Report of Assignment_04

胡兆平 12032344

4.1 Plotting with ggplot2

使用美国爱荷华州五个气象观测站的降水数据进行绘图

4.1.1 Boxplot

逐日的降水数据有很多 0 值，因此需要先将其转化为逐月的降水，再绘制箱线图。

以名称作为分组依据，月降水量为指标绘制箱线图。

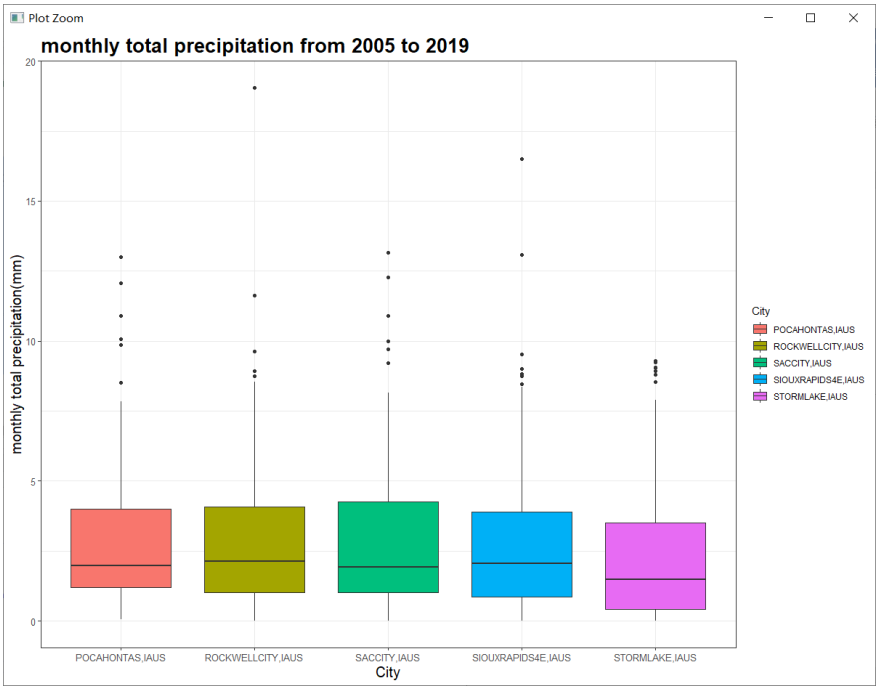

Fig 4.1.1 Boxplot

4.1.2 Time series

时间序列的绘制首先需要用 as.Date()命令将字符串转化为时间，之后以时间为横坐标，日降水量为纵坐标绘图，由于五个站点的降水趋势非常接近，因此将其绘制在不同的坐标系中。

