

# 中国城市降雨量时间序列分析

学生：孟宣佑 孟禹含 张雪茜

专业：统计学

学院：统计学院

指导老师：陆智萍

2018 年 6 月 14 日

## 摘要

月均降雨量时间序列具有明显的季节性，因此季节效应分析具有实际意义。为了更好地预测全国各城市降雨量分布，防范自然灾害，本文由南至北选择三个典型城市（广州、上海、哈尔滨）1997-2016 年的月均降雨量数据进行时间序列分析与建模，并基于最优模型进行静态与动态预测。由于三个地区降雨量数据存在差异，本文对三大城市数据通过 R 统计软件拟合多种模型，建模过程主要为以下步骤。

第一，对各城市月均降雨量数据进行总体描述性统计分析，并进行时间序列基本分析及预处理。基本分析如平稳性检验，相关性检验，谱图分析等。预处理主要包括信号分解、STL 分解，以及 12 步差分法。基于其可视化结果，确定序列总体特征——即无明显趋势但具有强季节性。提取或去除季节性影响，同时引入 Holt-Winters 三次指数平滑法进行初步拟合。

第二，构建简单季节模型与复杂季节模型。对去除季节性影响的数据进一步进行 AR,MA,ARIMA 的定阶与拟合，定阶方法均采取 AIC 最小原则。由于拟合模型存在模型不显著或阶数过高的问题，本文采取 SARIMA 复杂季节模型。

第三，检验复杂季节模型残差的 ARCH 效应。采用残差平方的 LB 检验，三城市中仅哈尔滨市的 SARIMA 模型残差具有 ARCH 效应，因此对其残差构建 ARCH、GARCH 模型。通过参数显著性、无条件方差收敛性、模型阶数有限性以及 ARCH 效应去除完成度四个准则对 ARCH、GARCH 模型进行定阶，优化其复杂季节模型。

第四，前馈神经网络模型的应用。对上海市降雨量数据，引入同时段的温度、相对湿度，以及年份、月份四个相关变量，结合试验选择法与逐步增加法确定隐含层节点数，并通过权重衰减精度以及最大迭代次数的限制，建立了一个具有相对高泛化能力和输出精度的前馈神经网络模型。

最后，本文基于各模型预测能力的比对分析，确定最优模型并进行样本内与样本外的预测。样本内预测包括各模型的动态预测与静态预测，其中对哈尔滨市的动态预测引入条件异方差性的影响。

**关键词：** STL 分解 Holt-Winters 复杂季节模型 GARCH 模型 前馈神经网络

# 目录

1 绪论 .....	1
1.1 研究背景及意义 .....	1
1.2 国内外模型研究现状 .....	1
1.2.1 STL 分解 .....	1
1.2.2 Holt-Winters 模型 .....	1
1.2.3 简单季节模型与复杂季节模型 .....	1
1.2.4 GARCH 模型 .....	2
1.2.5 前馈神经网络(FFNN)模型 .....	2
2 时间序列分析的模型建模方法及介绍 .....	3
2.1 信号分解 .....	3
2.2 STL 分解 .....	3
2.3 Holt-Winters 季节模型 .....	4
2.4 Loess 加法模型及混合模型 .....	4
2.5 简单季节模型 .....	5
2.5.1 AR 及 MA 模型 .....	5
2.5.2 ARMA 模型 .....	5
2.5.3 ARIMA 模型 .....	6
2.6 复杂季节模型 .....	6
2.7 GARCH 模型 .....	7
2.8 FFNN 模型 .....	7
3 数据来源及特征描述 .....	8
3.1 数据来源及数据补全 .....	8
3.2 描述性统计 .....	9
3.3 平稳性分析及长记忆性分析 .....	10
3.4 白噪声检验 .....	11
4 模型建立 .....	12
4.1 季节效应分析 .....	12
4.1.1 趋势分析 .....	12

4.1.2 信号分解 .....	13
4.1.2.1 Decompose 分解 .....	13
4.1.2.2 STL 分解 .....	14
4.2 Loess 加法模型及混合模型 .....	14
4.3 简单季节模型 .....	15
4.3.1 AR 及 MA 模型 .....	15
4.3.2 ARMA 模型 .....	16
4.3.3 模型显著性检验 .....	16
4.4 复杂季节模型 .....	17
4.5 GARCH 模型 .....	19
4.5.1 ARCH 效应检验 .....	19
4.5.2 GARCH(p,q)定阶 .....	19
4.5.2.1 ARCH(1)模型 .....	20
4.5.2.2 GARCH(p,q)模型 .....	21
4.6 FFNN 模型 .....	22
4.7 本章小结 .....	24
5 最优模型预测 .....	24
5.1 广州市降雨量预测——SARIMA 模型 .....	24
5.1.1 样本内预测 .....	24
5.1.2 样本外预测 .....	25
5.2 哈尔滨市降雨量预测——SARIMA+ARCH 模型 .....	25
5.2.1 样本内预测 .....	25
5.2.2 样本外预测 .....	28
5.3 上海市降雨量预测——FFNN 模型 .....	28
5.3.1 样本内预测 .....	28
5.3.2 样本外预测 .....	29
6 总结与展望 .....	30
6.1 总结 .....	30
6.2 创新点 .....	30
6.3 展望 .....	31

# 1. 绪论

## 1.1 研究背景及意义

降雨量是对天空降落下来的雨雪冰雹等液态或者固态水量的观测和记录。而降雨量的记录是以承受降水的地点水平面上积累的水层深度表示。降雨量的观测是水资源的重要基础资料之一，也是研究流域或者一个地区水文循环的重要指标。降雨量的多少极大影响到人类的生产生活等各个领域。对于工农业生产，若降水过少则会增加生活用水和农田灌溉用水还有工厂用水之间的矛盾；关于自然灾害，若降水量过多，会造成洪涝灾害、滑坡、泥石流等灾害，因此需要及时做好江河防洪。为了更好地预测全国各地区降雨量分布情况，实现自然灾害的及时预防，由北至南本文分别选择了哈尔滨市，上海市和广州市三个地区 1997-2016 年 20 年间的月均降雨量数据进行时间序列分析，选出可拟合的最优模型并作出相应预测。

## 1.2 国内外模型研究现状

### 1.2.1 STL 分解

STL 分解的核心思想——鲁棒局部加权回归算法，由 Cleveland W S. 于 1979 年提出，主要利用局部观测数据对欲拟合点进行多项式加权拟合，并用最小二乘法进行估计。它结合了传统线性回归的简洁性和非线性回归的灵活性，首次将鲁棒权重引入时间序列分解，有效规避了异常值对趋势项与季节项的影响，从而提高相应估计值的稳健性。

### 1.2.2 Holt-Winters 模型

近年来指数平滑方法由于其简单性和良好的综合性能而广泛使用。常见的应用范围从业务任务（例如预测销售或库存波动）到环境研究（例如大气成分测量或降雨数据）。这些研究领域通常不会具备比季节性模式趋势分析更多的先验知识，因此这样的方法有时也被称为 naïve。因为在模型中没有使用协变量，即数据被假定为是自解释的。他们的成功源自于它们属于一类局部模型，它们在估计过程中自动调整其参数，从而隐含地解释训练数据的结构变化。而且，由于新数据的影响是由超参数控制的，使其具有平滑原始时间序列的效果。当季节性变化在整个系列中大致保持不变时，加法方法是优选的，当季节性变化与时间序列的水平成比例变化时使用乘法方法。ETS 功能是 Holt-Winters 模型的特例，它使用先进的机器学习算法，如三重指数平滑。国外许多研究员研制出很多内置的 R 包。其中的 ETS 功能可帮助学者从 20 种可用选项中选择正确的型号。

### 1.2.3 简单季节模型与复杂季节模型

上个世纪英国统计学家 G.U.Yule 提出了自回归 AR(p)模型（G.U.Yule 等，1927），随后英国天文学家、数学家 G.T.Walker 在对印度的大气变化规律分析时采用了移动平均 MA(q)模型和自回归移动平均模型（ARMA(p,q)）（G.T.Walker 等，1931）。这些模型奠定了在时域上对时间序列进行分析的基础。美国统计学家 G.E.P.Box 和英国的统计学家

G.M.Jenkins 提出了对 ARIMA 模型的识别、估计、预测和检验方法。(G.E.P.Box 等, 1970)。随着专家学者对于事件序列更加深入的研究和探索, 黄文杰等通过对上海市气温的月度数据进行拟合与预测, 提出 SARIMA 模型对每年数据进行预测(黄文杰等, 1980)。随后, 田苗、黄毅、沈艳等运用季节差分自回归滑动平均模型进行拟合与预测(田苗等, 1980; 黄毅等, 2012; 沈艳等, 2012)。

#### 1.2.4 GARCH 模型

自回归条件异方差(ARCH)模型是由 Robert Engle 于 1982 年最早提出的, 模型提出后许多研究者对其进行了改进与延伸, 包括 GARCH、IGARCH、EGARCH、GARCH-M 等模型。2003 年, Engle 凭借 ARCH 模型获得了诺贝尔经济学奖。

近几年来, 由于市场频繁变化导致金融危机和高度波动, 使用非平稳模型对金融市场波动进行建模受到了广泛关注。在 Mandelbrot(1963)和 Fama(1965)的开创性工作之后, 许多研究人员发现股票收益的经验分布为非正态分布。(1)股票收益率时间序列的峰度明显大于正态分布的峰度;(2)股票收益的分布偏向右侧(正偏态)或左侧(负偏态);(3)股票收益的方差随时间变化, 或者波动是丛聚的。一些研究人员认为这是股市波动的持续性, 而金融分析师称其为不确定性或者风险。这种不确定性在现代金融理论中至关重要。

许多学者认为传统的时间序列和计量经济学模型只有在方差不变时才有效。直到最近, 金融和经济研究人员已经开始对二阶或更高阶矩的时间变化进行建模。Engle(1982)使用自回归条件异方差(ARCH)及相应扩展和修正的模型来表示变化的方差。自此, 数百名研究人员将这些模型应用于金融时间序列数据。

#### 1.2.5 前馈神经网络(FFNN)模型

前馈神经网络是一种最简单的神经网络, 也是目前神经网络研究与应用中最基本的网络之一, 其研究已有很长的历史。1958 年, 心理学家 Rosenblatt 提出了最早的感知器(Perceptron)模型, 在此基础上, 1965 年, Nilsson 提出了多层网络构想。1974 年, Werbos 提出了误差反向传播(Back Propagation Error, BP)算法。1986 年, 基于 Werbos 算法, Rumelhart 和 Mc-Clelland 等人提出了并行分布处理理论(PDP), 进一步发展并完善了多层网络 BP 算法, 由此对前馈神经网络的研究和应用掀起了新的高潮。

越来越多的研究表明, 结合多种优化方法, 前馈神经网络模型具有良好的非线性性质、灵活有效的学习方式以及完全分布式的存储结构, 对实现大规模并行信息处理以及非线性模拟具有里程碑意义。目前, 前馈神经网络被广泛应用于系统辨识、模式识别、信号处理、故障诊断及经济预测等领域。

## 2 时间序列分析的模型建模方法及介绍

### 2.1 信号分解

Decompose 分解法使用移动平均值将时间序列分解为季节性、趋势和不规则分量。常采用加法或乘法季节性模型。

使用的加法模型是：

$$x_t = T_t + S_t + I_t$$

使用的乘法模型是：

$$x_t = T_t \cdot S_t \cdot I_t$$

信号分解首先使用移动平均值确定趋势分量，并将其从时间序列中移除。然后，通过每个时间单位对所有时期进行平均来计算季节指数。所谓季节指数就是用简单平均法计算的周期内各时期季节性影响的相对数。季节指数反映了该季度与总平均值之间的一种比较稳定的关系。最后通过从原始时间序列中删除趋势和季节指数（根据需要再循环）来确定误差成分。

### 2.2 STL 分解

STL(Seasonal-Trend decomposition procedure based on Loess)基于 Loess 回归将时间序列分解为趋势项(trend component)、季节项(seasonal component)和余项(remainder component):  $Y_v = T_v + S_v + R_v (v=1, \dots, N)$ ，是一种以鲁棒局部加权回归作为平滑方法的时间序列分解方法。

STL 分为内循环（inner loop）与外循环（outer loop），其中内循环主要进行趋势拟合与周期分量的计算，外循环主要用于调节鲁棒权重（robustness weight）。

假定  $T_v^{(k)}$ 、 $S_v^{(k)}$  为内循环中第  $k-1$  次循环结束时的趋势分量、周期分量，设定初始值  $T_v^{(k)}=0$  以及参数：内层循环数  $n(i)$ ，外层循环数  $n(o)$ ，一个周期的样本数  $n(p)$ ，一次 Loess 平滑参数  $n(s)$ ，二次 Loess 平滑参数  $n(l)$ ，三次 Loess 平滑参数  $n(t)$ 。将每个周期相同位置的样本点组成一个子序列，则共有  $n(p)$  个子序列，称其为循环子列（cycle-subseries）。内循环主要分为以下六个步骤：

Step 1: 去除趋势（Detrending）。原始序列减去上一轮结果的趋势分量，得到  $Y_v - T_v^{(k)}$ ；

Step 2: 周期子序列平滑（Cycle-subseries smoothing）。用平滑参数为  $n(s)$  的 Loess 对每个子序列做回归，并向前向后各延展一个周期，平滑结果组成暂时性季节序列，记作  $C_v^{(k+1)}$ ；

Step 3: 周期子序列的低通量过滤（Low-Pass Filtering）。对上一步的结果序列  $C_v^{(k+1)}$  依次做长度为  $n(p)$ 、 $n(p)$ 、3 的滑动平均（moving average），然后进行平滑参数为  $n(l)$  的 Loess 回归，得到结果序列  $L_v^{(k+1)}$ ， $v=1, \dots, N$ ；

Step 4: 去除平滑周期子序列趋势（Detrending of Smoothed Cycle-subseries），得到  $S_v^{(k+1)} = C_v^{(k+1)} - L_v^{(k+1)}$ 。

Step 5: 去除周期（Deseasonalizing）。原始序列减去周期分量，得到  $Y_v - S_v^{(k+1)}$ ；

Step 6: 趋势平滑 (Trend Smoothing), 对于去除周期之后的序列进行平滑参数为 $n_{(t)}$ 的 Loess 回归, 得到趋势分量 $T_v^{(k+1)}$ 。

