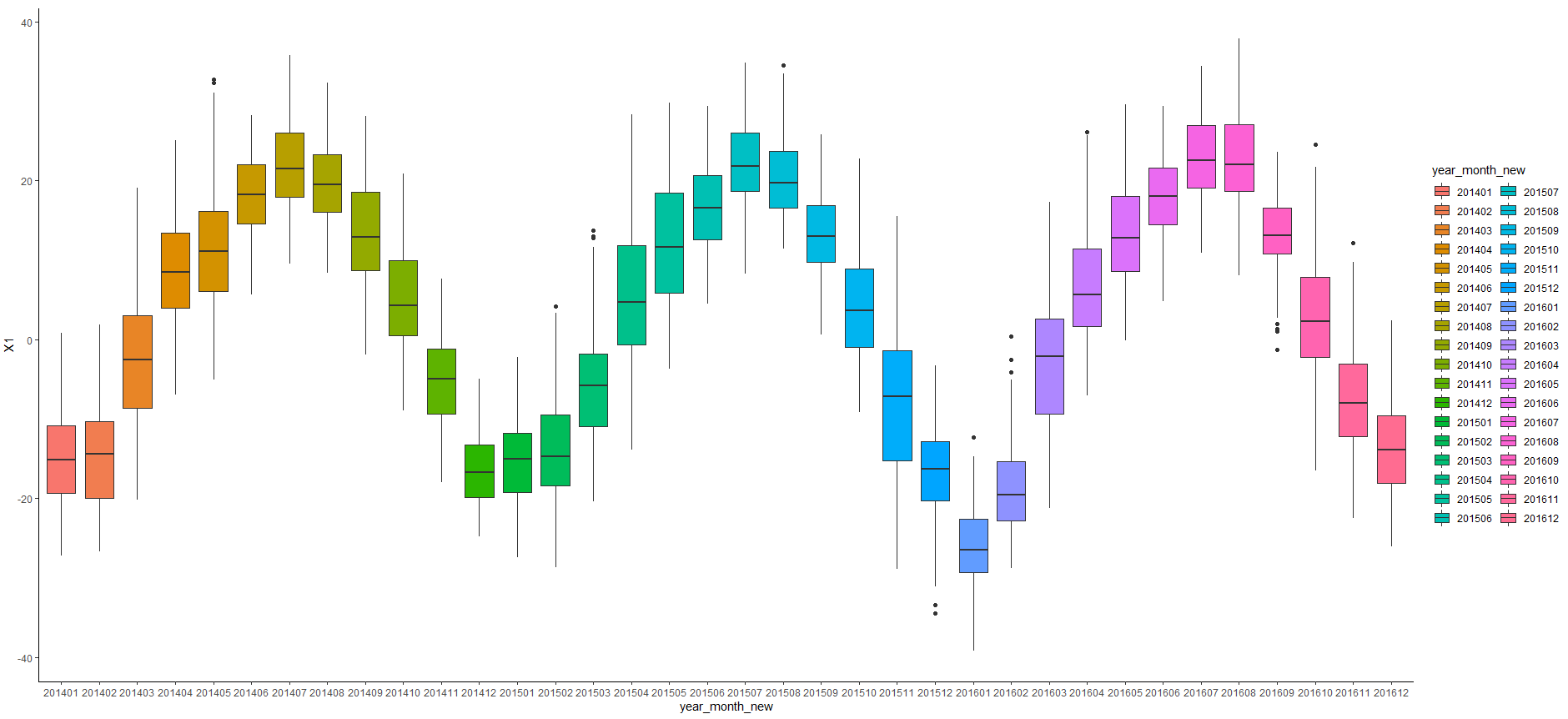
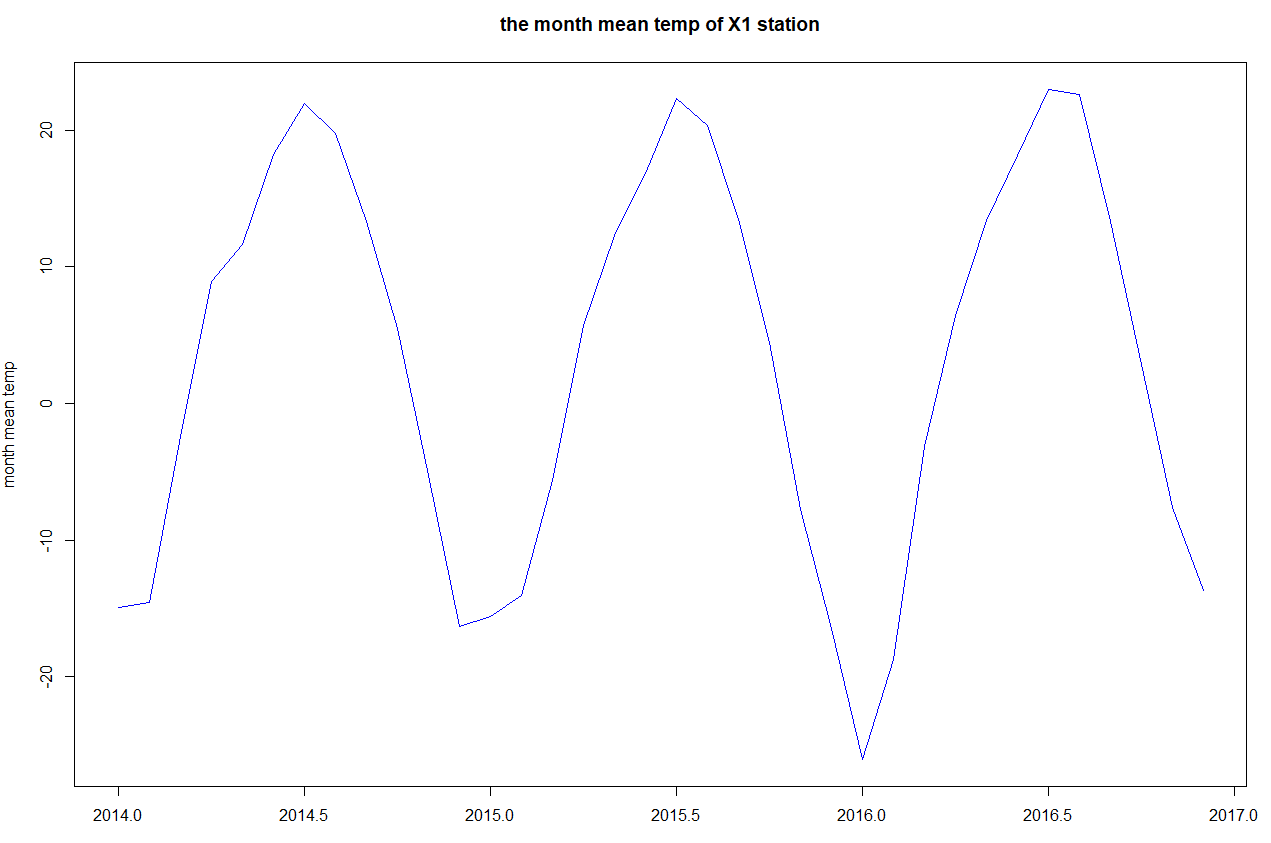