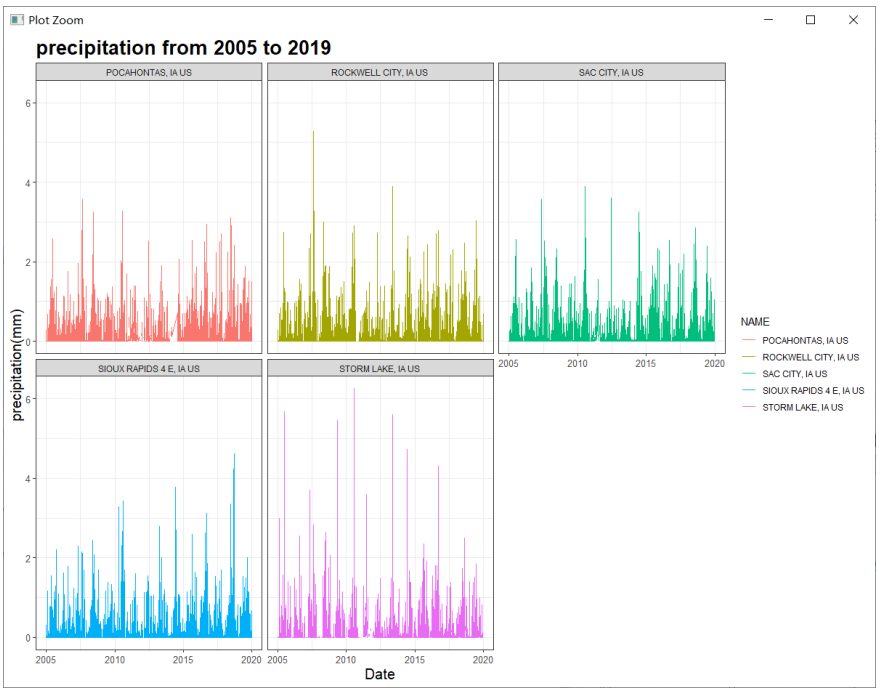

Fig 4.1.2 Time series

4.1.3 Histogram

直方图使用逐日的降水数据效果也不好，因此先将其转化为逐月的降水数据。以逐月的降水数据为变量，将五个站点的数据分别绘制到不同坐标系。

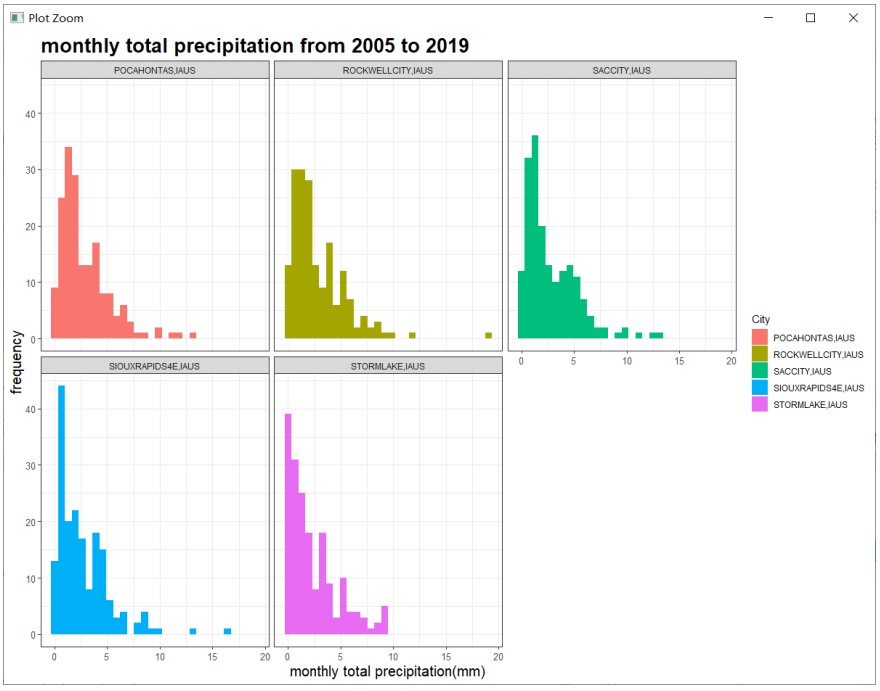

Fig 4.1.3 Histogram

4.1.4 Scatter plot

逐日的散点图线性关系太差，因此使用逐月的数据。选择的城市为 ROCKWELLCITY 和 SACCITY

由于所有站点的数据均在一列，因此需要先单独导出 ROCKWELLCITY 和 SACCITY 的数据。再将这两组数据合并到新的矩阵中，并求线性回归，用于添加趋势线。