外循环主要用于调节鲁棒权重, 定义  $h = 6 \times \text{median}(|R_v|)$ 。位置位于  $v$  的数据点的鲁棒权重设为 $\rho_v = B(|R_v|/h)$ , 其中 $B$ 为双平方函数:

$$B(u) = \begin{cases} (1-u^2)^2, & 0 \leq u < 1 \\ 0, & u \geq 1 \end{cases}$$

每一次迭代的内循环中, 在第二步与第六步中以鲁棒权重 $\rho_v$ 进行鲁棒局部加权回归, 以减少异常值对回归的影响。

## 2.3 Holt-Winters 季节模型

Holt-Winters 季节模型是一种指数平滑方法, 可以同时修正序列的季节性和倾向性, 被看作是简单 ES 的一个普通扩展, 并带有一个额外的趋势成分。在 Holt-Winters 季节模型中有三个部分必须在每个时期进行更新——水平, 趋势和季节效应。水平是每个期末数据价值的平滑估计, 而趋势是每个期末平均增长的平滑估计。Holt-Winters 包含季节加法模型和季节乘法模型。其基本思想是对同时具有趋势效应和季节效应的序列分别进行指数平滑, 目的是分解确定性信息, 然后结合每种形式的平滑结果, 对原时间序列进行预测。“Holt-Winters 法”有时特指三次指数平滑法, 可以处理趋势与季节因素。其常见的两种数学模型如下:

Holt-Winters 季节加法模型:

$$\begin{aligned} S_t &= \alpha(x_t - p_{t-k}) + (1-\alpha)(S_{t-1} + T_{t-1}), \\ T_t &= \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\ p_t &= \gamma(x_t - S_t) + (1-\gamma)p_{t-k} \end{aligned}$$

Holt-Winters 季节乘法模型:

$$\begin{aligned} S_t &= \alpha \frac{x_t}{p_{t-k}} + (1-\alpha)(S_{t-1} + T_{t-1}), \\ T_t &= \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \\ p_t &= \gamma \frac{x_t}{S_t} + (1-\gamma)p_{t-k} \end{aligned}$$

上述公式中,  $0 < \alpha, \beta, \gamma < 1$ , 是平滑参数。 $\alpha$  控制水平估计的平均长度,  $\beta$  控制趋势的平滑,  $\gamma$  控制季节的平滑。 $S_t$ 、 $T_t$ 、 $p_t$ 在这个模型中分别表示水平性部分、趋势性部分和周期性 (季节性) 部分。

## 2.4 Loess 加法模型及混合模型

Loess 加法模型与混合模型是常用的季节性综合分析模型, 其中对趋势的拟合均采用了局部加权回归 Loess (locally weighted regression)。

Loess 加法模型  $x_t = T_t + S_t + I_t$  的构造分为两步:

Step 1: 计算并去除原始序列季节影响 $\hat{S}_t$ , 得到 $x_t - \hat{S}_t$ ;

Step 2: Loess 多元回归拟合趋势 $T_t$ , 得到平滑趋势拟合序列 $\hat{T}_t$ 。

同 Loess 加法模型类似, Loess 混合模型  $x_t = S_t(T_t + I_t)$  的构造也分为两步:

Step 1: 计算并去除原始序列季节影响 $\hat{S}_t$ , 得到 $\frac{x_t}{\hat{S}_t}$ ;

Step 2: Loess 多元回归拟合趋势 $T_t$ , 得到平滑趋势拟合序列 $\hat{T}_t$ 。

对模型的拟合检验主要为对拟合残差的检验, 其中加法模型拟合残差为 $\hat{I}_t = x_t - \hat{S}_t -$



$\hat{T}_t$ ，混合模型拟合残差为  $\hat{I}_t = \frac{x_t}{\hat{S}_t} - \hat{T}_t$ 。

## 2.5 简单季节模型

平稳时间序列可以分为严平稳时间序列和宽平稳时间序列，而一般应用的平稳时间序列指的是宽平稳时间序列。在对平稳时间序列进行简单季节分析预测时，通常采用的模型为 AR 模型（自回归模型）、MA 模型（滑动平均模型）、ARMA 模型（自回归移动平均模型）和 ARIMA 模型

### 2.5.1 AR 及 MA 模型

AR(p)模型，即为自回归模型，被广泛应用于经济学、信息学、自然现象的分析和预测。当一个系统在时刻  $t$  的响应变量  $x_t$  与其之前时刻的响应变量  $x_{t-1}, x_{t-2}, \dots$  以及  $t$  时刻的随机扰动  $\varepsilon_t$  有关，且与之前的  $\varepsilon_{t-j}, j = 1, 2, \dots$  没有直接的关系，这类系统被称作自回归模型(Auto Regresion Model)，简记为 AR 模型。

AR(p)具有如下结构：

$$\begin{aligned} x_t &= \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t \\ \phi_p &\neq 0 \\ E(\varepsilon_t) &= 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) &= 0, \forall s < t \end{aligned}$$

其中， $\phi_0, \phi_1, \dots, \phi_p$  为未知参数。特别的当  $\phi_0 = 0$  时，则称为中心化 AR(p)模型。

引入延迟算子，中心化的 AR(p)模型又可以记为  $\Phi(B)x_t = \varepsilon_t$ ,

自回归系数多项式  $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ 。

若系统在  $t$  时刻的响应变量  $x_t$  与它之前时刻的响应变量  $x_{t-1}, x_{t-2} \dots$  没有直接关系，而和之前时刻的随机扰动  $\varepsilon_{t-1}, \varepsilon_{t-2} \dots$  存在一定的关系，那么可以使用移动平均模型(Moving Average, MA)来拟合。在金融模型模型中，MA 经常用来刻画冲击效应，例如预期外的时间。

MA(q)具有如下结构：

$$\begin{aligned} x_t &= \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \\ \theta_q &\neq 0 \\ E(\varepsilon_t) &= 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \end{aligned}$$

特别的，中心化 MA(q)可以记为  $x_t = \Theta(B)\varepsilon_t$ ，当  $\mu = 0$  时。其中， $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$ 。 $\theta_1, \theta_2, \dots, \theta_q$  被称作移动系数。

### 2.5.2 ARMA 模型

自回归移动平均模型（Auto-Regression Moving Average Model, ARMA）是常用的一类随机时序模型，在 1970 年由美国统计学家伯克斯和詹金斯首次提出，并随后广泛地应用于各类时间序列的分析。ARMA 模型仅适用于平稳时间序列，对于非平稳序列可以通过差分使其平稳化。

若系统在  $t$  时刻的响应变量  $x_t$  不仅与它之前时刻的响应变量  $x_{t-1}, x_{t-2} \dots$  有关系，还与之前时刻的随机扰动  $\varepsilon_{t-1}, \varepsilon_{t-2} \dots$  存在一定的关系，可用自回归移动平均（ARMA）模型拟合。

ARMA(p,q)具有如下结构：

$$\begin{aligned} x_t &= \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \dots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \\ \phi_p &\neq 0, \theta_q \neq 0 \\ E(\varepsilon_t) &= 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) &= 0, \forall s < t \end{aligned}$$

当  $\phi_0 = 0$  时，则称为中心化 ARMA(p,q)模型。简记为  $\Phi(B)x_t = \Theta(B)\varepsilon_t$ ,  $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ ,  $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$

### 2.5.3 ARIMA 模型

当时间序列经过  $d$  阶差分后拟合得到自回归移动平均模型，那么原序列对应拟合的这类模型称为自回归积分滑动平均模型(Auto-Regressive Intergrated Moving Average Model, ARIMA)。对于  $d$  阶齐次非平稳序列  $x_t$ ,  $\nabla^d x_t$  是一个平稳序列，可以应用 ARMA(p,q)模型，所以可以推出 ARIMA 模型如下：

$$\begin{aligned} \Phi(B)\nabla^d x_t &= \Theta(B)\varepsilon_t \\ E(\varepsilon_t) &= 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) &= 0, \forall s < t \end{aligned}$$

## 2.6 复杂季节模型

当序列本身具有短期相关性时，往往可以用 ARMA(p,q)模型来拟合；季节相关性可以使用周期步长为单位的 ARMA(P,Q)拟合。当序列的季节效应，长期趋势效应和随机波动之间存在复杂关系时，就不能简单的用可加性拟合，两者往往是具有乘积关系的。因此模型拟合的实质是 ARMA(p,q)与 ARMA(P,Q)的乘积。再结合差分运算，通过低阶差分(d)，充分提取趋势信息，而季节信息则以周期步长  $S$  的  $D$  阶简单季节差分提取充分。因此 ARIMA 乘积季节模型可以简记为 SARIMA(p,d,q)×(P,D,Q)<sub>s</sub>，式中  $p$  和  $q$  为自回归和移动平均阶数， $P$  和  $Q$  分别为季节性自回归和移动平均阶数， $d$  代表差分次数， $D$  是季节性差分次数， $s$  为季节周期和循环长度。此类模型已经被用于多项技术和科学研究。结构如下：

$$\Phi(B)\Phi_S(B)\nabla^d \nabla_S^D x_t = \Theta(B)\Theta_S(B)\varepsilon_t$$

其中：

$\nabla = 1 - B$  是正常差异，而  $\nabla_S$  代表季节差异， $\varepsilon_t$  是白噪声。

$\Theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ ,  $\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ ，分别是  $q$  阶移动平均多项式和  $p$  阶自回归多项式，

$\Theta_S(B) = 1 - \theta_1 B^S - \dots - \theta_Q B^{QS}$ ,  $\Phi_S(B) = 1 - \phi_1 B^S - \dots - \phi_P B^{PS}$ 。其中  $\Theta_S(B)$  是

Q 阶季节移动平均多项式， $\Phi_S(B)$  是 P 阶季节自回归多项式。

## 2.7 GARCH 模型

ARCH(Auto-Regressive Conditonal Heteroskedasticity)模型全称为自回归条件异方差模型。在许多应用中，由于自相关系数较高并且衰减速度较慢，线性 ARCH(q)模型需要很长的滞后长度 q 来刻画。常见的 ARCH(q)的模型结构有以下两种：

ARCH(q)模型结构 1:

$$\begin{cases} \varepsilon_t = \sqrt{h_t} e_t \\ h_t = \omega + \sum_{j=1}^q \lambda_j \varepsilon_{t-j}^2 \\ e_t \sim WN(0,1) \end{cases}$$

ARCH(q)模型结构 2:

$$\begin{aligned} a_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_q a_{t-q}^2 \\ \varepsilon_t &\sim WN(0,1) \\ \alpha_0 &> 0, \text{ 对 } i > 0, \alpha_i > 0 \end{aligned}$$

另一种更灵活的滞后结构是由 Bollerslev (1986) 引入的广义 ARCH (GARCH)。他证明了带有很小的滞后项的 GARCH (1,1) 模型也可以很好地模拟长期样本期间的变化 (French et al, 1987; Franses and Van Dijk, 1996)。GARCH 模型的一般表达式是含有 p 个 GARCH 项和 q 个 ARCH 项的和，即 GARCH (p,q)。常见的 GARCH(p,q)的模型结构有以下两种：

GARCH(p,q)模型结构 1:

$$\begin{cases} \varepsilon_t = \sqrt{h_t} e_t \\ h_t = \omega + \sum_{i=1}^p \eta_i h_{t-i} + \sum_{j=1}^q \lambda_j \varepsilon_{t-j}^2 \\ e_t \sim WN(0,1) \end{cases}$$

GARCH(p,q)模型结构 2:

$$\begin{aligned} a_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^q \alpha_i a_{t-i}^2 + \sum_{i=1}^p \beta_i \sigma_{t-i}^2 \\ \varepsilon_t &\sim WN(0,1) \\ \alpha_0 &> 0, \alpha_i \geq 0, \beta_j \geq 0, \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1 \end{aligned}$$

## 2.8 FFNN 模型

前馈神经网络 (Feedforward Neural Networks, FFNN) 是一种最简单的神经网络模型，多个神经元分层排列，且每个神经元只与前一层的神经元相连，接收前一层的输出，并输出给下一层，各层间没有反馈。因此，在前馈神经网络中，信息以单一方向流动，即从输入层，经过隐含层 ( 可为一个或多个 )，再流向输出层。作为一种静态非线性映射，FFNN 能够以任意精度逼近任意连续函数及平方可积函数，通过简单非线性处理单元的复合映射，可获得复杂的非线性处理能力。其中，多层神经网络 ( Multilayer Feedforward Neural Networks, MFNN) ( 含有一个输入层、一个输出层以及多个隐含层 ) 是目前神经网络应用最广泛、发展最迅速的神经网络模型之一。

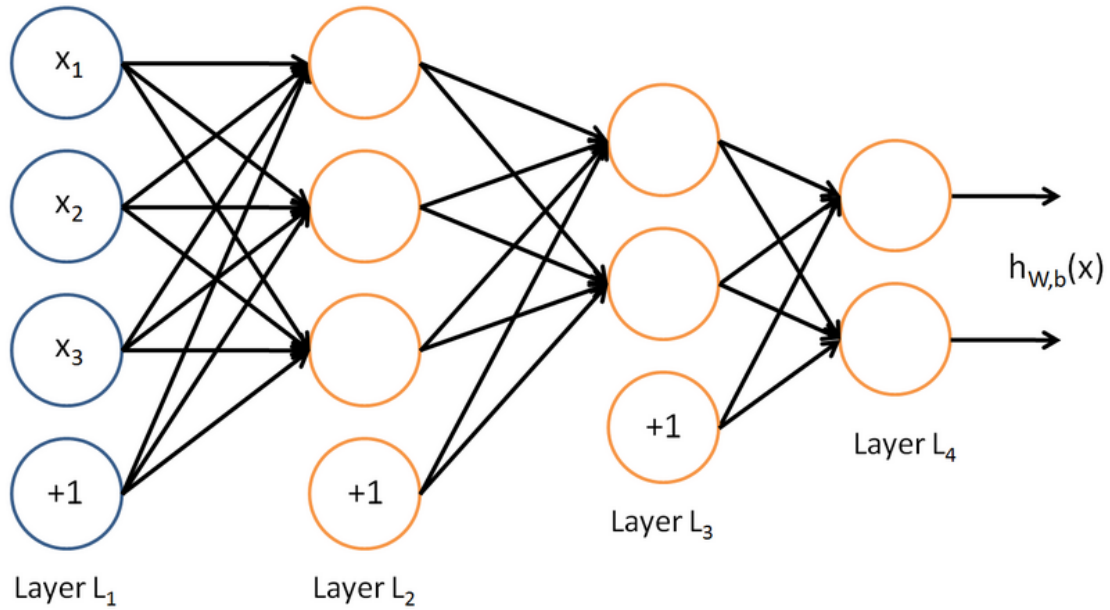


图 1 多层前馈神经网络示意图

如图 1 所示的 MFNN 的输入层 (Layer  $L_1$ ) 具有三个输入节点 ( $x_1$ 、 $x_2$ 、 $x_3$ )，第一隐含层 (Layer  $L_2$ ) 具有三个隐含节点，第二隐含层 (Layer  $L_3$ ) 具有两个隐含节点，输出层 (Layer  $L_4$ ) 具有两个输出节点。信息自  $L_1$ ，以单一方向经由  $L_2$ 、 $L_3$ ，最终流向  $L_4$ 。赋予该网络初始权值和阈值后，按照给定的样本前向计算网络输出，同时基于实际输出与期望输出之间的误差反向修改网络权值和阈值，通过反复训练得到最终的网络拟合模型（即使得实际输出与期望输出的误差达到最小）。