# Report

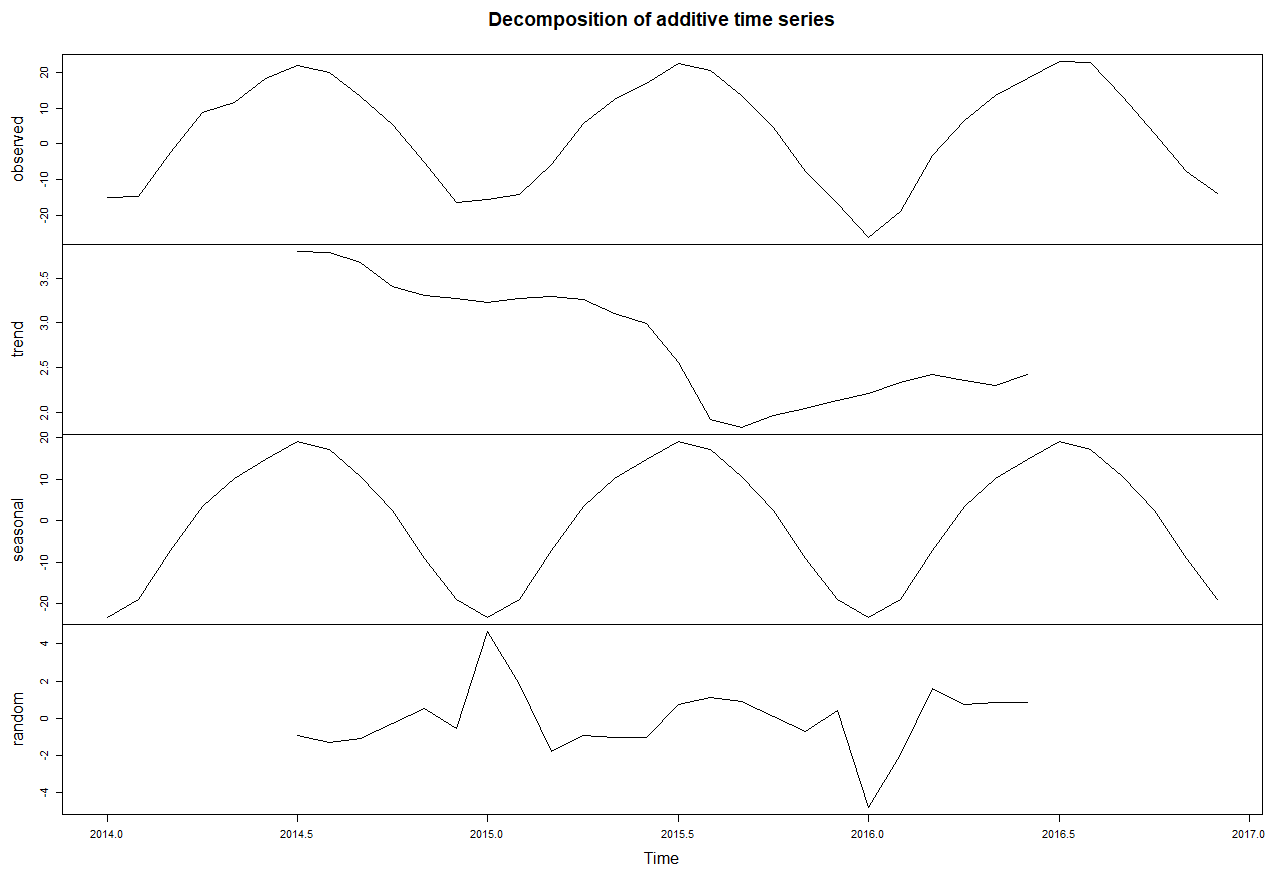