以 ROCKWELLCITY 为 x 轴，以 SACCITY 为 y 轴绘制逐月降水数据的散点图，并添加趋势线。

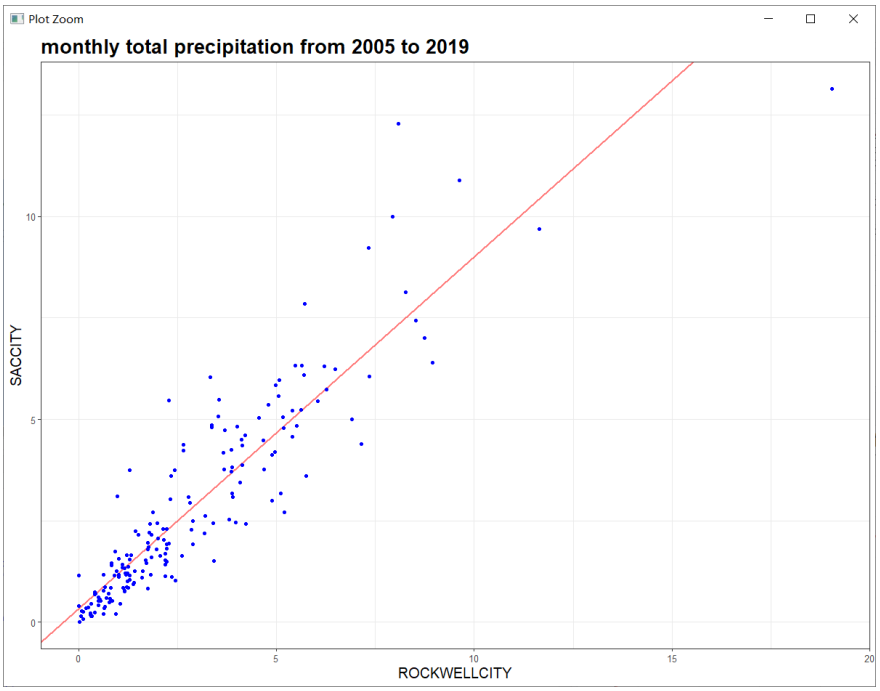

Fig 4.1.4 Scatter plot

4.1.5 Image plot

绘制今年一月份全球的 ndvi（归一化植被指数），数据来源为 Giovanni 的 MODIS 数据。由于数据量较大，绘图过程可能会很慢。

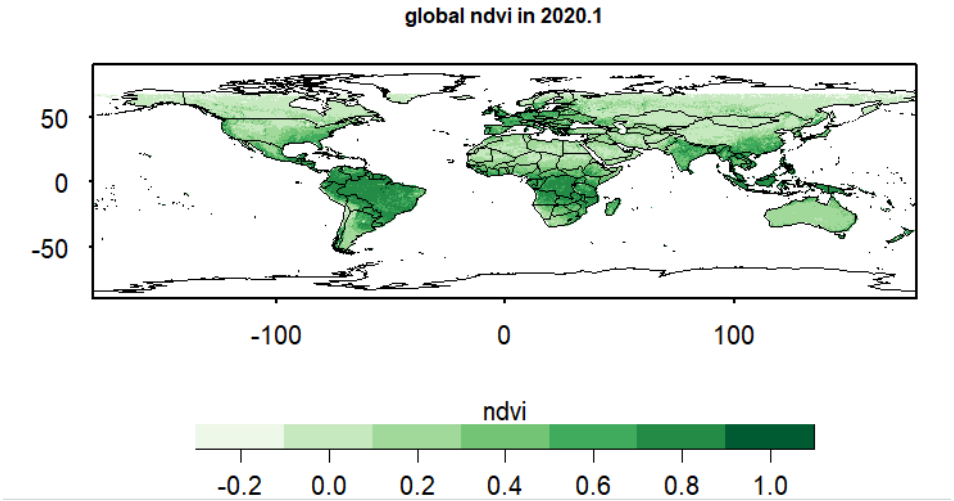

Fig 4.1.5 Image plot

4.2 Analysis of the time series of monthly temperature

首先进行数据的预处理，补充 2020 年 9 月和 10 月的数据，去除温度的异常值，计算月平均气温。