FFNN 应用的数学基础是 Kolmogorov 定理：给定任意连续映射  $H: E_m$  ( $m$  维单位立方体)  $\rightarrow R_n$ ， $H$  可以精确地由一个三层神经网络实现，此神经网络第一层有  $m$  个神经元，中间层有  $(2m+1)$  个神经元，输出层有  $n$  个神经元。由定理可知，在给定变量输入输出关系的前提下，如  $Y = f(x_1, x_2, \dots, x_m)$ ，其中  $Y = (y_1, y_2, \dots, y_n)$  为输出向量， $X = (x_1, x_2, \dots, x_m)$  为输入向量，则必存在一个  $m$ —— $(2m+1)$ —— $n$  的前馈神经网络  $N$ ，使得  $Y = N(x_1, x_2, \dots, x_m)$ ，即用一个前馈神经网络可以模拟给定变量的输入输出关系，得到相应的前馈神经网络模型。

### 3 数据来源及特征描述

#### 3.1 数据来源及数据补全

本文研究的数据为 1997 到 2016 年广州市、上海市及哈尔滨市月度降水量、温度、湿度时间序列数据，该数据是从国泰安数据库获取的。其中，广州市 2008 年 12 月降雨量缺失，哈尔滨 2004 年 2 月及 2008 年 1、2 月降雨量缺失，而上海市降雨量数据完整。对于数据中的缺失值，用缺失值前一年同月及后一年同月两者的平均值填入。

对降雨量进行测量一直以来是记录水资源的重要基础资料之一，同时也是研究流域

和地区水文循环的重要指标，对人类生产生活有重要意义。而通过历史数据对降雨量进行预测有助于对自然灾害的预防。如图 2 本文从南至北选取三个城市，其降雨量依次递减，研究其降水量的模型拟合及预测。

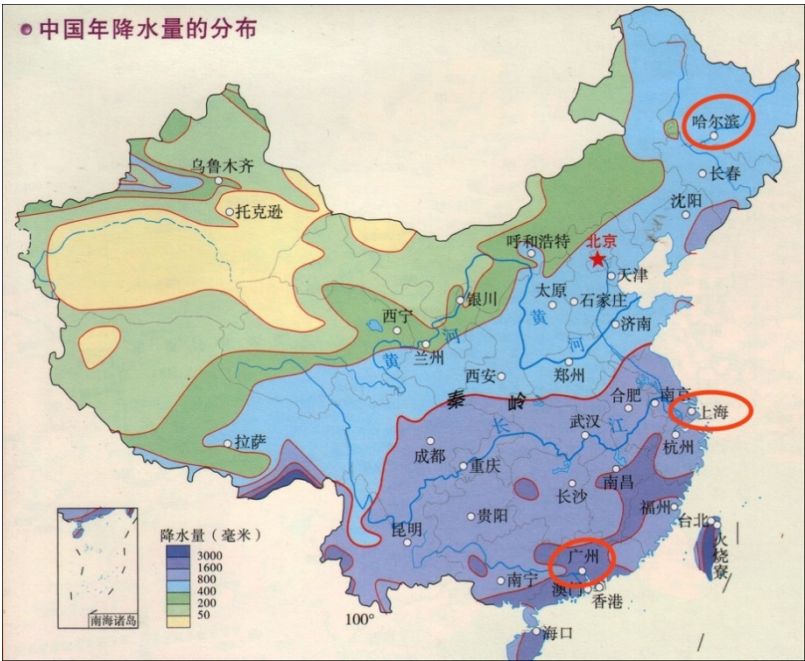


图 2 中国年降雨量分布图

3.2 描述性统计

	Min	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max
广州	0.10	42.10	116.60	164.90	260.10	834.60
上海	1.50	44.52	81.55	101.77	121.38	613.10
哈尔滨	0.40	6.99	21.25	42.79	63.13	249.50

表 1 降雨量描述性统计

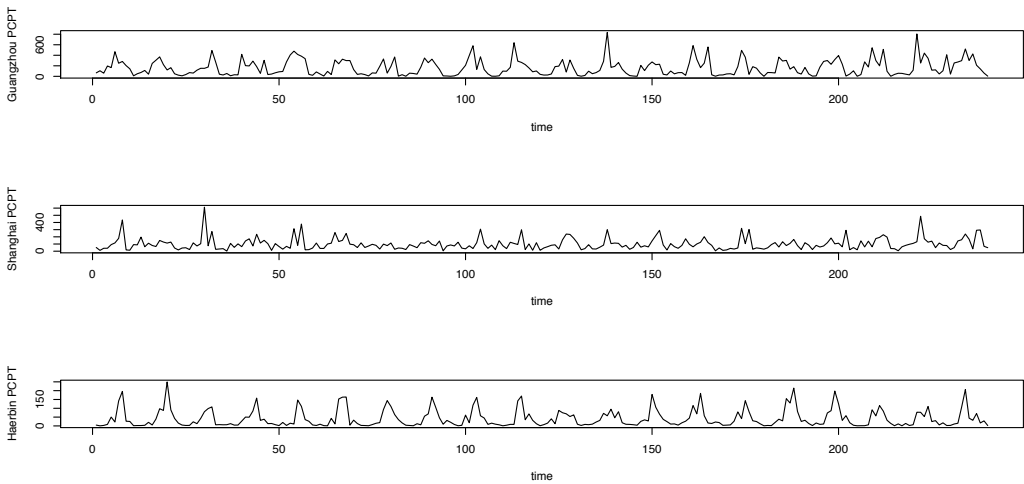


图 3 三个城市降雨量时序图

从图 3 可以发现三个城市的降雨量都具有季节性，没有明显上升或者下降趋势，同时从表 1 可以看出三个城市的降雨量从南至北依次降低，符合正常认知。

### 3.3 平稳性分析及长记忆性分析

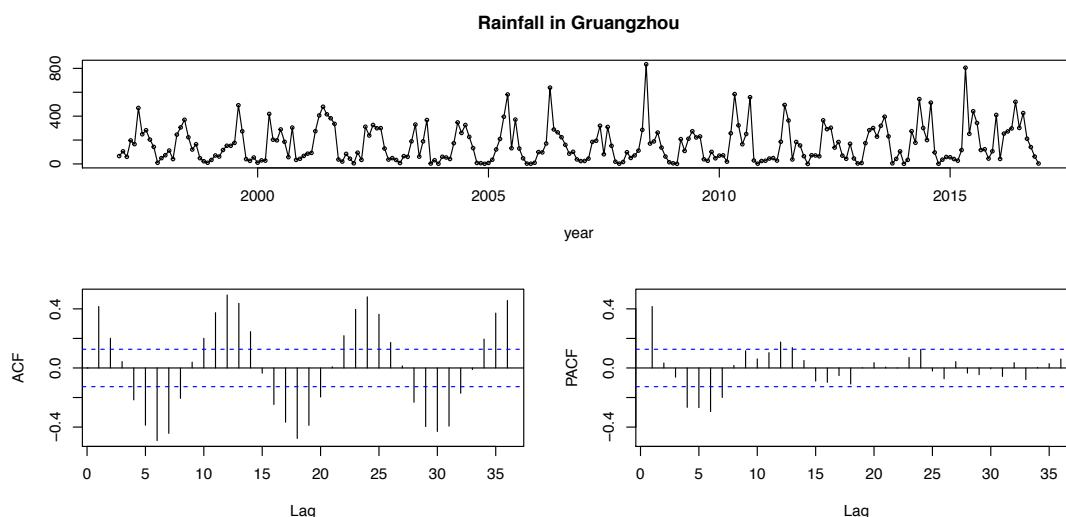


图 4 广州市时序图、自相关图及偏自相关图

为了检测可能存在的季节性，趋势和其他非线性现象，本文将观察数据的时序图与样本自相关函数（ACF）和样本偏自相关函数（PACF）的图进行对比。这将帮助我们确定差分阶数以及对数变换的必要性。非平稳性表现为 ACF 值  $\rho_k$ ，和 PACF 值  $\phi_{kk}$  快速消失。并且季节性差异由大的 ACF 值出现在  $\text{Lag}=s, 2s \dots ns$  处体现。

为了确定趋势和异方差是否存在，本文观察了降雨量时序图，并且与 ACF、PACF 一同比较。通过对广州市降雨量的时序图分析，猜测降雨量平稳且存在着季节性，而 ACF 图呈现的间隔 12 个月的周期性变化证明了序列具有季节性，PACF 图中， $\phi_{kk}$  快速地衰减为 0，表明了序列的平稳性。

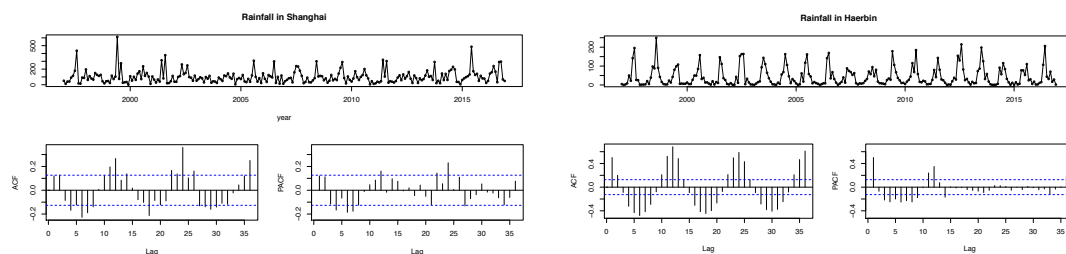


图 5 上海市、哈尔滨市时序图、自相关图及偏自相关图

同理，上海市、哈尔滨市的降雨量同样具有季节性，且序列平稳。

城市	单位根检验		平稳性检验
	adf.test p-value	pp.test p-value	kpss.test p-value
广州	0.01	0.01	0.1
上海	0.01	0.01	0.1
哈尔滨	0.01	0.01	0.1

表 2 平稳性检验

单位根检验是在原假设 $\beta = 1$ 和备择假设 $\beta < 1$ 的情况下进行的，计算出检验统计量为  $ADF_{\beta} = \frac{\hat{\beta}-1}{SE(\hat{\beta})}$ ，将它与 Dickey-Fuller 测试的相关临界值进行比较，若检验统计量小于临界值，则拒绝原假设即没有单位根存在。同时计算出 p 值，当  $p < 0.05$  时，拒绝原假设说明没有单位根的存在。

分别对三个城市降雨量进行单位根检验，来检验序列的平稳性。当时间序列含有单位根时，序列为非平稳时间序列。通过 R 软件对三个城市的单位根进行检验，发现 p 值均小于 0.05，拒绝原假设，不存在单位根故序列平稳。再对城市进行 Phillips-Perron 单位根检验和 KPSS 平稳性检验均证明序列平稳。

图 6 显示三个城市的谱密度均未在 0 处趋近于无穷，故三个城市无需采用长记忆时间序列模型。

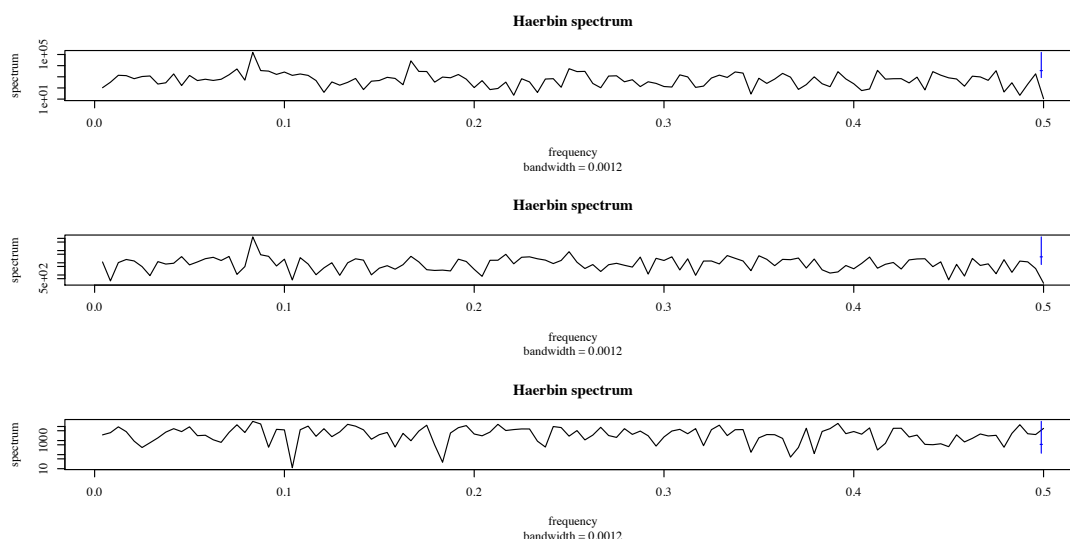


图 6 三个城市谱图

### 3.4 白噪声检验

本文采用 Ljung-Box 方法对序列进行白噪声检验。通过 Ljung-Box 检验在时间段内的一系列观测值是否为纯随机的独立观测值。如果观测值不是相互独立，一个观测值会在 m 时间后与另外一个观测值相关，则会形成一种称为自相关的关系。自相关会降低



基于时间的预测模型（例如时间序列模型）的准确性，也会得出错误的数据解释。

城市	白噪声检验(Box.test)					
	Lag=1		Lag=6		Lag=12	
	$\chi^2$ 值	P 值	$\chi^2$ 值	P 值	$\chi^2$ 值	P 值
广州	156.82	<0.0001	156.82	<0.0001	316.17	<0.0001
上海	3.4967	0.06149	31.999	<0.0001	75.233	<0.0001
哈尔滨	60.389	<0.0001	196.95	<0.0001	448.78	<0.0001