## PS4\_1

# 1. Boxplot

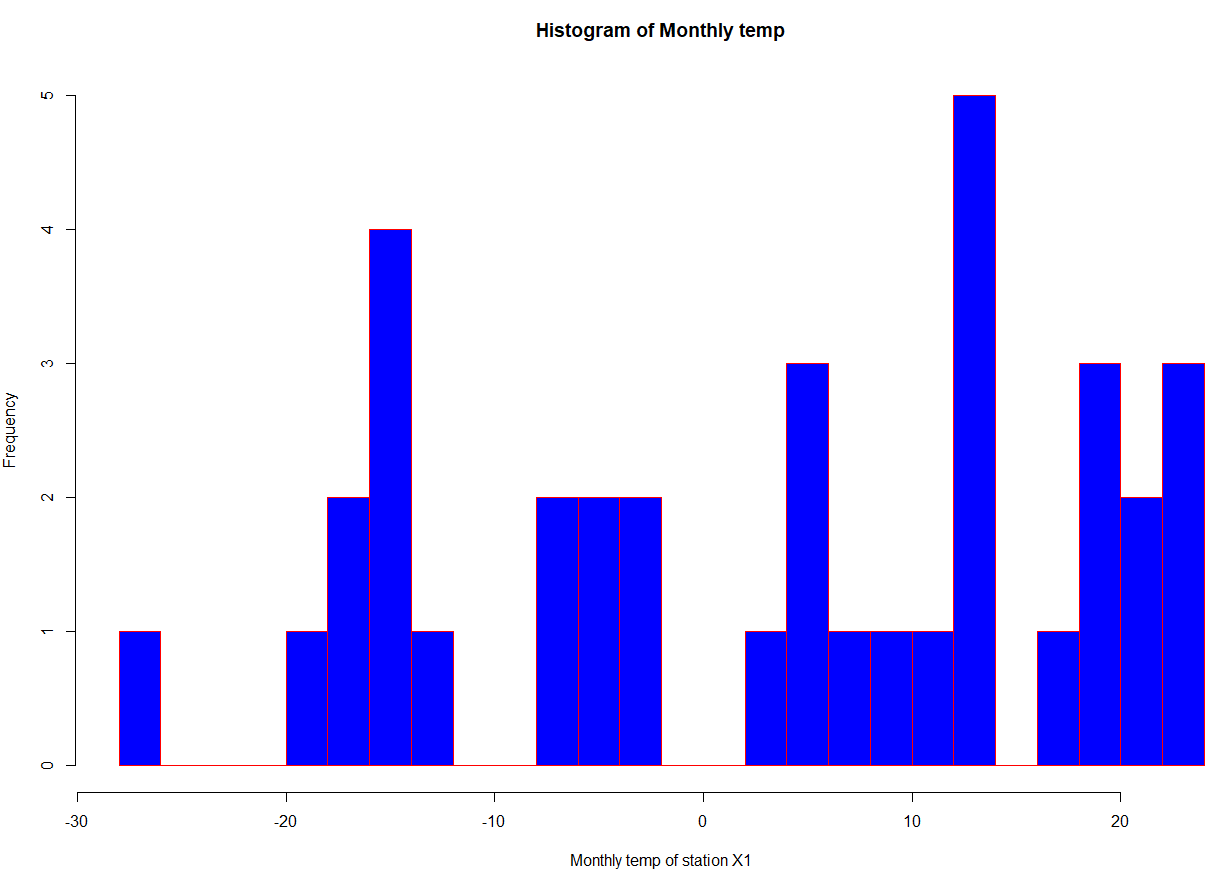


