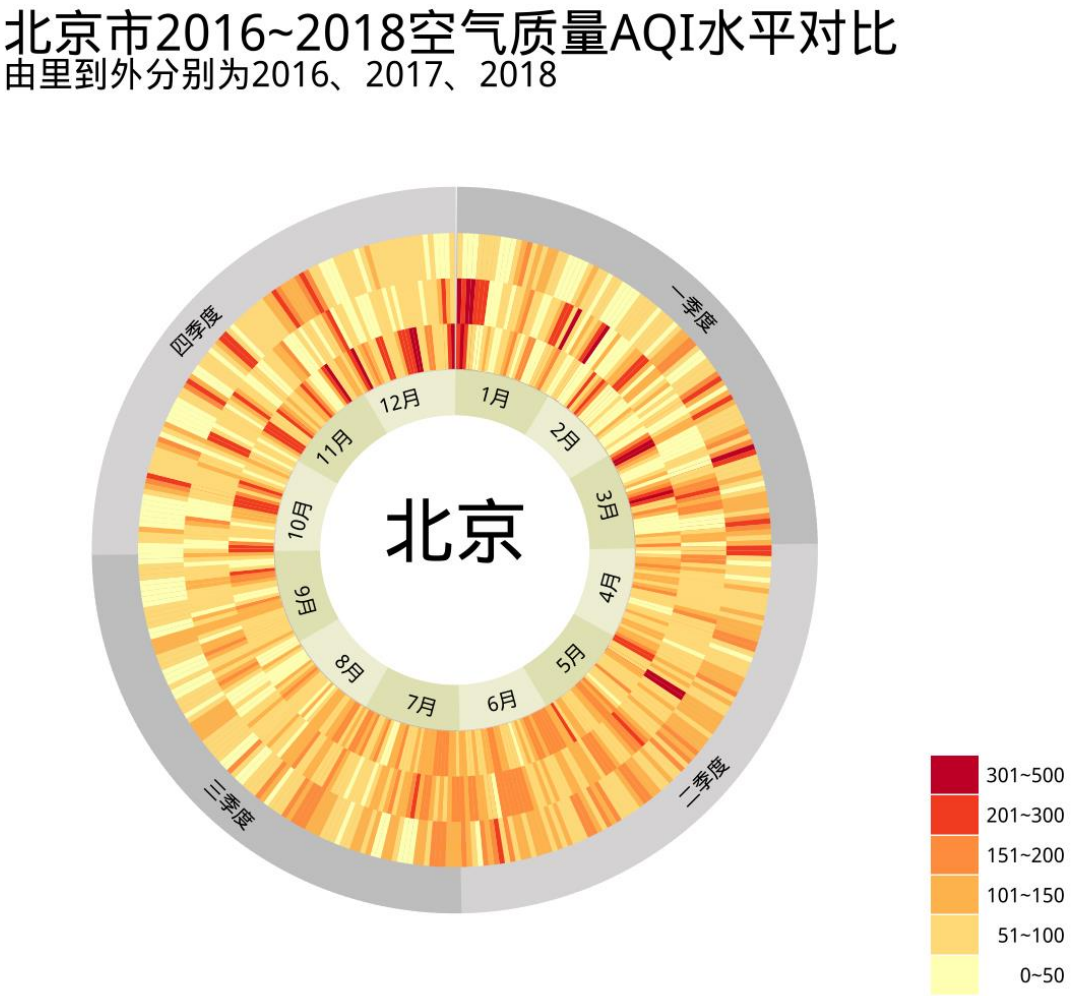
使用日历的形式，用颜色为度表示 AQI 数据如下：




2017
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月
serious pollution
heavy pollution
moderate pollution
Mild pollution
good
excellent
2018
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月
serious pollution
heavy pollution
moderate pollution
Mild pollution
good
excellent

2、2016 年～2018 年每月、季横向对比可视化

进一步，下图基于 ggplot2 的极坐标变换，将北京三年空气质量指标 AQI 值，共 1095 个数据点完全呈现在了一张图表上[1]