表 3 白噪声检验

对三个城市分别进行 Lag=1,6,12 的白噪声检验，表 3 中 p 值显示，系数的显著水平为 0，因此时间序列存在自相关性。

通过对上述的平稳性检验、长记忆检验，研究的三个城市降雨量均为平稳非白噪声序列，且应使用短记忆模型对序列拟合。

4 模型建立

4.1 季节效应分析

4.1.1 趋势分析

由于数据波动比较平稳，且明显具有季节性。用具有季节性的 Holt-Winters 加法模型来对三个城市进行拟合，拟合的平滑参数如下表 4。

	$\alpha$	$\beta$	$\gamma$
哈尔滨市	0.0639	0.0414	0.3188
广州市	0.0399	0.0355	0.0688
哈尔滨市	0.0115	0.1326	0.2604

表 4 Holt-Winters 加法模型系数

从平滑参数值可以看出，水平参数 $\alpha$ 接近于 0。因为  $\alpha$  越接近 1，平滑后的值越接近当前时间的数据值，数据越不平滑， $\alpha$  越接近 0，平滑后的值越接近前 i 个数据的平滑值，数据越平滑。因此该 Holt-Winters 的平滑性很强。趋势参数 $\beta$ 的值接近于 0，说明不具有明显的趋势性。而季节参数 $\gamma$ 对应的值相对于前两个参数是比较大的，也可以说明数据具有比较强的季节性但无很强的趋势性。

以哈尔滨月均降雨量为例，采用加法模型的带季节性的 Holt-Winters 模型，用 R 统计软件作出对应的拟合图（图 7）。



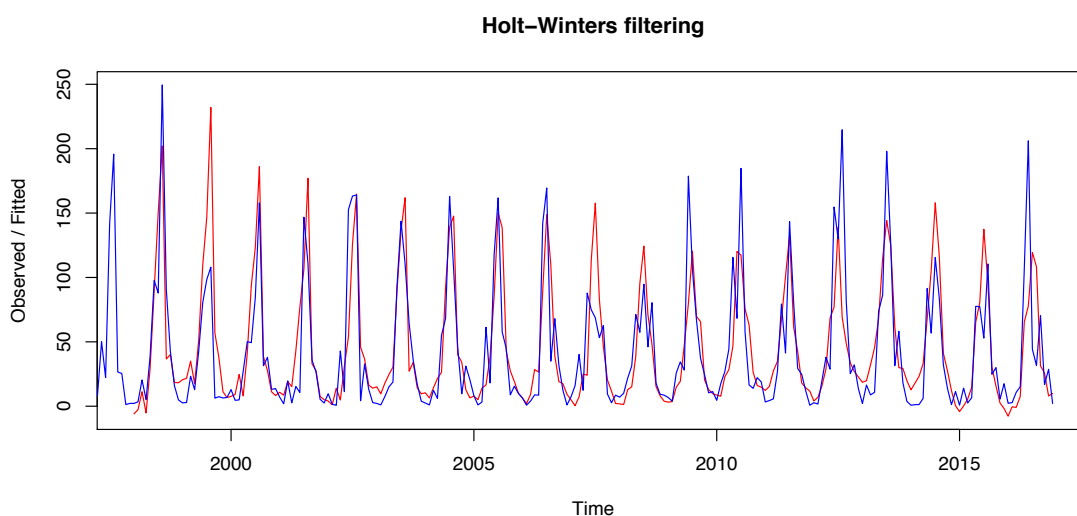


图 7 Holt-Winters 拟合图

从图 7 可看出，拟合效果并不是非常理想，同理得到广州市和上海市的 Holt-Winters 拟合图，拟合结果同样不是很理想。

#### 4. 1. 2 信号分解

##### 4. 1. 2. 1 Decompose 分解

由于数据波动比较平稳，采用加法模型的 Decompose 处理哈尔滨月均降雨量数据，用 R 统计软件作出 Decompose 的拟合图如下：

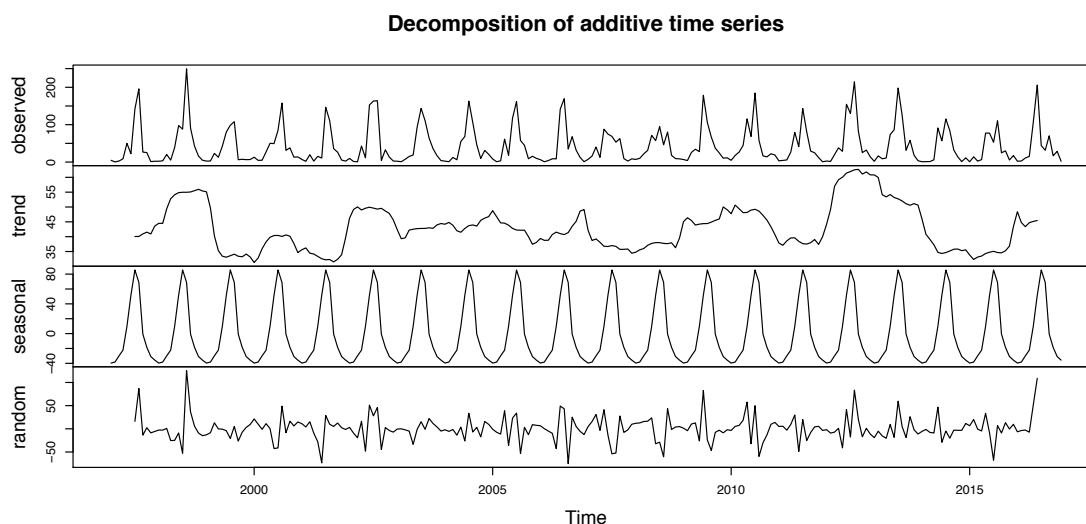


图 8 Decompose 分解图

从图中可以看出，数据具有明显的季节性，但是不具有趋势性，因此后续处理数据本文进行了 12 步差分（图 9）及 Decompose 法（图 10）去除季节性。

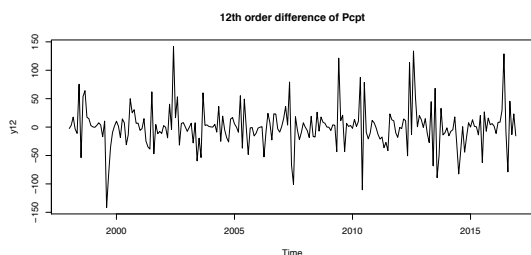


图9 12步差分结果图

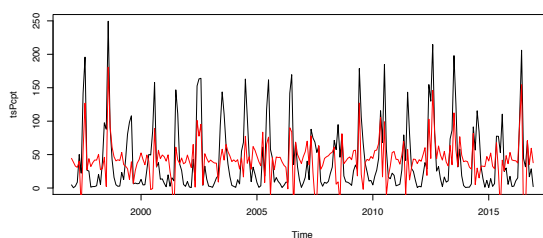


图10 Decompose 去除季节性

两种结果都不再具有明显的周期性，故数据处理是合理的且去除了部分季节性。同理得到广州市和上海市也不具有明显的趋势性，但是具有季节性，因此对数据可进行相同的处理。

#### 4.1.2.2 STL 分解

在 R 语言环境下，应用 `stl` 函数对原始时间序列进行 STL 分解，结果如图 11 所示。

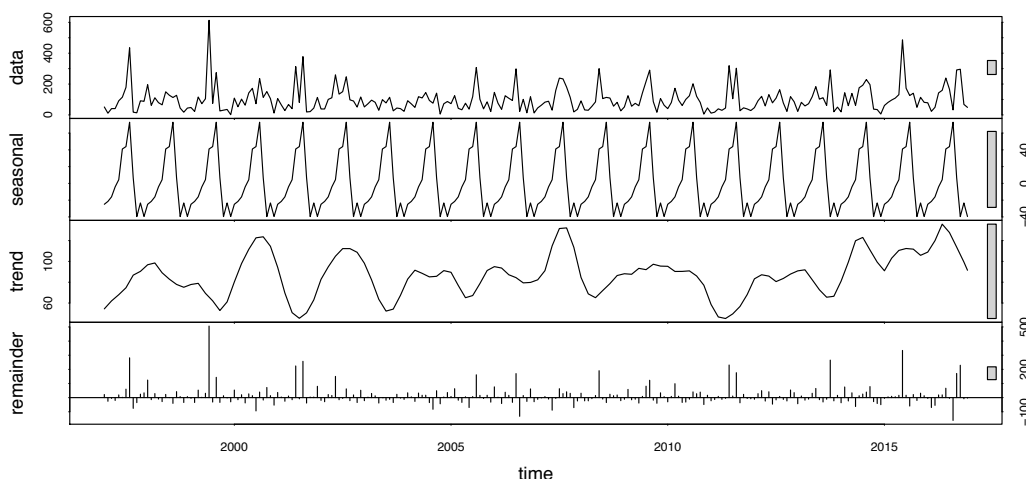


图11 STL 分解图

## 4.2 Loess 加法模型及混合模型

在 R 语言环境下，应用 `decompose` 函数得到 Loess 加法模型中的季节影响  $\hat{S}_t$ ，原始序列去除季节影响后，应用 `loess` 函数对趋势进行 Loess 回归，并将相应结果可视化(图 12)。Loess 混合模型的构建中，引入季节指数  $\hat{S}_t$ ，通过对原始序列( $x_t$ )除以  $\hat{S}_t$  从而去除季节影响。对序列  $\frac{x_t}{\hat{S}_t}$  进行 Loess 回归，得到对趋势的拟合结果(图 13)。

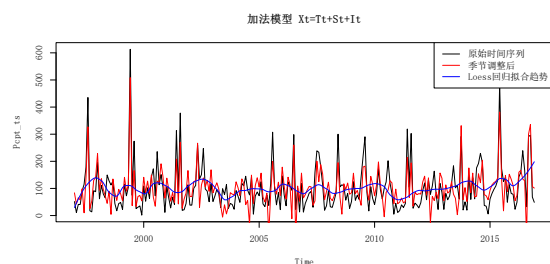


图12 Loess 加法模型拟合图

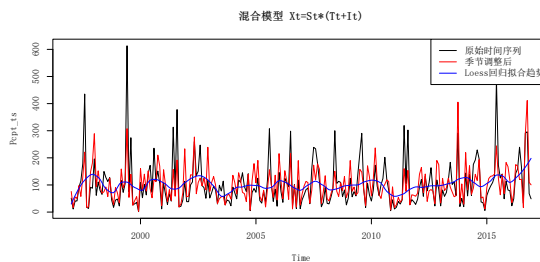


图13 Loess 乘法模型拟合图

对两个模型的拟合残差分别进行 Lag=12 的白噪声检验(加法模型,  $P=0.027$ ; 混合模型,  $P=0.0013$ ) 得出 Loess 加法模型以及混合模型对上海市月均降水量的拟合均不显著。

### 4.3 简单季节模型

由于月度降雨量有明显的季节性, 对于简单季节模型的选择, 使用 12 步差分后数据进行拟合。

#### 4.3.1 AR 及 MA 模型

在 R 语言环境下, 运用 `ar()` 函数根据广州市月度降雨量分别拟合出 AR(1)-AR(20) 模型, 可视化模型对应的 AIC 值, 并运用 AIC 信息准则 (赤池信息准则) 选择模型阶数。

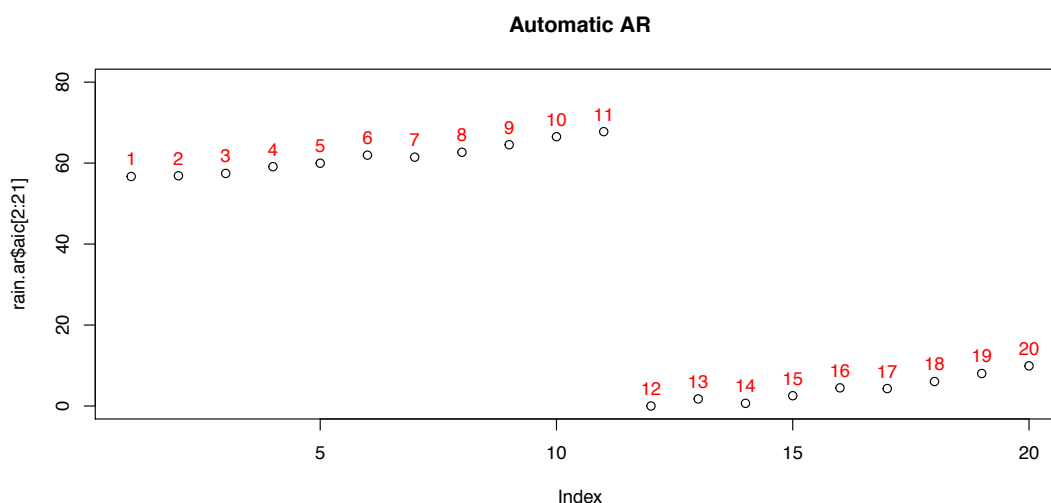


图 14 AR 模型定阶

AIC 是用来衡量统计模型拟合优良性的一种标准, 在 1974 年由日本统计学家赤池弘次提出。AIC 是建立在熵的概念上, 为权衡估计模型复杂度和拟合数据优良性提供了标准。通常情况下, AIC 定义为:  $AIC = 2k - 2\ln(L)$ 。其中  $k$  是模型参数个数,  $L$  是指似然函数。在几个可供选择的模型中, 通常选择 AIC 最小的模型做为最佳模型。通过图 14, 可以看出 AR 模型在阶数为 12 的时候 AIC 值最小, 因此选择 AR(12) 作为在 AR 模型的选择。

同理, 对于哈尔滨市的差分后的降雨量数据通过 AIC 最小信息准则选择 AR(2) 模型拟合。

对于广州市月度降雨量 MA 模型阶数的选择, 同样采用 AIC (最小信息准则) 来选择。通过下表各模型 AIC 值的比较, 选择 MA(6) 作为去除季节趋势的数据的 MA 模型拟合。因此, 对于原始降雨量数据, 选择 MA(6,12) 作为模型拟合。

模型	广州市 AIC	哈尔滨市 AIC	通过 AIC 最小的原则，对广州市降雨量选择模型 MA(6,12)，对哈尔滨市降雨量模型选择 MA(2,12)
MA(1,12)	2940.316	2290.641	
MA(2,12)	2940.819	2290.715	
MA(3,12)	2940.737	2287.222	
MA(4,12)	2941.971	2289.076	
MA(5,12)	2943.608	2291.073	
MA(6,12)	2938.627	2295.026	