# 2.Time series

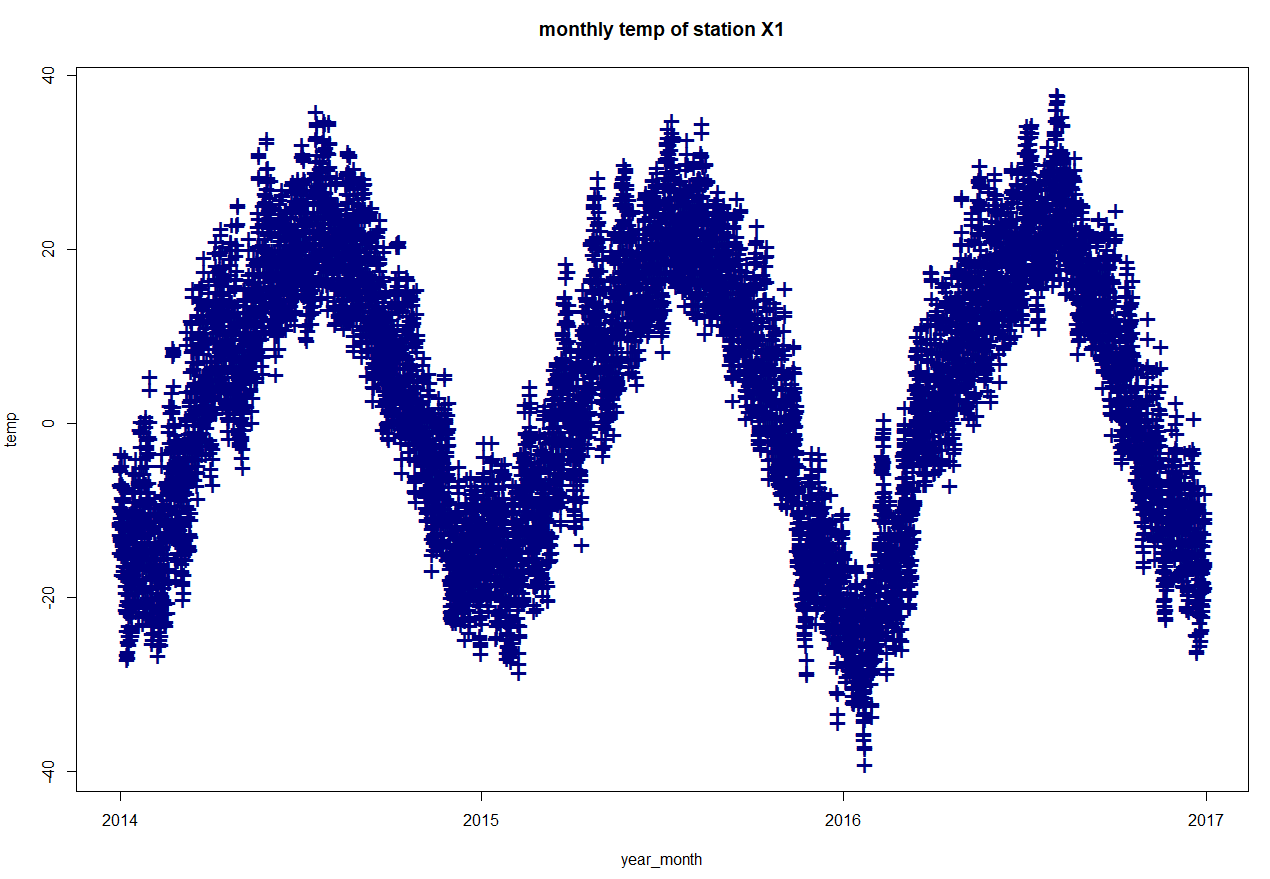