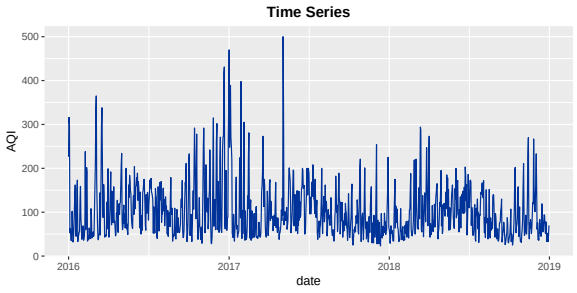


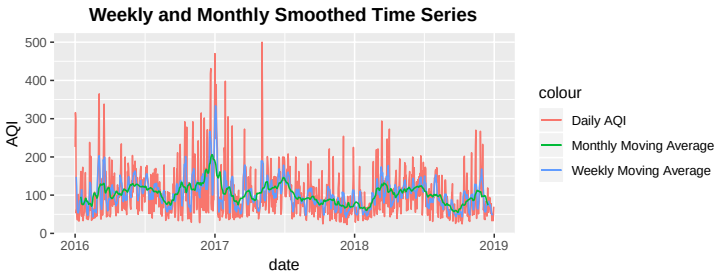
从图中我们可以看出，相较于 2017 年和 2016 年，2018 年的空气污染在最高值方面有所改善，没有出现 AQI 超过 300 的“严重污染”情况；不同年份同期对比显示，12 月～1 月以及 3 月初，往往是污染相对严重的时期（2018 年的冬季有显著改善）；而 6～7 月则长期处于有一定污染，9 月～10 月秋季通常空气质量较好。这也揭示了数据时间序列中的季节性因素，故在建模时笔者首先考虑了时间序列模型。

[1] 注：因作图需要，改图可视化过程中去掉 2016 作为闰年 2 月 29 日的数据，并对极个别日期进行微调整。

3、时间序列图

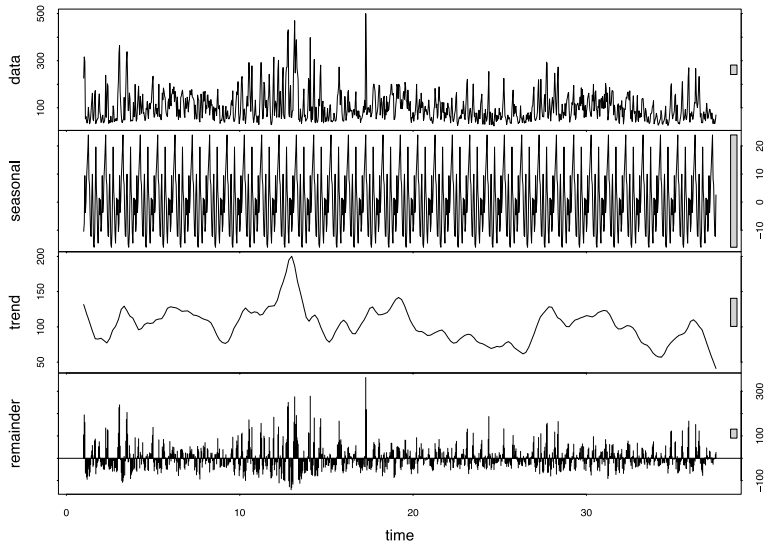
时间序列中图显示，数据波动较大，进一步分析每周和每月的移动平均曲线，可以看到数据集中在 100 上下波动。