4.2.1 Construct a time series

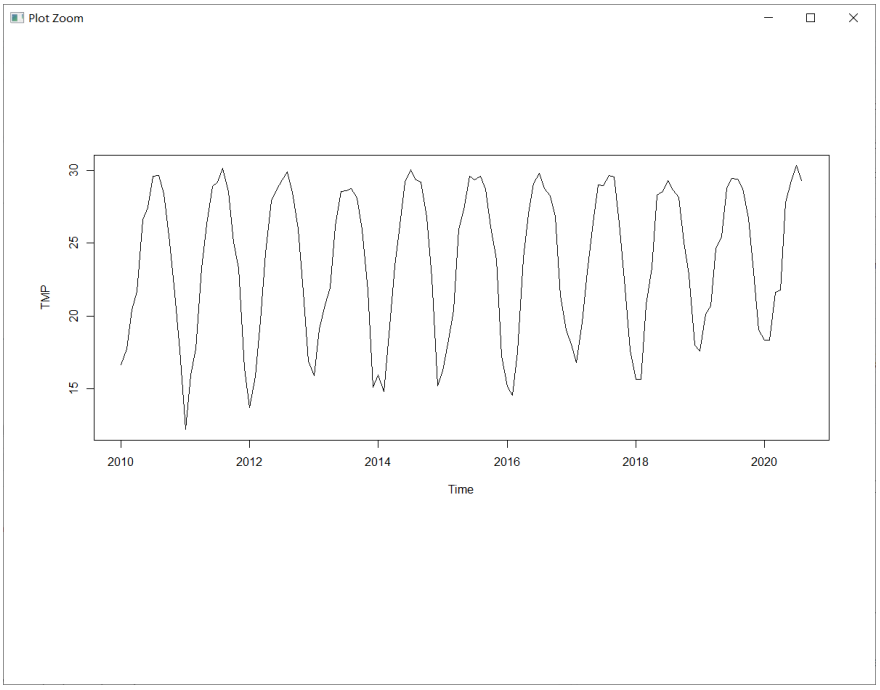

Fig 4.2.1 time series

4.2.2 Decompose the time series

分解时间序列得到结果如图 4.2.2。

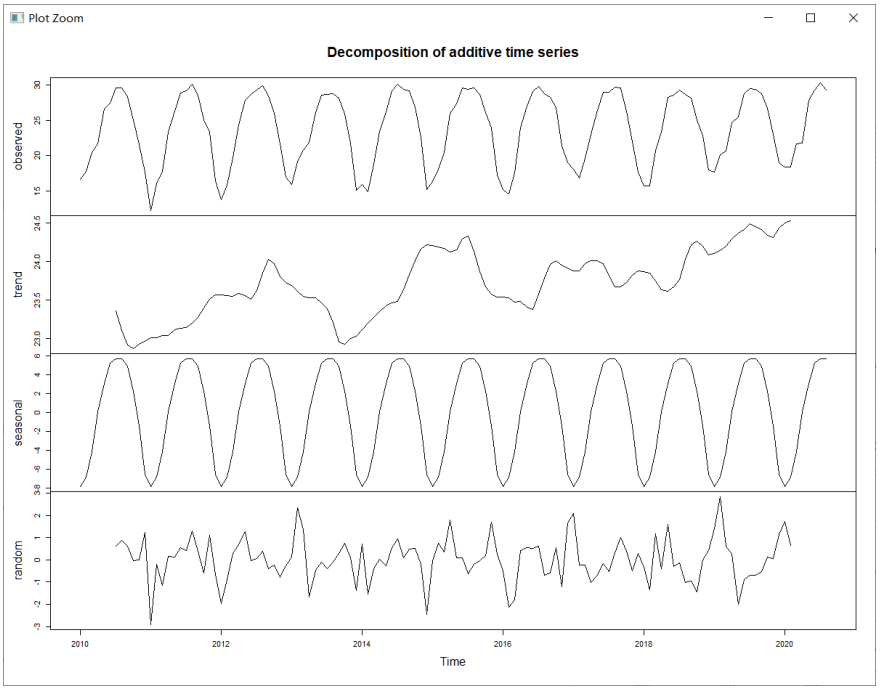

Fig 4.2.2 Decomposition od additive time series

根据直方图和 Box.test（p-value = 0.01904 < 0.05）,可以判断误差遵循高斯白噪声分布

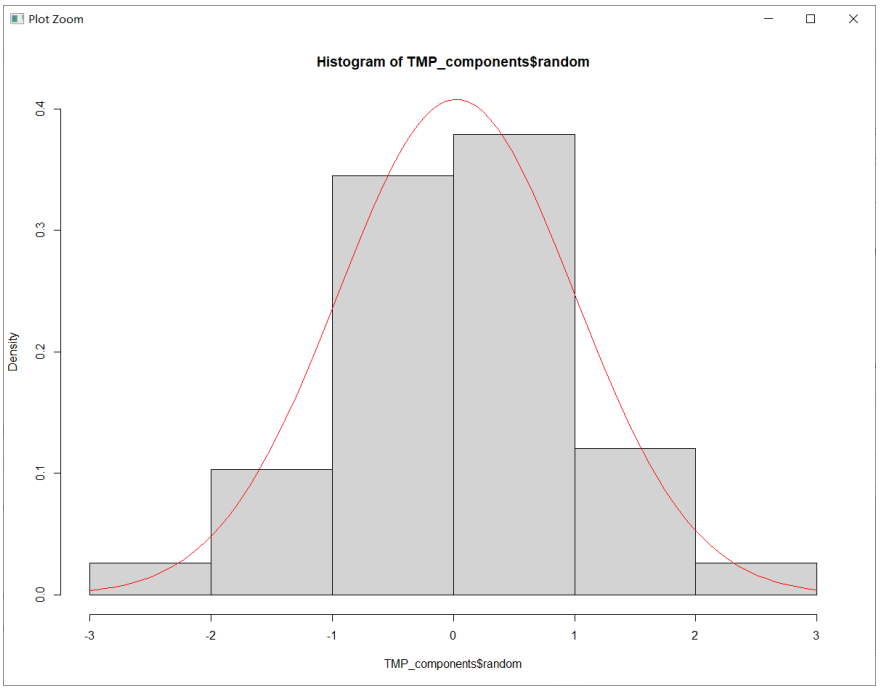

Fig 4.2.3 Histogram of random

4.2.3 ARIMA model

先通过 acf()和 pacf()进行观察，可以发现 TMP 是平稳的，因此可以不进行差分。