表 5 MA 模型定阶

#### 4.3.2 ARMA 模型

同理，根据 AIC 最小原则选择 ARMA((6,12),(6,12))对广州市原始月度降雨量拟合。

MA AR	1	2	3	4	5	6
1	2942.112	2934.094	2941.887	2943.825	2937.544	2940.453
2	2941.653	2943.544	2923.759	2945.857	2921.759	2926.826
3	2942.535	2923.883	2917.051	2918.851	2922.846	2921.320
4	2944.454	2925.732	2919.246	2921.195	2923.573	2923.291
5	2945.628	2922.981	2921.080	2916.393	2918.373	2920.287
6	2939.210	2920.043	2922.641	2924.564	2920.111	2907.016

表 6 ARMA 模型定阶

对于哈尔滨市降雨量的模型通过 AIC 准则采用 ARMA((5,12),(3,12))。

#### 4.3.3 模型显著性检验

延迟阶数	拟合残差白噪声检验					
	AR(12,12)		MA(6,12)		ARMA((6,12),(6,12))	
	$\chi^2$ 值	P 值	$\chi^2$ 值	P 值	$\chi^2$ 值	P 值
6	0.97	0.9867	6.29	0.3916	6.81	0.3392
12	11.14	0.5169	58.09	5.01e-08	40.97	4.966e-05
结果	模型阶数过高		模型不显著		模型不显著	

表 7 白噪声检验

如上表 7，分别对广州市模型拟合的 AR(12,12)，MA(6,12)，ARMA((6,12),(6,12))进行残差白噪声检验。若残差为白噪声，证明模型对信息提取充分。结果显示仅有 AR(12,12)的残差是白噪声，但是由于 AR 模型是短记忆性的，12 阶模型的阶数选择过高，因此上述三种模型均不适用。

对于哈尔滨市降雨量的 AR(2,12)模型进行残差白噪声检验，发现在 Lag=6 时，p 值为 0.9928，而在 Lag=12 时，p 值为 0.0070，拒绝原假设。在 AR(2,12)模型下，残差非白噪声。而 MA(2,12)和 ARMA((5,12),(3,12))对于哈尔滨市的降雨量拟合，在 Lag=12 时，p 值分别为 0.0087 和 0.0351，两个模型的残差均非白噪声，模型均不显著。

#### 4.4 复杂季节模型

通过降雨量时序图与 ACF、PACF 图的综合比较，可以得出每 12 个月为一个季节周期，且季节长度没有超过 24 个单位，因此可以使用复杂季节模型进行拟合。模型中需要确定的参数有：差分次数 d，自回归阶数 p 和移动平均阶数 q，以及季节性自回归阶数 P、移动平均阶数 Q、季节性差分次数 D 和季节周期 s。这些参数很难通过 ACF、PACF 图的可视化方法识别。这些参数可以通过反复实验比较 AIC 来确定。

在 R 语言环境下，使用 auto.arima()函数来确定复杂季节模型的阶数。最终确定 SARIMA(0,0,0)(2,0,0)<sub>12</sub> 为广州市月度降雨量的拟合模型。

模型如下：

$$(1 + 0.3222B^{12} + 0.3801B^{24})X_t = \varepsilon_t$$

对模型拟合残差进行白噪声检验验证模型是否显著。通过 Ljung-Box 检验，显示在 Lag=6, 12 时 P>0.05，不拒绝原假设。由于在 Lag=12 时，残差为白噪声，可以明确得出在 Lag<12 时，残差仍为白噪声。再检验模型拟合的参数是否显著，若存在不显著参数则删掉不显著参数使得模型达到最精简。通过 t-ratio test 进行检验，发现两个参数均显著。相较于简单季节模型的拟合，复杂季节模型是相对最优模型。

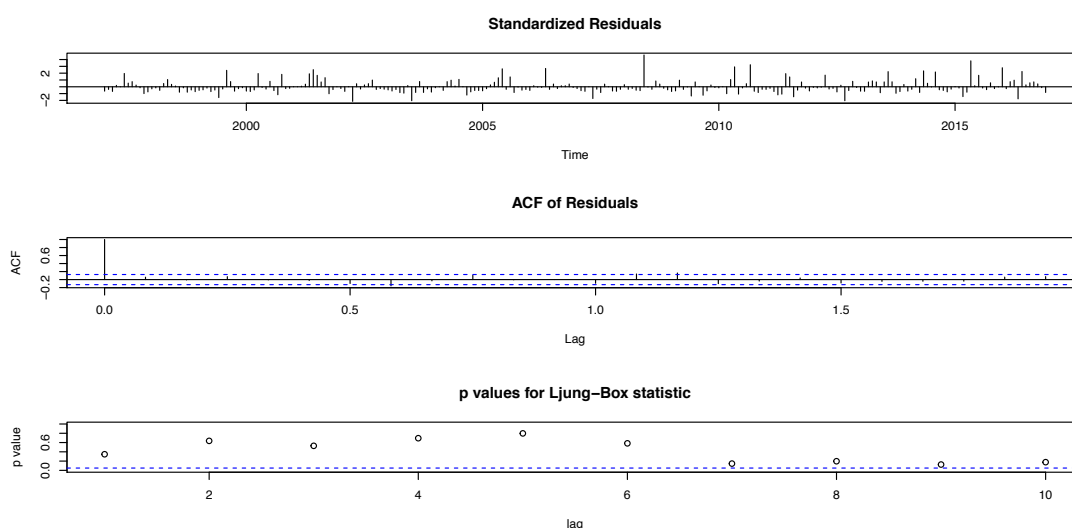


图 15 复杂季节模型的残差分析

残差白噪声检验			参数显著性检验		
延迟阶数	$\chi^2$ 值	P 值	待估参数	t 统计量	P 值
6	4.58	0.5987	$\phi_1(0.3222)$	5.42	0.0324
12	15.458	0.2173	$\phi_2(0.3801)$	6.02	0.0265
结果	模型显著		参数均显著		

表 8 广州市模型检验

为确定模型残差是否为高斯白噪声，对残差进行正态性检验。Normal Q-Q Plot 显示残差非正态。同时比较残差的分布图与正态分布图（图 16），残差尖峰，差异很大。故而残差非高斯白噪声。

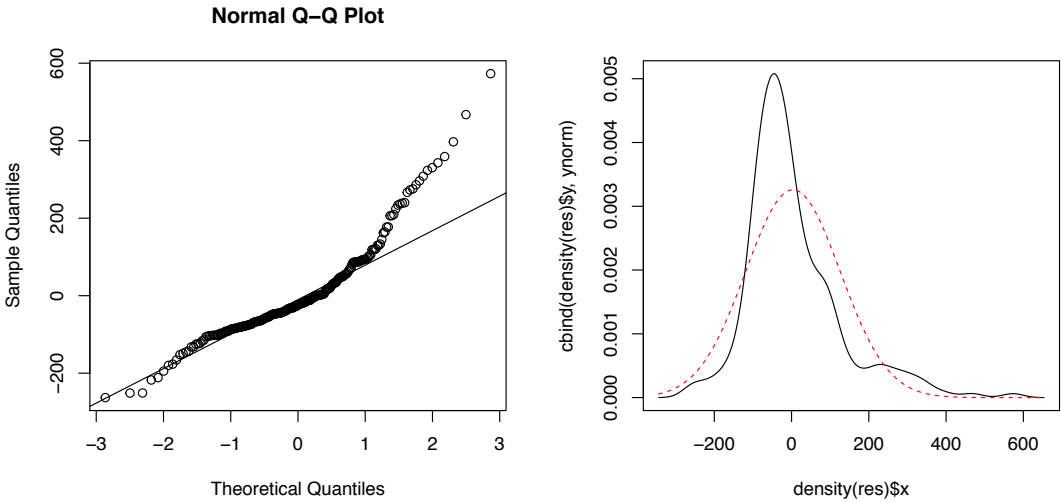


图 16 残差正态性检验

同样对于哈尔滨市的降雨量数据进行复杂季节模型拟合，拟合结果同样是 SARIMA(0,0,0)(2,0,0)<sub>12</sub>，且残差分布为非正态。

残差白噪声检验			参数显著性检验		
延迟阶数	$\chi^2$ 值	P 值	待估参数	t 统计量	p 值
6	4.49	0.6109	$\phi_1(0.5247)$	8.28	0.0143
12	11.02	0.5274	$\phi_2(0.3011)$	4.53	0.0454
结果	模型显著		参数均显著		

表 9 哈尔滨市模型检验

对模型进行检验，发现模型显著，参数显著。模型如下：

$$(1 + 0.5247B^{12} + 0.3011B^{24})X_t = \varepsilon_t$$

同理，对于上海市的降雨量数据进行复杂季节模型拟合，结果为 $(1 - B^{12})X_t =$

$(1 - 0.8774B^{12})\varepsilon_t$ 。进行模型检验后，证明模型、参数均显著。

但由于残差图（图 17）显示了一个最大异常值（1998 年 6 月），本文引入一个哑变量  $X_{dummy}$  对该异常值进行处理。哑变量长度与原始上海降雨量数据长度相同，正常值对应赋值为 0，最大异常值对应赋值为 1。将该哑变量引入  $SARIMA(0,0,0)(0,1,1)_{12}$ ，重新拟合后得到的新复杂季节模型（哑变量对应参数 449.7）与原模型阶数相同，残差图如图 18 所示，并且 AIC 值降低（原模型：2627.166；新模型：2587.215）。

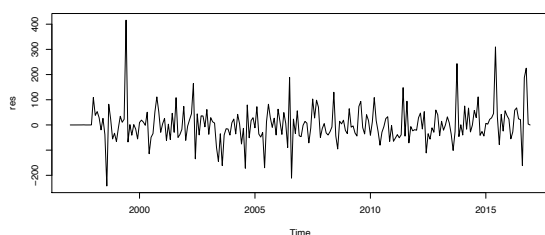


图 17 原 SARIMA 残差图

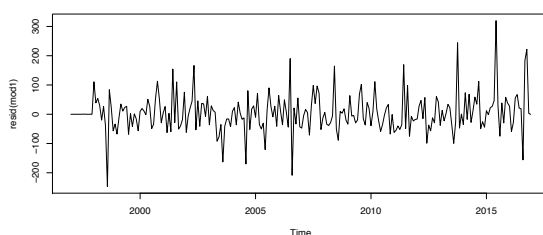


图 18 新 SARIMA 残差图

经检验，新 SARIMA 模型与参数均显著。

## 4.5 GARCH 模型

### 4.5.1 ARCH 效应检验

根据上述拟合效果较好的季节模型，即通过残差白噪声检验及参数检验的复杂季节模型，进一步检测它们的残差是否具有 ARCH 效应。检验 ARCH 效应的方法有残差平方的 Ljung-Box 检验，ARCH-LM 检验等。本文采用第一种方法，即检验模型残差的平方是否具有相关性。在后续 GARCH 模型定阶中，本文引入 ARCH-LM 检验来判断对应 GARCH 模型是否消除了自回归条件异方差的影响。

以哈尔滨市的月度降雨量数据为例，根据  $SARIMA(0,0,0)(2,0,0)_{12}$  模型残差平方的 LB 检验，如下表 10 所示：

延迟阶数	LB 统计量	P 值	检验结论
6	17.88	0.0066	具有 ARCH 效应
12	34.41	0.0006	

表 10 残差平方白噪声序列检验

从表数据可看出，p 值在 Lag=6 和 12 的情况下均小于 0.05，拒绝原假设，即复杂季节模型的残差平方是具有相关性的。同理对广州市及上海市复杂季节模型的残差进行相同的 ARCH 效应检验。由于这两个城市的对应 p 值均大于 0.05，表明模型残差均无 ARCH 效应，因此仅对哈尔滨市复杂季节模型的残差进一步拟合 GARCH 模型。

### 4.5.2 GARCH(p, q) 定阶

由于研究的数据量有限，且均值方程的拟合模型（复杂季节模型）较为复杂，因此对均值方程对应残差的 GARCH 模型，这里只考虑最基本的 GARCH (p,q) 模型。相应

的推广模型如 EGARCH, TGARCH, GARCH-M 等不在考虑范围之内。GARCH 阶数选择本文主要根据以下三个原则:(1)模型参数显著, 且满足无条件方差的收敛性, 即模型相应系数项和满足不大于  $1 (\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1)$ ; (2)模型拟合后的残差不再具有 ARCH 效应, 即模型去除了 ARCH 效应; (3)模型拟合的阶数合理。

#### 4.5.2.1 ARCH(1) 模型

首先拟合 ARCH(1), 即 GARCH(0,1)。拟合模型为:

$$\begin{cases} \varepsilon_t = \sqrt{h_t} e_t \\ h_t = \omega + \alpha_1 \varepsilon_{t-1}^2 \\ e_t \sim WN(0,1) \end{cases}$$

用 R 统计软件运行的检验结果如下表所示。

	参数估计	标准差	t 值	p 值
$\omega$	601.18	93.7304	6.41	<0.0001
$\alpha_1$	0.74	0.2505	2.96	0.0031

表 11 参数估计及检验

t 检验的 p 值均小于 0.05, 所以参数显著。同时参数小于 1 所以满足无条件方差收敛。进一步对模型进行检验, 对 ARCH(1)模型标准化的残差进行 LB 检验, ARCH-LM 检验, Jarque-Bera 和 Shapiro-Wilk 正态性检验, 残差的平方进行 LB 检验。R 软件运行的检验结果如下表 12 所示。

	残差	统计量名	统计量值	P 值
Jarque-Bera 检验	R	$\chi^2$	627.47	<0.0001
Shapiro-Wilk 检验	R	W	0.85	<0.0001
Ljung-Box 检验	R	Q(10)	6.37	0.7833
Ljung-Box 检验	R	Q(15)	24.98	0.0502
Ljung-Box 检验	R	Q(20)	28.21	0.1045
Ljung-Box 检验	$R^2$	Q(10)	8.50	0.5804
Ljung-Box 检验	$R^2$	Q(15)	22.34	0.0991
Ljung-Box 检验	$R^2$	Q(20)	27.26	0.1281
ARCH-LM 检验	R	$TR^2$	18.14	0.1114

表 12 残差平方项白噪声检验

从表中可以看出, LB 检验的 p 值均大于 0.05, ARCH-LM 检验的 p 值也大于 0.05,



均接受原假设，即 ARCH（1）模型拟合后的残差是无 ARCH 效应的，模型是显著的。同时 Jarque-Bera 检验和 Shapiro-Wilk 检验的 p 值均显著小于 0.05，得到拟合后的残差不具有正态性，即非高斯白噪声。

ARCH(1)模型建立如下。 拟合效果较好。接下来尝试 GARCH(p,q)模型。

$$\begin{cases} \varepsilon_t = \sqrt{h_t}e_t \\ h_t = 601.18 + 0.74\varepsilon_{t-1}^2 \\ e_t \sim WN(0,1) \end{cases}$$