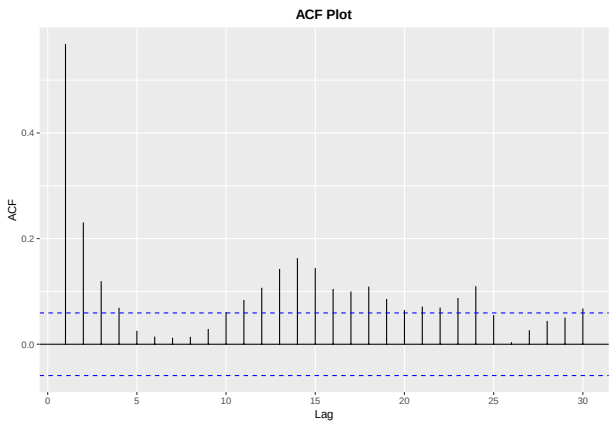


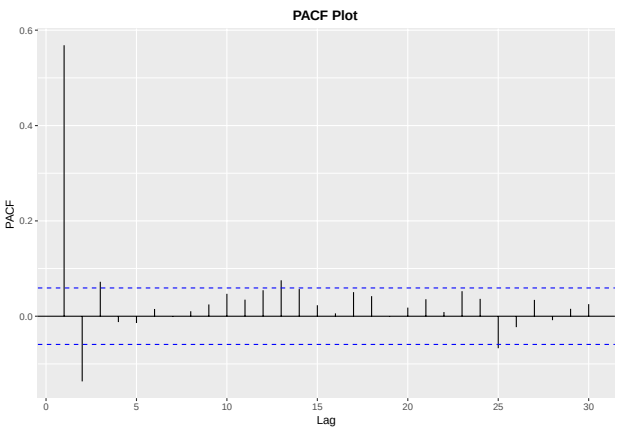
二、 模型一：Autoregressive Integrated Moving Average model（ARIMA）

模型建立前先经过 ADF 稳定性检验，得到的 p 值为 0.01，说明数据是平稳性的，将时间序列拆解为季节性因素和趋势波动后，可以得到：



从 ACF 和 PACF 图中，可以考到 2 或 3 阶截尾，可以使用 AR(2)或 AR(3)





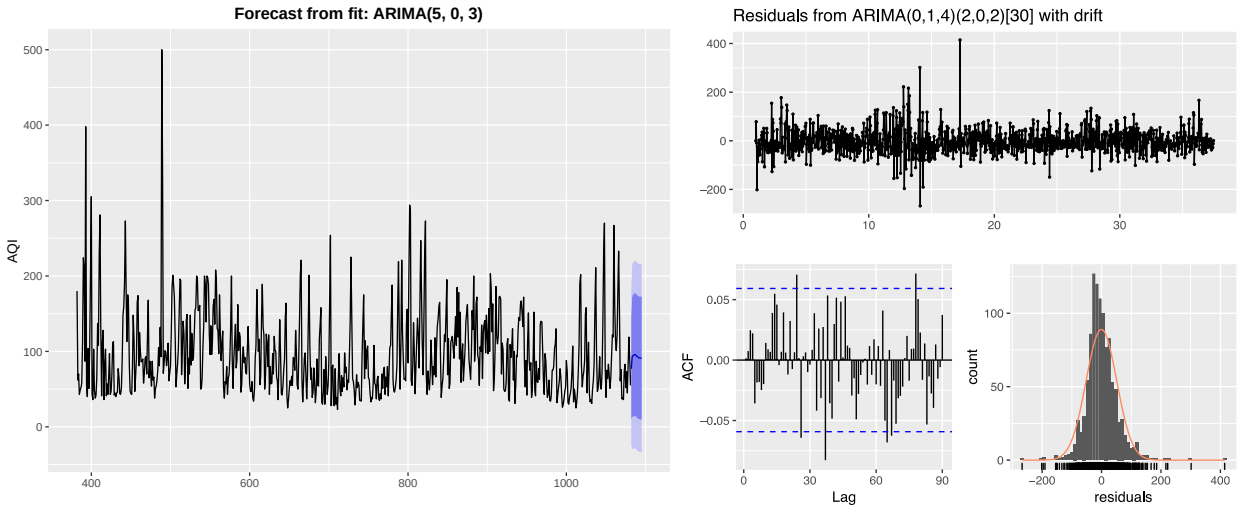
再综合三个不同训练集平均值区间下，auto.ARIMA()方法的输出效果比较：

	ARIMA (2,0,0)	ARIMA (3,0,0)	**ARIMA (5,1,3)**	ARIMA(1,1,2)(1,0,1)[30]	ARIMA(3,1,1)(0,0,1)[90]
AIC	11620.03	11616.46	11606.86	11793.76	11792.91
Log likelihood	-5806.02	-5803.23	-5794.43	-5890.88	-5890.45