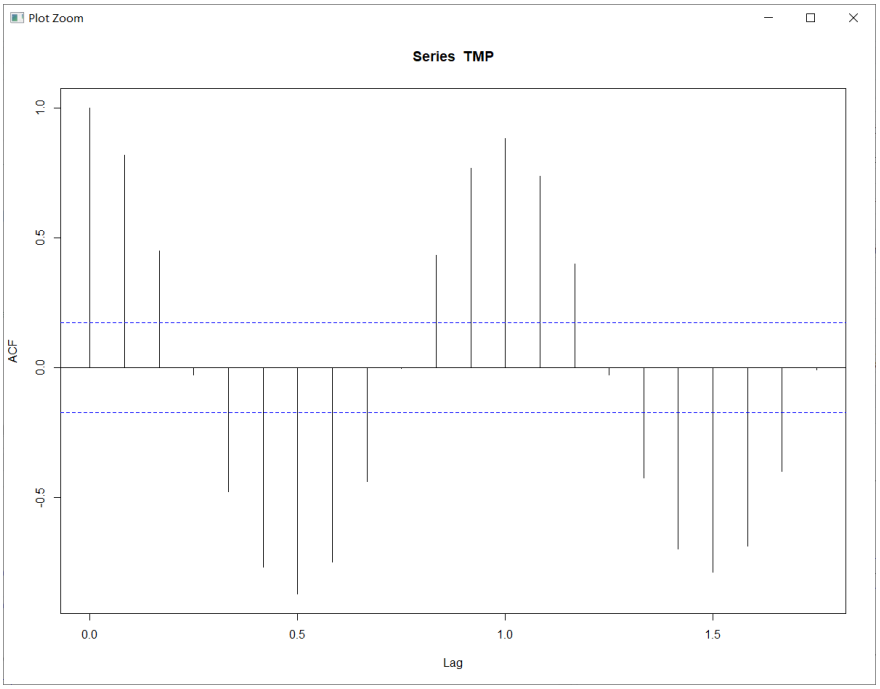

Fig 4.2.4 acf

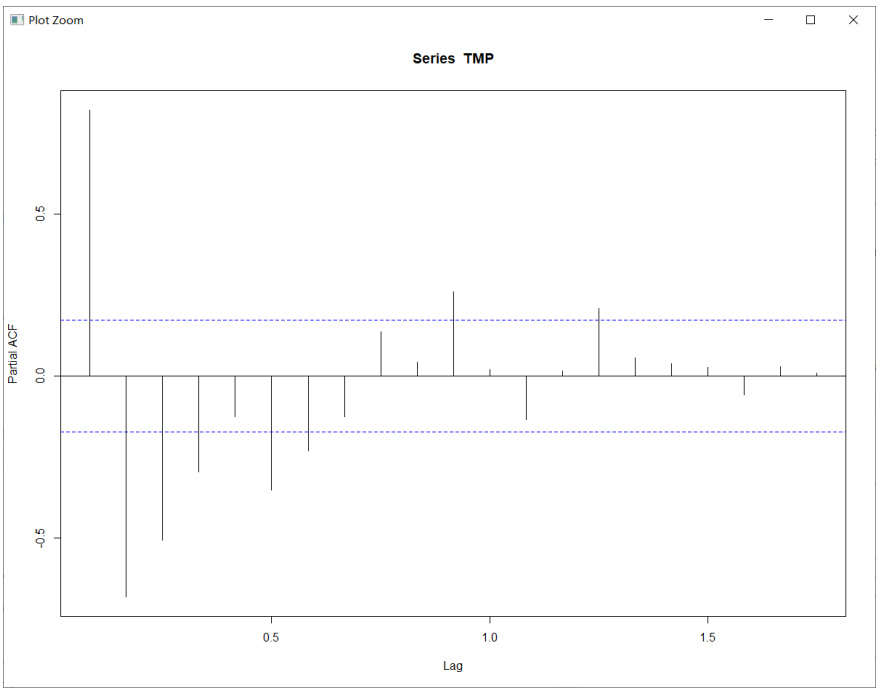

Fig 4.2.5 pacf

由于一开始对建模过程不太清楚，因此我们先用 auto.arima()获取最佳模型，再反推建模过程，推导过程中和王超同学进行了讨论。

auto.arima()得到的模型为 ARIMA(0,0,2)(1,1,1)[12]，共分为两个部分：

1)、第一组参数(0,0,2)代表去除季节性的 ARIMA 模型，是满足 p = 0 : 5,q = 0 : 5,d = 0 : 2 的条件下获得所有模型中，AIC 最小的那个。其中，p，q 取值均为 0 : 5;由于 TMP 是平稳的，因此 d = 0 : 2，否则要根据差分情况确定 d 的取值。

AIC 最小可以通过以下方法证明，首先用 TMP 减去分解时间序列时得到的季节因素，得到校准的 TMP_adjust，再求 TMP_adjust 不同 p、d 和 q 值的模型，比较不同模型的 AIC，可以发现(0,0,2)确实是最小的一组。

```
> min_aic <- which(aic[,1]==min(aic[,1]))
> aic[min_aic,]
[1] "376.600079452426" "0 0 2"
>
```

Fig 4.2.6 the result of min AIC

2)、第二组参数(1,1,1)[12] 代表季节性，由 TMP_components$seasonal 可知，它是一个周期性循环的时间序列。由于季节性变化的周期为 12，因此对原数据进行间隔为 12 的差分，观察其 acf 和 pacf，显然 P、Q 和 D 均为 1