4. 5. 2. 2 GARCH (p, q) 模型

我们以 GARCH(1,1)模型为例，代表性地分析 GARCH 模型。拟合模型为：

$$\begin{cases} \varepsilon_t = \sqrt{h_t}e_t \\ h_t = \omega + \alpha_1\varepsilon_{t-1}^2 + \beta_1h_{t-1}^2 \\ e_t \sim WN(0,1) \end{cases}$$

用 R 统计软件运行的检验结果如下表 13 所示。

	参数估计	标准差	t 值	p 值
$\omega$	349.667	61.9956	5.64	<0.0001
$\alpha_1$	0.94	0.2487	3.78	0.0002
$\beta_1$	0.12	0.0470	2.61	0.0090

表 13 参数估计及检验

t 检验的 p 值均小于 0.05，所以参数显著。然而因为参数和 $\alpha_1 + \beta_1 > 1$ ，所以不满足无条件方差收敛。进一步同 ARCH（1）模型一样进行下列检验。R 软件运行的检验结果如下表所示。

	残差	统计量名	统计量值	P 值
Jarque-Bera 检验	R	$\chi^2$	279.17	<0.0001
Shapiro-Wil 检验	R	W	0.87	<0.0001
Ljung-Box 检验	R	Q(10)	6.19	0.7994
Ljung-Box 检验	R	Q(15)	27.14	0.0276
Ljung-Box 检验	R	Q(20)	29.67	0.0754
Ljung-Box 检验	$R^2$	Q(10)	11.74	0.3029
Ljung-Box 检验	$R^2$	Q(15)	34.50	0.0029
Ljung-Box 检验	$R^2$	Q(20)	42.41	0.0024
ARCH-LM 检验	R	$TR^2$	27.31	0.0070

表 14 标准化残差检验

从表 14 中可以看出，LB 对残差检验的 p 值在 Lag=15 时小于 0.05，残差平方检验的 p 值在滞长为 15，20 时小于 0.05，ARCH-LM 检验的 p 值小于 0.05，均接受原假设，即 GARCH(1,1)模型拟合后的残差是仍然具有 ARCH 效应的，模型不是显著的。

同理用 R 进行了 GARCH(1,2)，GARCH(2,1)等模型的拟合，发现仍会出现模型参数不显著或者拟合模型后的残差仍然具有 ARCH 效应的问题。通过以上分析，本文最终确定用 ARCH (1) 模型来对 SARIMA(0, 0, 0)(2, 0, 0)<sub>12</sub> 均值模型的残差进行拟合。

最终模型为 SARIMA(0, 0, 0)(2, 0, 0)<sub>12</sub>+ARCH(1):

$$\begin{cases} X_t = -0.5247X_{t-12} - 0.3011X_{t-24} + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t}e_t \\ h_t = 601.1780 + 0.7418\varepsilon_{t-1}^2 \\ e_t \sim WN(0,1) \end{cases}$$

## 4.6 FFNN 模型

在 R 语言环境下，利用 nnet 包，以上海市 1997 年-2016 年的月平均气温（Temperature, TEMP）、相对湿度（Relative Humidity, RH）、时间变量年份（Year）以及月份（Month）作为输入层的四个节点，以上海市 1997 年-2016 年的月平均降雨量（PRCP）作为输出层的一个节点。其中，基于月平均降雨量呈现较强的季节效应，本文引入时间变量 Year 和 Month 来拟合季节效应。

FFNN 模型的构造，实则是网络输入、输出以及隐含层节点的选择。在已知训练样本（TEMP, RH, Year, Month, PRCP）的前提下，输入层与输出层的构建已经初步完成，网络结构的决定性因素则为隐含层节点个数的选择。而隐含层神经元个数的选择对网络的拟合能力、预测能力、泛化能力以及训练速度等都具有重要的影响。纵观国内外的相关研究，目前主要有三种隐含层节点个数的选择方法：

- (1)试验选择法。基于具体应用，通过多次试验来选择具有相对高泛化能力和输出精度的隐含层节点个数；
- (2)逐步增加法。从具有较少隐含层节点数的网络开始训练，逐步增加节点数，当网络输出误差减少到接受范围内时，确定隐含层节点个数。
- (3)网络裁剪法。从具有较多隐含层节点数的网络开始训练，根据各节点对网络输出误差的贡献大小，适当减少节点数从而确定隐含层节点个数。

本模型的隐含层节点个数选择采用第二种方法，即逐步增加法。隐含层节点最初设定为 1，以 1 为阶梯依次增加，增加至 5 时，误差结果如图 19。通过 Jarque Bera

Test ( $P<2.2e-16$ ) 以及 Q-Q Plot (图 20) 检验得出其不服从正态分布。

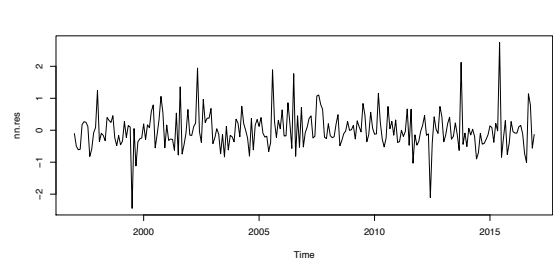


图 19 FFNN 训练误差

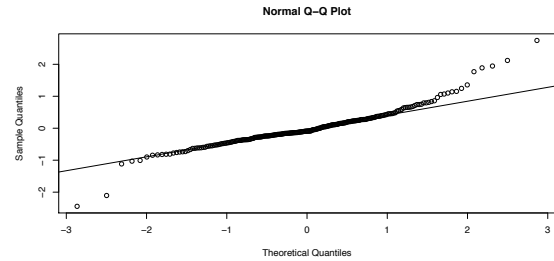


图 20 Q-Q 图

通过对训练误差及误差平方的 Box Test ( $\text{Lag}=12, P=0.8221$ ;  $\text{Lag}=12, P=0.8951$ ) 检验得出误差为白噪声序列并且无 ARCH 效应。说明隐含层包含 5 个节点的前馈神经网络模型显著，误差在可接受范围内。为确定其为最优的 FFNN，我们也对含有 6 个隐含层节点的网络进行了残差检验，发现误差衰减不明显。故最终确定隐含层节点数为 5，与试验选择法（隐藏层节点数为输入层节点数的 1.2-1.5 倍）的结果基本一致。

至此，输出月平均降雨量的前馈神经网络模型构建已基本完成。

在保证模型精度的前提下，为提高模型训练速度，我们对模型权重值的衰减精度以及最大迭代次数也有一定的限制。最终，我们选定权重衰减精度 (weight decay) 为 0.001，并设置模型最大迭代次数 (maximum number of iterations) 为 1000，即当模型权重值每次衰减小于 0.001 或迭代次数超过 1000 时，模型将不再进行迭代。

基于以上条件构造的 FFNN 模型结构如图 21 所示，图 22 为该模型的训练结果。

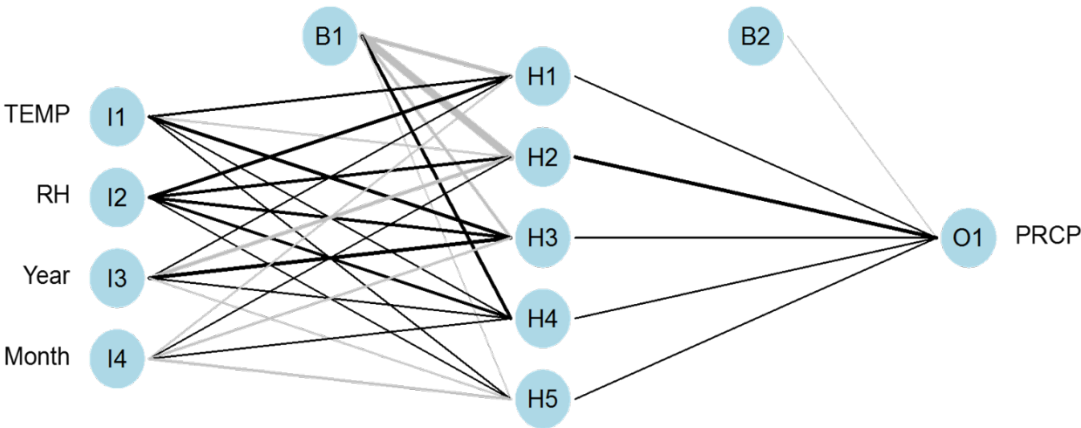


图 21 FFNN 模型结构图

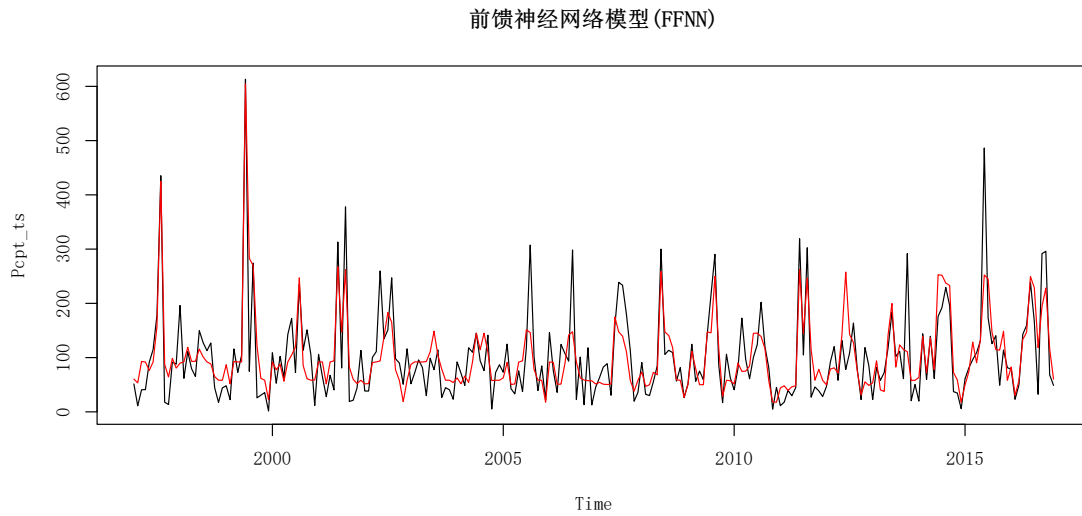


图 22 FFNN 模型拟合图

## 4.7 本章小结

对三个城市降雨量均进行了信号分解、Holt-Winters 三次指数平滑拟合，以及简单季节模型、复杂季节模型的拟合。三个城市的复杂季节模型均显著，但哈尔滨市对应模型的残差具有 ARCH 效应，因此进行了 GARCH 拟合，ARCH(1)拟合效果良好，成功去除 ARCH 效应。考虑到温度以及相对湿度对降雨量的影响，对上海市进行了前馈神经网络的拟合，并与复杂季节模型进行了预测误差比对分析。

对三个城市的最优模型选择如下：

广州市：  $(1 + 0.3222B^{12} + 0.3801B^{24})X_t = \varepsilon_t$

上海市：  $\text{TEMP} + \text{RH} + \text{Year} + \text{Month} \xrightarrow{\text{FFNN}} \text{PRCP}$

哈尔滨市：

$$\begin{cases} X_t = -0.5247X_{t-12} - 0.3011X_{t-24} + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t}e_t \\ h_t = 601.1780 + 0.7418\varepsilon_{t-1}^2 \\ e_t \sim WN(0,1) \end{cases}$$

## 5 最优模型预测

### 5.1 广州市降雨量预测——SARIMA 模型

#### 5.1.1 样本内预测

对广州市降雨量模型拟合选择的 SARIMA(0,0,0)(2,0,0)<sub>12</sub> 进行样本内预测。本文首先对去除 2 年的数据用 SARIMA(0,0,0)(2,0,0)<sub>12</sub> 模型向外预测两年。得到广州市 2015-2016 两年 24 个月度的降雨量，并与实际降雨量进行对比。由于没有找到样本残差的分布

布族，因此对于预测区我们采用加减一倍标准差进行处理。结果如下：

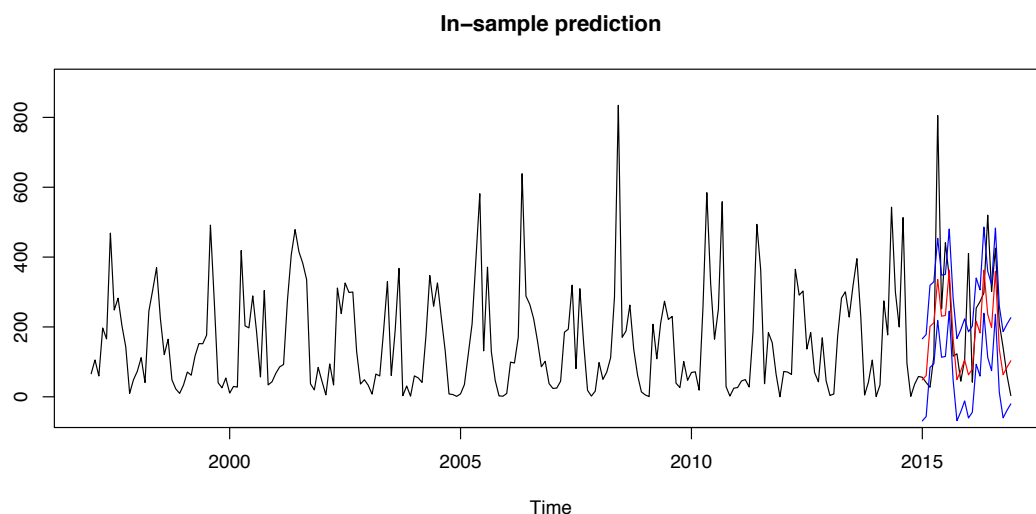


图 23 静态样本内预测

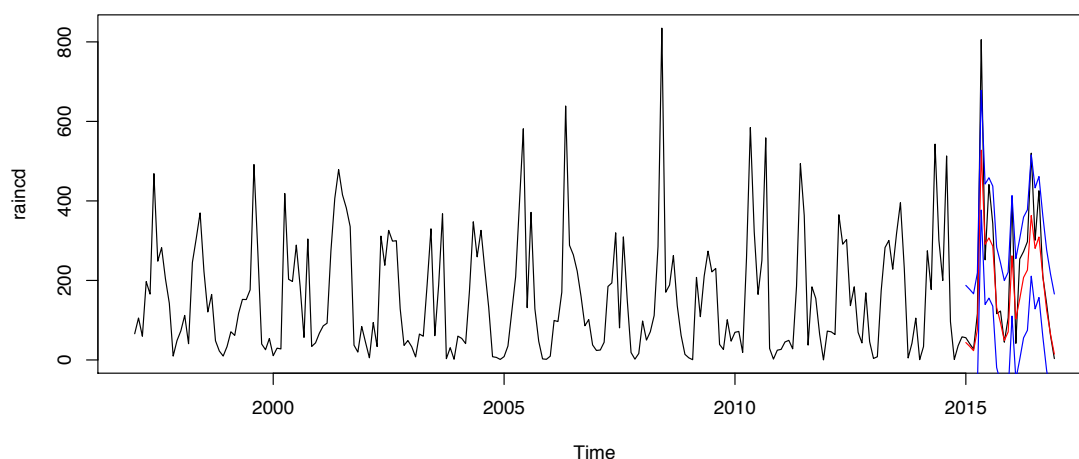


图 24 动态样本内预测

图中黑色为原始数据，红色为拟合的预测值，而蓝色为预测区间。发现静态预测图对于预测的效果没有动态预测的效果好。为确定原始数据与预测数据的区别，首先画出静态预测时间段的残差，发现 3 个异常值。对异常值所对应的原始数据进行分析，发现用于通过预测 2015-2016 月度降雨量的前 18 年数据并未有超过 650 的降雨量，而 2015 年 5 月的降雨量为 805.6，与前 18 年相比降雨量极大，通过原始数据难于预测。而除 3 个异常点其余的预测残差均在可接受范围内。分析动态预测时间段的残差，发现其残差绝对值小于静态预测残差绝对值，动态预测更优，接受模型  $SARIMA(0,0,0)(2,0,0)_{12}$ 。