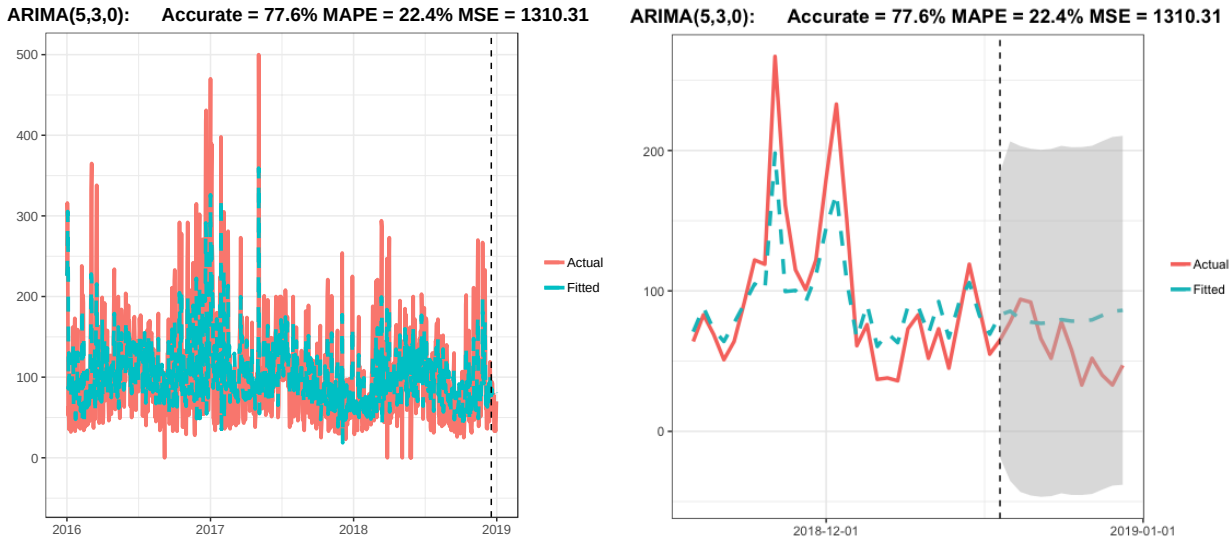
选择 AIC 值最小、最大似然估计值的 ARIMA(5,1,3)建立如下模型：

$$(1-\varphi_1 B^1-\varphi_2 B^2-\varphi_3 B^3-\varphi_4 B^4-\varphi_5 B^5)(1-B)y_t = c+(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)$$