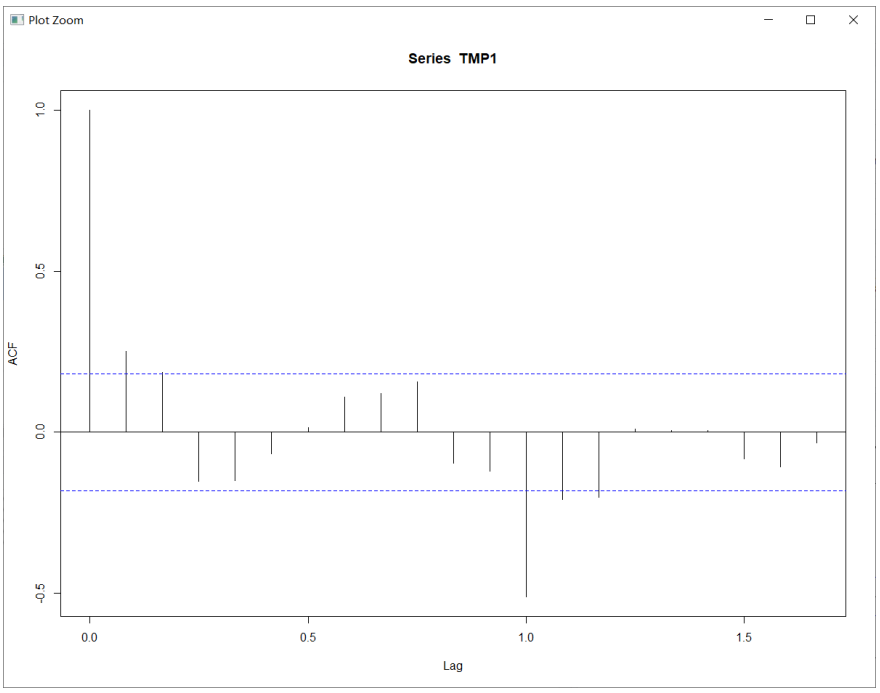

Fig 4.2.6 acf of TMP1

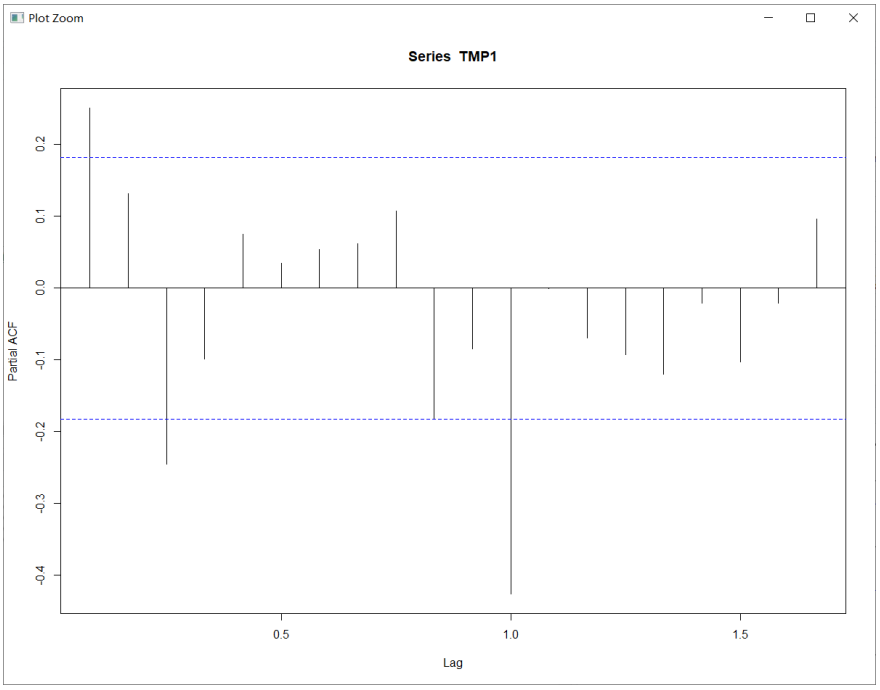

Fig 4.2.7 pacf of TMP1

综上，可以得知最优模型为 ARIMA(0,0,2)(1,1,1)[12]。

4.2.4 Predict

预测结果图如下

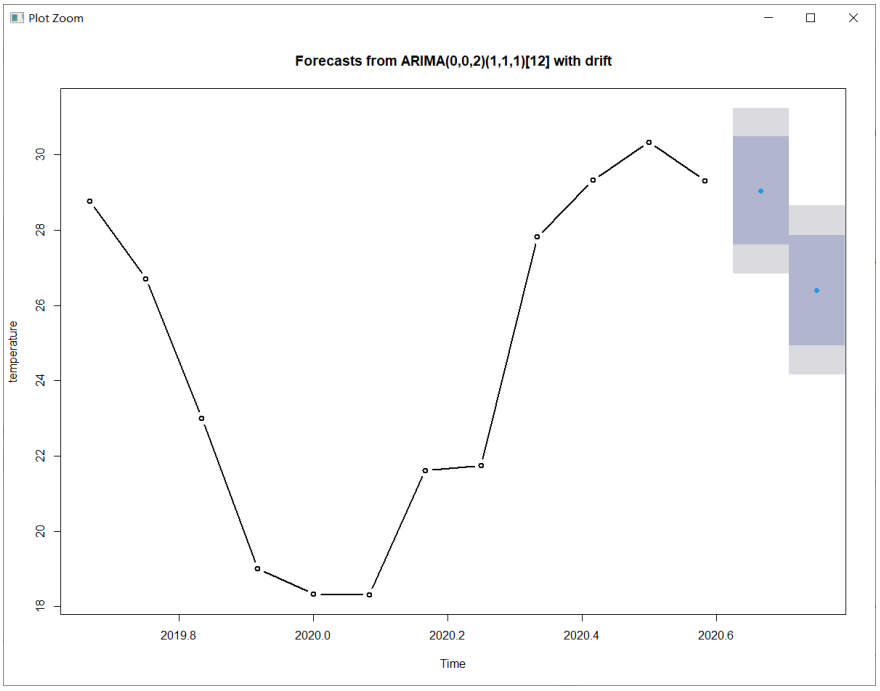

Fig 4.2.8 predication of Sep and Oct

测得九月份相对误差为 0.84%，十月份相对误差为 2.47%。