残差图如下：

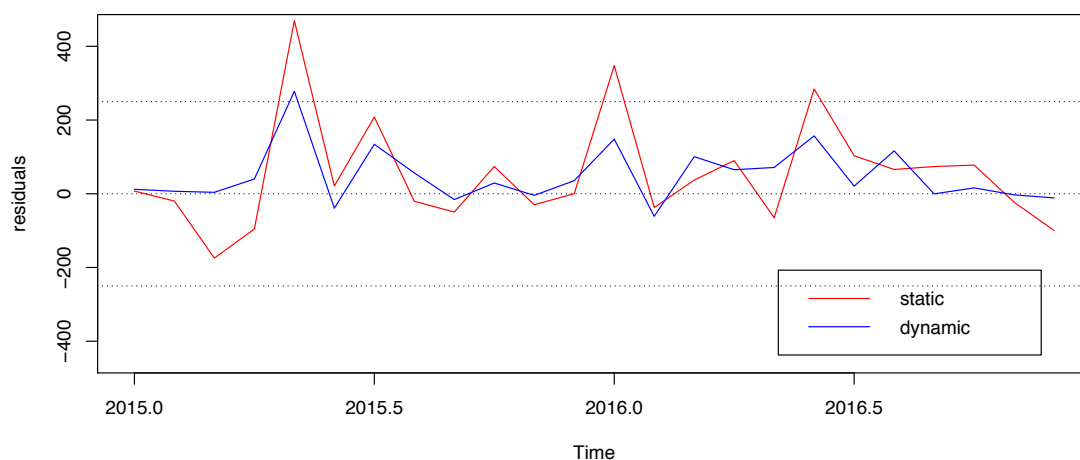


图 25 原始数据与预测值残差

### 5.1.2 样本外预测

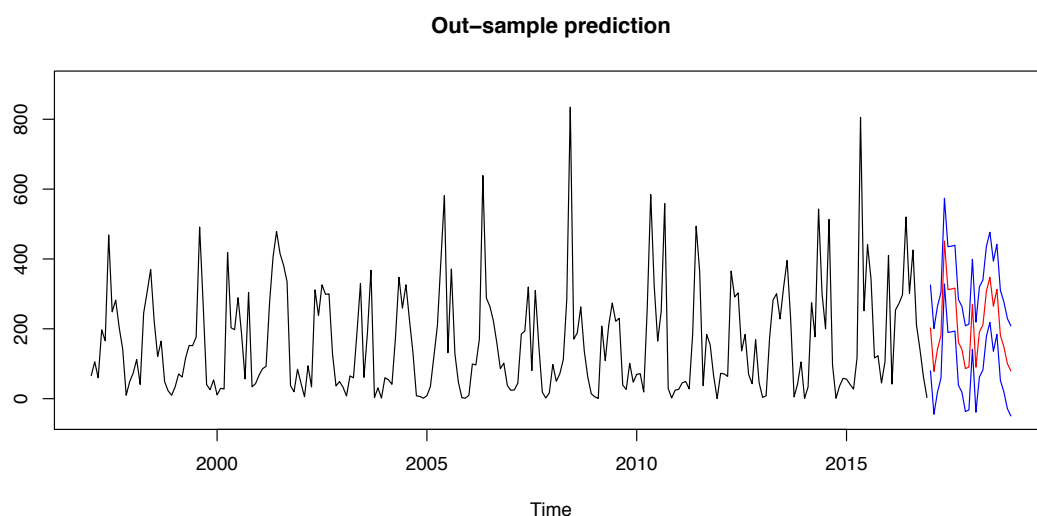


图 26 样本外预测

通过 R 语言，对拟合的模型进行样本外预测，结果如上图，红色为预测数据而蓝色部分是预测区间。认为在接下来的两年里，广州市的降雨量仍有周期为 12 的季节性，且在 6、7 月份降雨量最高。

## 5.2 哈尔滨市降雨量预测——SARIMA+ARCH 模型

### 5.2.1 样本内预测

SARIMA(0, 0, 0)(2, 0, 0)<sub>12</sub> 复杂季节模型静态样本内预测，本文先对去除 2 年的数据用 SARIMA(0, 0, 0)(2, 0, 0)<sub>12</sub> 模型拟合后向外预测两年。即得到 2015 年 1 月到 2016 年 12 月的哈尔滨市月均降雨量，与实际相应时间段的数据进行对比。同时因为样

本残差不服从正太分布，同广州市降雨量预测，预测区间仍采用加减一倍的标准差。R 统计软件运行结果如图 27 所示。

动态预测及每预测出一个数据，对应的复杂季节模型重新拟合得到新的参数，下一个数据根据新的拟合模型来预测。对应的波动率及标准差同理。我们预测遇见也采用加减一倍的标准差来处理。用 R 软件运行结果如图 28 所示。

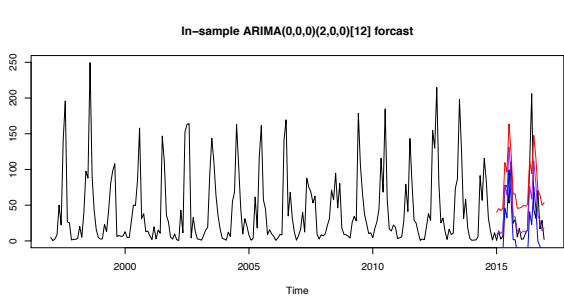


图 27 静态样本内预测

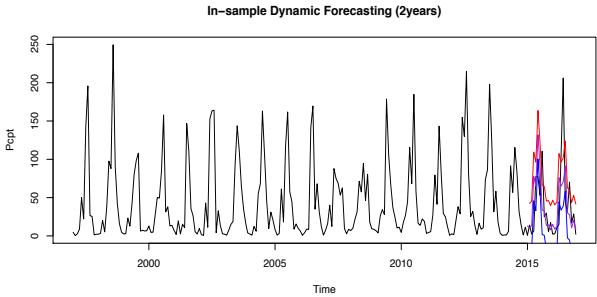


图 28 动态样本内预测

图中紫色为拟合的预测数据，黑色为原始数据，红色和蓝色分别为预测区间的上下界，从图中可以看出动态预测拟合数据波动性体现更明显。但是预测区间对于极端高于均值的数据还是无法包含，预测存在缺陷。

对复杂季节模型动态预测和静态预测的残差做出如下对比图，其中蓝线代表动态预测，红线代表静态预测。

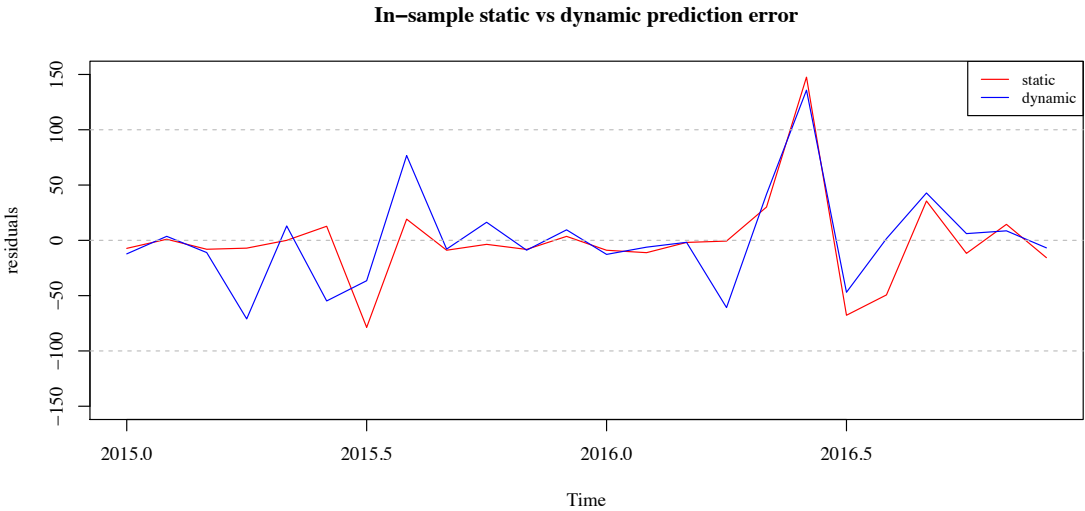


图 29 静动态预测误差分析

从图 29 可看出两种预测差距不大，预测后的残差都在 0 附近，除去一个异常值点，大多残差落在可接受范围内。所以均值方程复杂季节模型预测效果较好。



上文最优模型是包含 ARCH(1)模型的复杂季节模型，引入条件异方差对数据预测的标准差的影响。图 30 为标准差的预测，图 31 为加上条件异方差影响的样本内动态预测。

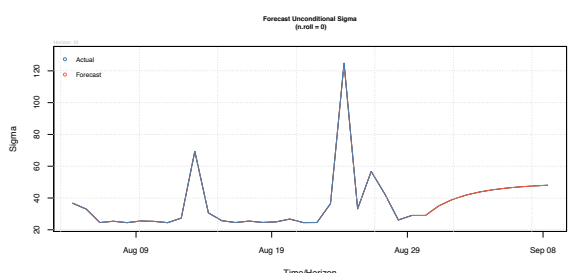


图 30 无条件方差预测图

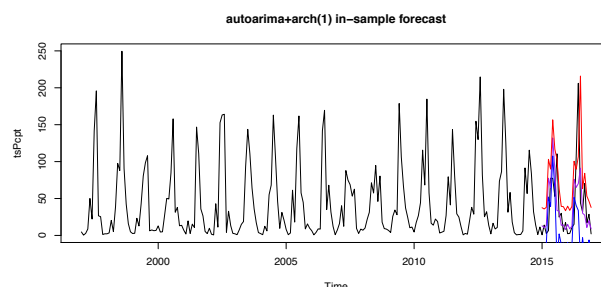


图 31 样本内动态预测（含异方差）

从图中看出加上条件异方差的影响后预测区间近似涵盖所有样本点，预测效果比不引入 ARCH(1)的效果更好。综上 SARIMA(0,0,0)(2,0,0)<sub>12</sub>+ARCH(1)动态预测最好。

### 5.2.2 样本外预测

SARIMA(0,0,0)(2,0,0)<sub>12</sub> 样本外预测如图 32，其中蓝色部分为预测均值，阴影部分为预测区间。从图中看出 SARIMA(0,0,0)(2,0,0)<sub>12</sub> 的样本外预测效果比 Holt-Winters 要很多。

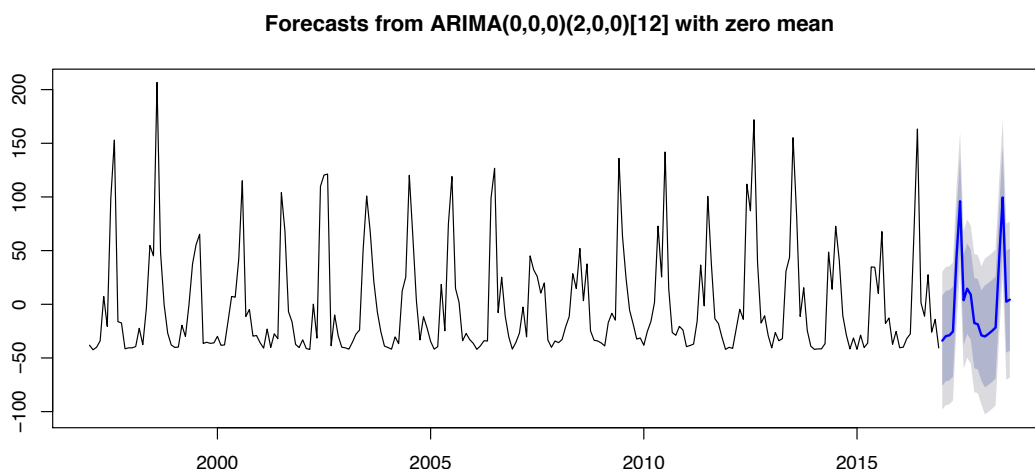


图 32 SARIMA 样本外预测

通过样本内和样本外预测，本文认为 SARIMA(0,0,0)(2,0,0)<sub>12</sub>+ARCH(1)模型预测效果较好。

## 5.3 上海市降雨量预测——FFNN 模型

### 5.3.1 样本内预测

为选择上海市月均降雨量的最优模型，本文分别进行了复杂季节模型与前馈神经网络模型样本内 24 步预测误差分析。图 33 直观显示了复杂季节模型与前馈神经网络模型 2015.1-2016.12 的预测误差，可以看出前馈神经网络模型的预测误差绝对值整体