对最后 14 天数据进行点估计和区间估计得到结果：



残差检测中，Box.test 的 p 值为 0.9617，说明季节性因素已经消除，残擦为白噪声。对测试集拟合，与真实值比较得：



注：准确率计算方式为 1-平均绝对百分误差（ MAPE ）

尽管得到了不错的结果，但 ARIMA 模型估计的波动范围较小，与真实情况有出入，下面使用贝叶斯模型进一步探究。

三、贝叶斯

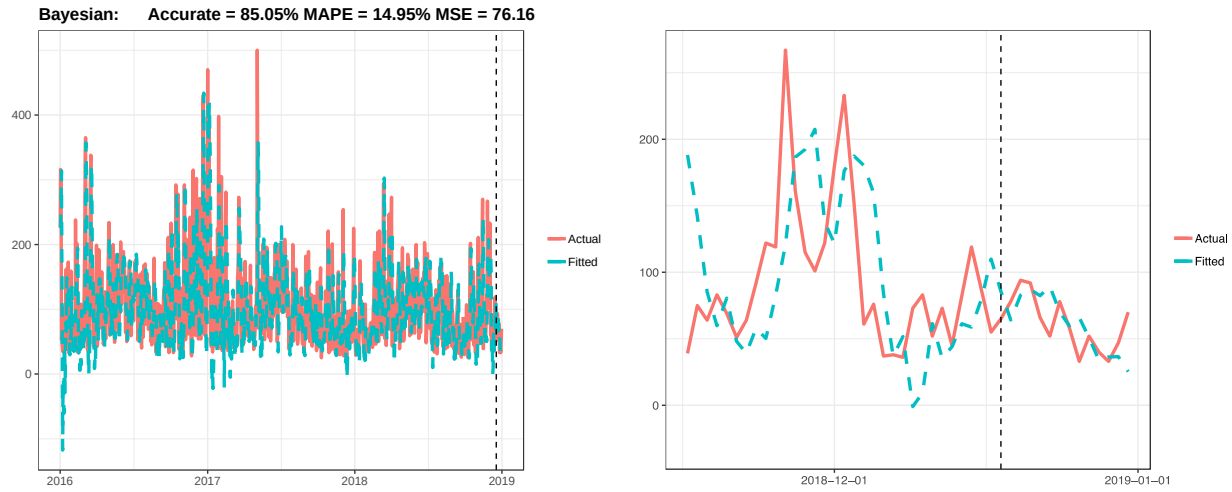
1、模型形式

$$Y_t=\mu_t+x_t\beta+S_t+e_t, e_t\sim N(0,\sigma_e^2)$$

$$\mu_{t+1}=\mu_t+\nu_t, \nu_t\sim N(0,\sigma_\nu^2).$$

- xt：表示一组回归向量
- St：表示季节性
- μt：局部水平量，定义了潜在状态随着时间的变化，表示为观察到的趋势

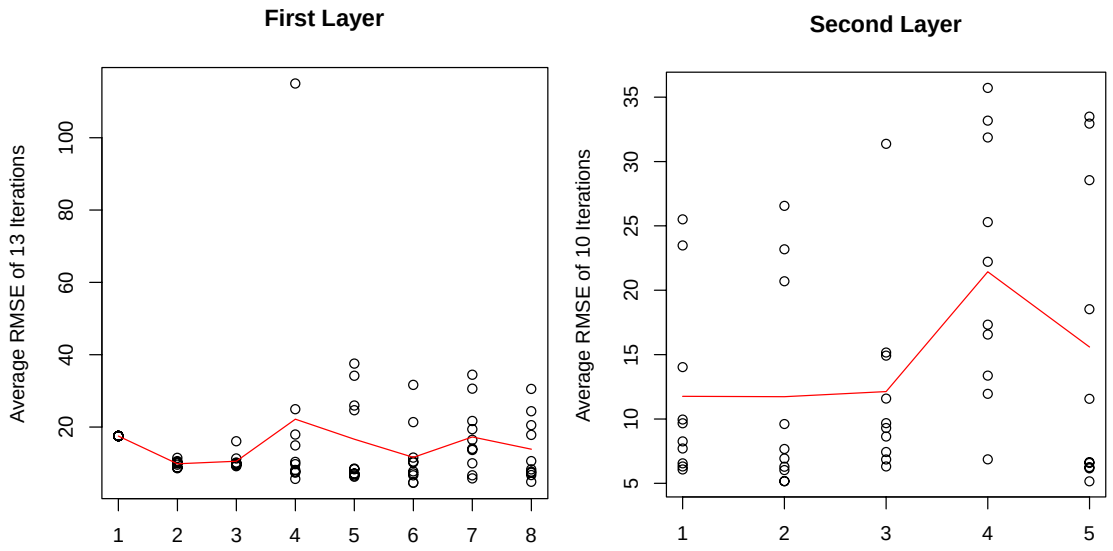
2、拟合结果



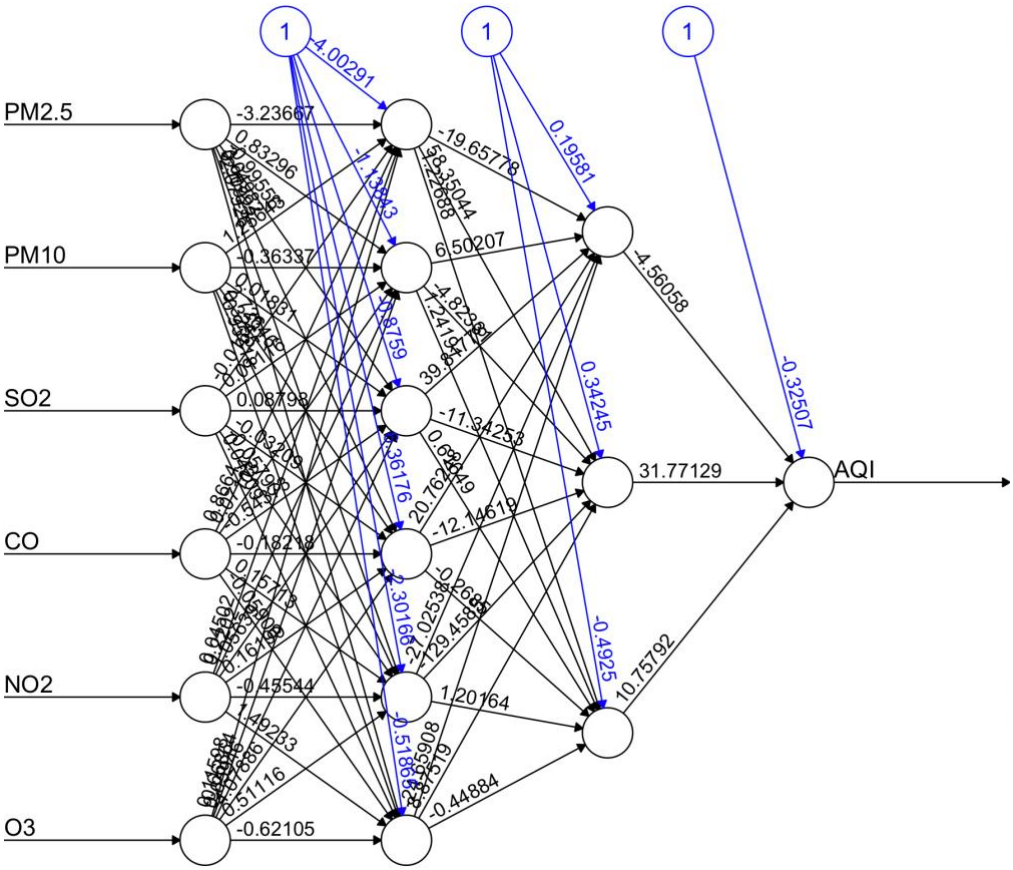
贝叶斯模型存在一定的过拟合问题，准确率并没有相较于 ARIMA 模型提高，由于缺乏先验信息，模型优化较难完成，下面考虑使用两个神经网络模型来进一步提高预测的准确率。