小于复杂季节模型。因此，本文认为 FFNN 的预测能力优于复杂季节模型，并选择 FFNN 为最优模型。

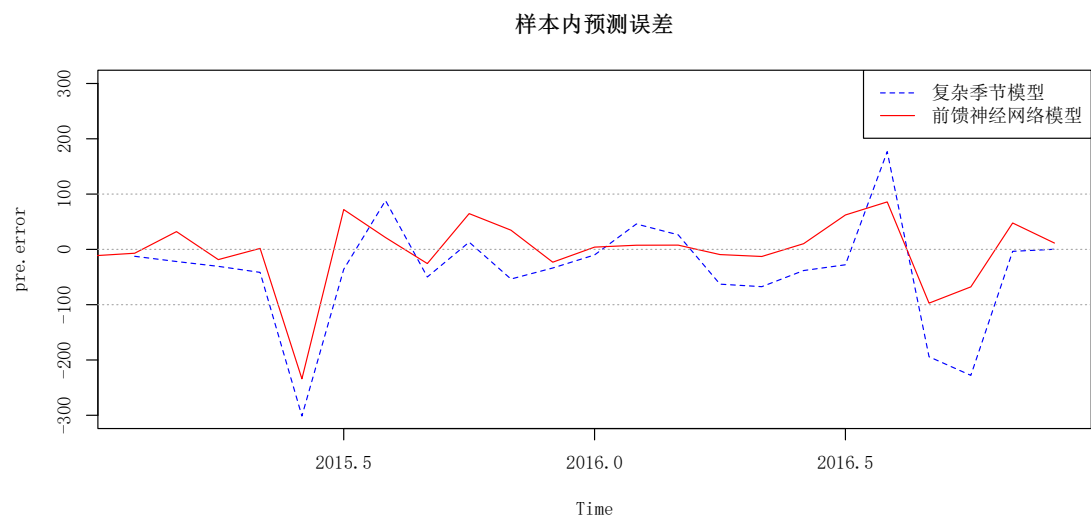


图 33 SARIMA 与 FFNN 预测误差比较

图 34 展示了 FFNN 模型的样本内 24 步预测结果，红色为预测部分。

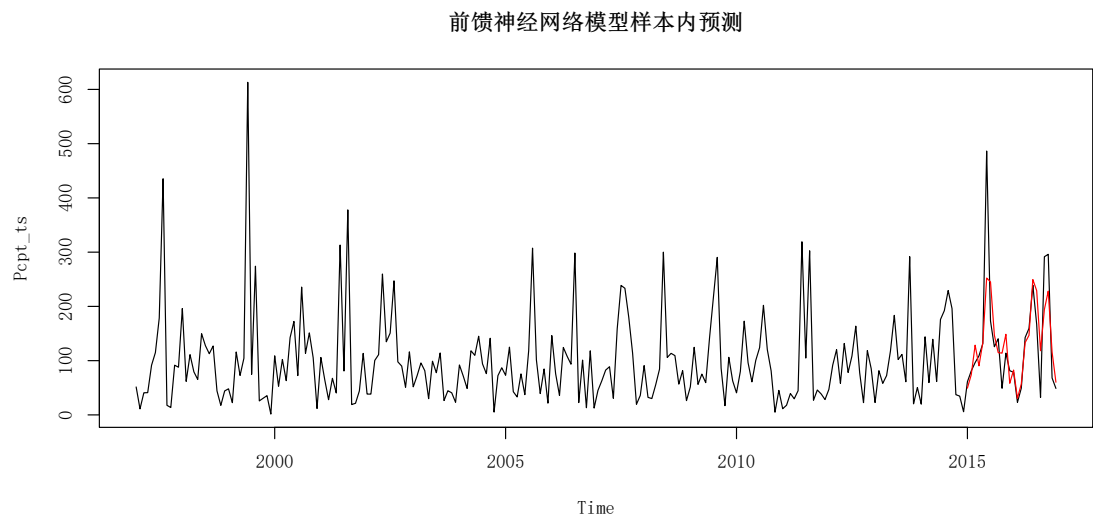


图 34 FFNN 样本内预测

5.3.2 样本外预测

图 35 展示了 FFNN 模型的样本外 24 步预测结果，红色为预测部分。

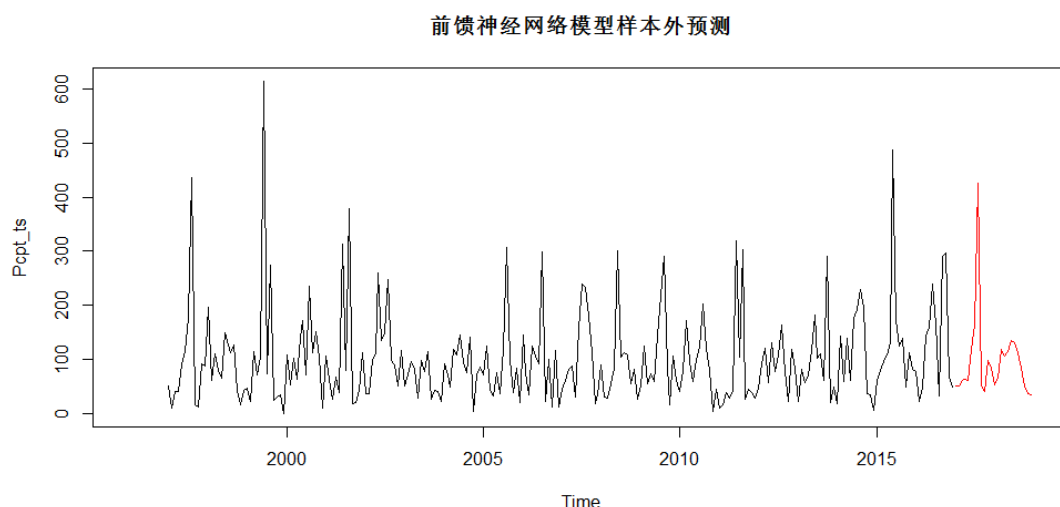


图 35 FFNN 样本外预测

## 6 总结与展望

### 6.1 总结

通过对三个城市的降雨量月度数据进行分析，可以明确从广州市到哈尔滨市，由南至北降雨量依次降低。为确定地区不同，季节模型的拟合是否相似，对三个城市首先进行了时序图和 ACF、PACF 图进行联合分析，三者均具有季节性。验证序列平稳性后，通过谱图证明三者均无长记忆性。其次，选择简单季节模型拟合，均不是优良的拟合模型。随后进行复杂季节模型拟合分析，前两者均可采用  $SARIMA(0,0,0)(2,0,0)_{12}$  进行拟合，其中哈尔滨市的模型拟合残差具有 ARCH 效应，采用 ARCH(1)成功去除 ARCH 效应，而上海市的月度降雨量由于存在一个最大异常值，模型阶数略有不同，拟合模型为  $SARIMA(0,0,0)(0,1,1)_{12}$ 。最后为了研究温度与相对湿度对降雨量的预测是否有实际意义，采用前馈神经网络模型进行拟合并验证二者显著增强降雨量模型预测能力。选出三个城市的最优拟合模型后，进行样本内预测与样本外预测。为分析拟合效果，对静态样本内预测和动态样本内预测的残差进行比对，模型的动态预测优于静态预测且与原始数据的残差均在可接受范围内，模型预测能力较强。并且样本外预测了后续两年的降雨量，具有现实意义。

### 6.2 创新点

本文针对从南至北三个城市（广州市、上海市、哈尔滨市）的月均降雨量数据，进行了多角度的季节效应分析与建模。

对去除季节影响后的降雨量趋势进行了鲁棒局部加权回归，通过鲁棒权重的调整减少数据异常值对回归的影响。并引入温度、相对湿度两个降雨量影响变量以及年份、月份两个时间变量，结合试验选择法与逐步增加法确定五个隐含层节点，建立了一个具有

相对高泛化能力和输出精度的前馈神经网络模型。在保证模型精度的前提下，通过权重衰减精度以及最大迭代次数的限制，提高模型训练速度。

### 6.3 展望

本文最终对三个城市的月均降雨量建立了不同的最优模型，但由于时间的限制，在建立模型过程中存在一些不足有待改进，主要有以下几点：

- (1)模型进行预测时，由于模型残差没有找到合适的分布，预测区间我们选择加减一倍标准差来进行预测。关于数据分布的拟合，可以查阅相关文献并进行深入研究。
- (2)模型预测的误差没有运用合理的统计量进行量化估计，在预测误差分析过程中，考虑到均方误差对模型预测误差并不能很好的做到量化描述，本文并没有采用均方误差来进行研究。
- (3)通过多次训练，本文采取单层前馈神经网络模型。关于隐含层层数的确定，有待深入分析。

## 参考文献

- [1]F. Rosenblatt. The Perceptron:A Probabilistic Model for Information Storage and Organization in the Brain[J]. Psychological Review,1958,65(6):386-408.
- [2]Kolmogorov A N. On the Representation of Continuous Function of Many Variables by Superposition of Continuous Functions of One Variable and Addition[J]. American Math Society Trans,1963,28(1):55-59.
- [3]M. Kendall and A. Stuart .The Advanced Theory of Statistics[M].1983(3):California: Griffin:410-414.
- [4]TPChen, HChen, RWLiu. Approximation Capability in  $C(R^n)$  by Multilayer Feedforward Networks and Related Problems[J].IEEE Trans Neural Networks,1995,6:25-30.
- [5]Jin-Chuan Duan. Augmented GARCH(p,q) Process and its Diffusion Limit[J].Journal of Econometrics,1997,79:97-127.
- [6]Zhang Daiyuan.Zero Cost Function Training Algorithms for Three-Layered Feedforward Neural Networks[J].Journal of Systems Engineering and Electronics,2000,11(2):32-38.
- [7]张德贤.前向神经网络合理隐含层结点个数估计[J].计算机工程与应用,2003,39(5): 21-23.
- [8]Charles C. Holt.Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages[J]. International Journal of Forecasting,2004,20:5-10.
- [9]McLeod A I.Robust Loess: S Lowess[EB/OL].<http://www.stats.uwo.ca/faculty/aim/2004/04-259/notes/RL.pdf>.
- [10]蒋乐天等基于强局部加权回归算法的软件老化趋势提取[J].上海交通大学学报,2006,11(4):1951-1954.
- [11]赵凌,张健,陈涛.基于 ARIMA 的乘积季节模型在城市供水量预测中的应用[N].水资源与水工程学报,2011,22(1):59-62.
- [12]涂雄苓,黄月玲.旅游需求模型的 ARIMA 乘积季节模型构建及实证分析[J].广西财经学院学报,2011,24(1):112-117.
- [13]A.L.S.Maia, F.A.T.de Carvalho. Holt's Exponential Smoothing and Neural Network Models for Forecasting Interval-Valued Time Series[J]. International Journal of Forecasting, 2011,27:740-759.
- [14]M.Szmit and A. Szmit. Usage of Modified Holt-Winters Method in the Anomaly Detection of Network Traffic: Case Studies[D]. Journal of Computer Networks and Communications, 2012:1-5.
- [15]虞乐,肖基毅.数据挖掘中强局部加权回归算法实现[J].电脑知识与技术,2012,8(7): 1493-1495.

- [16]于志军,杨善林.基于误差校正的 GARCH 股票价格预测模型[J].中国管理科学,2013,21:341-344.
- [17]Elsevier B.V. A Time Varying GARCH(P,Q) Model and Related Statistical Inference[J]. Statistics and Probability Letters,2013(83):1983–1990.
- [18]Paulo Rotela Junior,Fernando Luiz Riêra Salomon,Edson de Oliveira Pamplona.ARIMA: An Applied Time Series Forecasting Model for the Bovespa Stock Index[J].Applied Mathematics,2014:3383-3391.
- [19]Chulin Likasiri,Eakkapong Duangdai,Radom Pongvuthithum.Mathematical Model on the Effects of Global Climate Change and Decreasing Forest Cover on Seasonal Rainfall in Northern Thailand[J].Ecological Modelling,2014(272):388-393.
- [20]谭立云,刘海生,谭龙.逐步线性回归与神经网络预测的算法对比分析[J].华北科技学院学报,2014,11(5):60-65.
- [21]Rong-Fang Xu,Thao-Tsen Chen,Shie-Jue Lee. Weighted Learning for Feedforward Neural Networks[J].Journal of Electronic Science and Technology,2014,12(3):299-304.
- [22]Daniel Eni,Fola J.Adeyeye.Seasonal ARIMA Modeling and Forecasting of Rainfall in Warri Town,Nigeria[J].Journal of Geoscience and Environment Protection,2015:91-98.
- [23]Jaesung Choi,David C.Roberts,EunSu Lee.Forecasting Oil Production in North Dakota Using the Seasonal Autoregressive Integrated Moving Average (S-ARIMA)[J].Natural Resources,2015:16-26.
- [24]Quang Thanh Tran,Zhihua Ma, Hengchao Li,Li Hao,Quang Khai Trinh. A Multiplicative Seasonal ARIMA/GARCH Model in EVN Traffic Prediction[J].Int.J.Communications, Network and System Sciences,2015:43-49.
- [25]Dominika Dąbrowska, Marek Sołtysiak, Jan Waligóra. Short-Term Forecasting of the Chloride Content in the Mineral Waters of the Ustroń Health Resort Using SARIMA and Holt-Winters Models[J]. Environmental & Socio-economic Studies, 2015, 3(4): 57-65.
- [26]王玉全.地表温度时间序列分析方法的研究[D].山东农业大学,2016.
- [27]林文,王宏,谢祖欣,陈彬彬,余永江.福州市低能见度天气特征及季节模型[N].自然灾害学报,2016,23(1):147-157.
- [28]Liu Qiang,Gao Ming,Zhang Tao,Zhang Qijun. Feedforward Neural Network Models for FPGA Routing Channel Width Estimation[J]. Chinese Journal of Electronics,2016,25(1):71-76.
- [29]Mladenović, Lepojević,Janković-Milić.Modelling and prognosis of the export of the republic of Serbia by using seasonal Holt-Winters and ARIMA method[J].Economic themes, 2016,54(2):233-260.
- [30]黄艳华.乘积季节模型  $ARIMA(p,d,q) \times (P,D,Q)_s$  在 CPI 分析中的应用[N].重庆工商大学

学报(自然科学版),2016,33(3):70-75.

[31]Nicholas Fournier, Eleni Christofa,Michael A.Knodler Jr.A Sinusoidal Model for Seasonal Bicycle Demand Estimation[J]. Transportation Research Part D,2017(50):154-169.

[32]Eakkapong Duangdai,Chulin Likasiri.Rainfall Model Investigation and Scenario Analyses of the Effect of Government Reforestation Policy on Seasonal Rainfalls:A Case Study from Northern Thailand[J]. Atmospheric Research,2017(185):1-12.

[33]张涛,邓晓卫,李凡一,张苏婧.基于 GARCH 模型的融资融券对我国股市的影响研究[J]. 金融,2018,8(1):26-55.

[34]马晓梅,史鲁斌,其木格等.基于 ARIMA 乘积季节模型和 Holt-Winters 季节模型的梅毒月发病率预测[J].郑州大学学报,2018,53(1):80-84.