四、神经网络

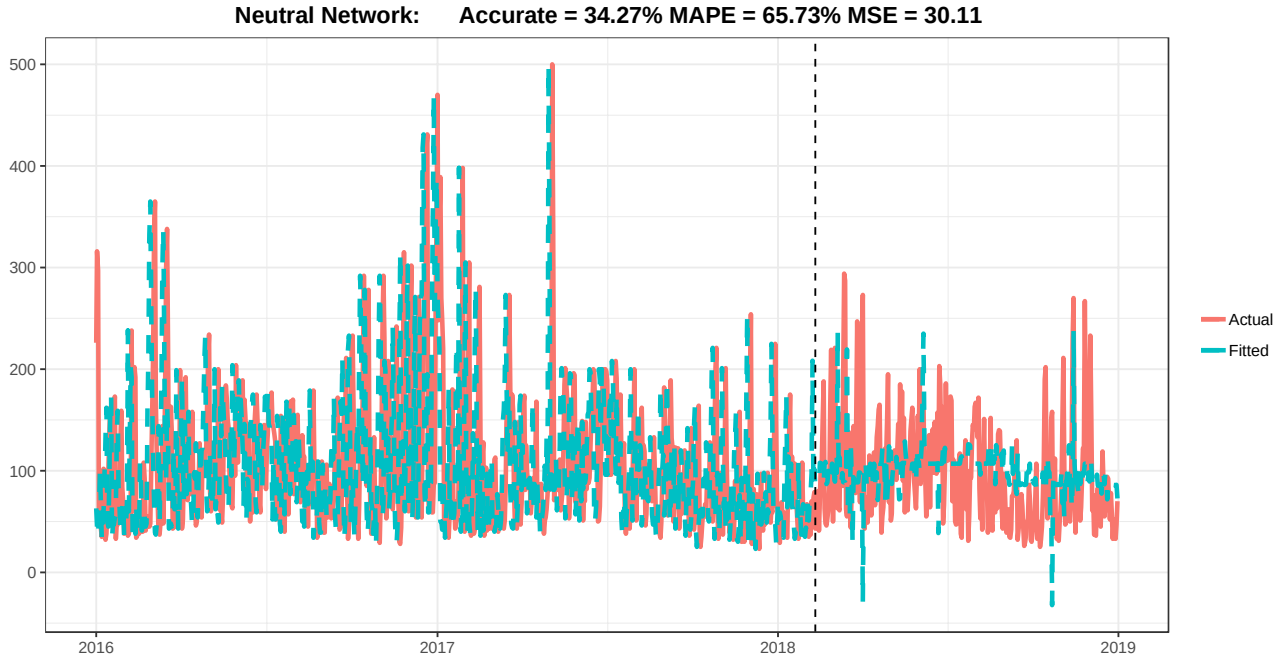
首先，在基础神经网络模型上进行简单的参数搜索，得到两层神经网络最佳的节点数，让隐藏层的层次设计尽可能提高模型拟合准确度。



根据上述的结果，选择第一层为 6，第二层 4 为得到了神经网络模型结构图如下：



预测的结果得到 34.27%的总体精确率：



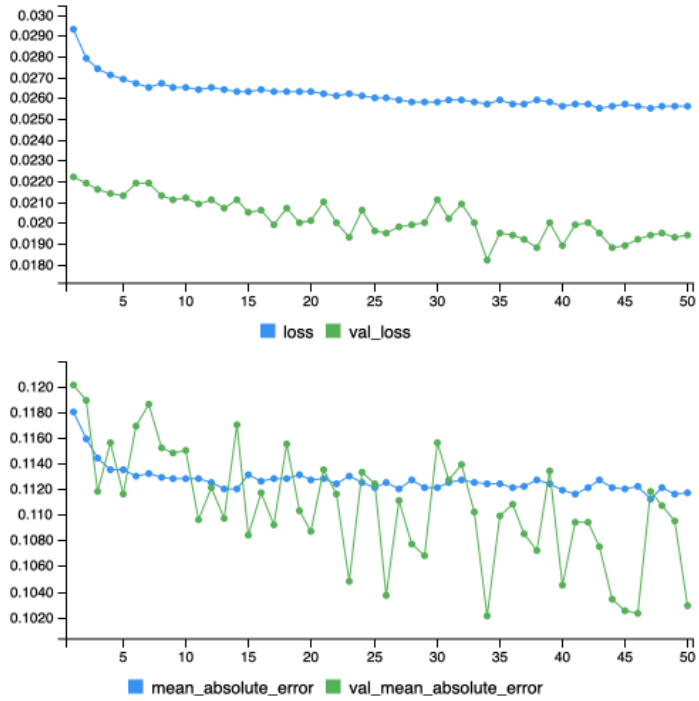
在参数搜索的时候，受到计算能力的限制，迭代次数有限，超参数问题无法解决，也与数据本身适应程度有关，普通的神经网络模型的准确率仍然不足，因此在此基础上，引入新的算法——LSTM。

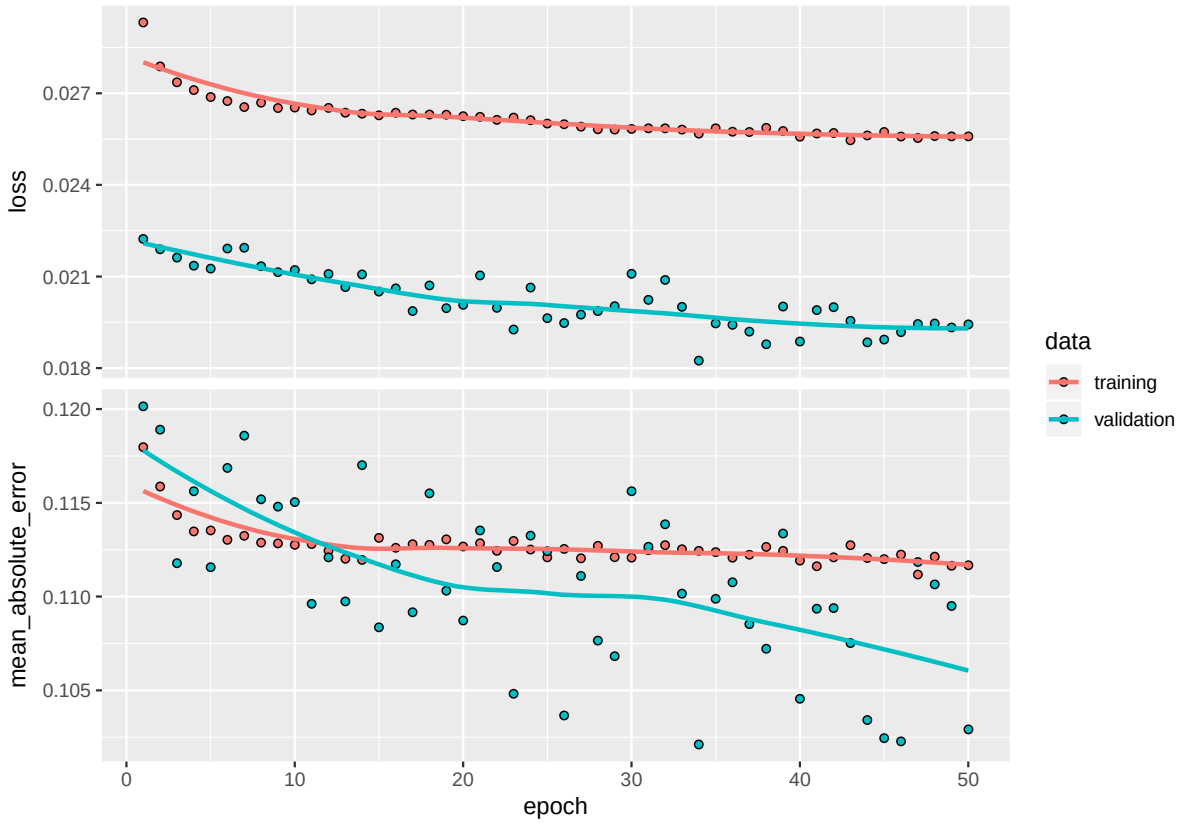
五、Long Short-term Memory（LSTM）

LSTM 是在循环神经网络（RNN）上，进行改进，可以很大程度上避免常规 RNN 的梯度消失问题。使用 LSTM 对模型拟合的结果如下：

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (1, 1)                    12
_________________________________________________________________
dense_1 (Dense)              (1, 1)                    2
=================================================================
Total params: 14
Trainable params: 14
Non-trainable params: 0
```

在该模型的基础上，对原始数据进行 50 次拟合，模型的损失值逐渐收敛，而 MAPE 和标准误则呈震荡收敛，MAPE 值总体保持在 11%左右的水平，模型基本稳定。





模型的预测结果：以 1-MAPE 衡量的准确率达到 92.72%，为前述四个模型中最准确、稳定的，对于北京市空气污染数据而言，LSTM 为最有效的模型，预测结果如下：