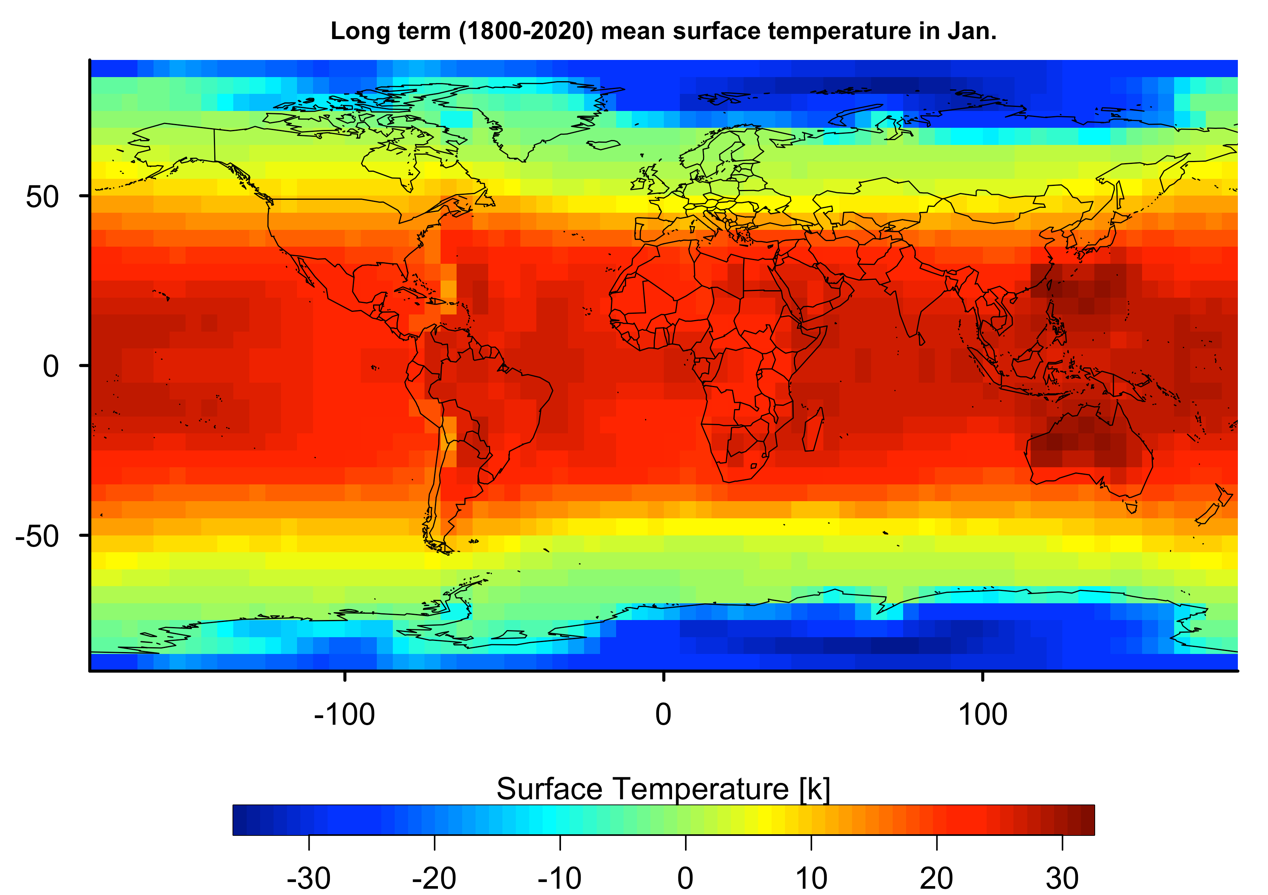
# 3 hisgram



# 4 Scatterplots

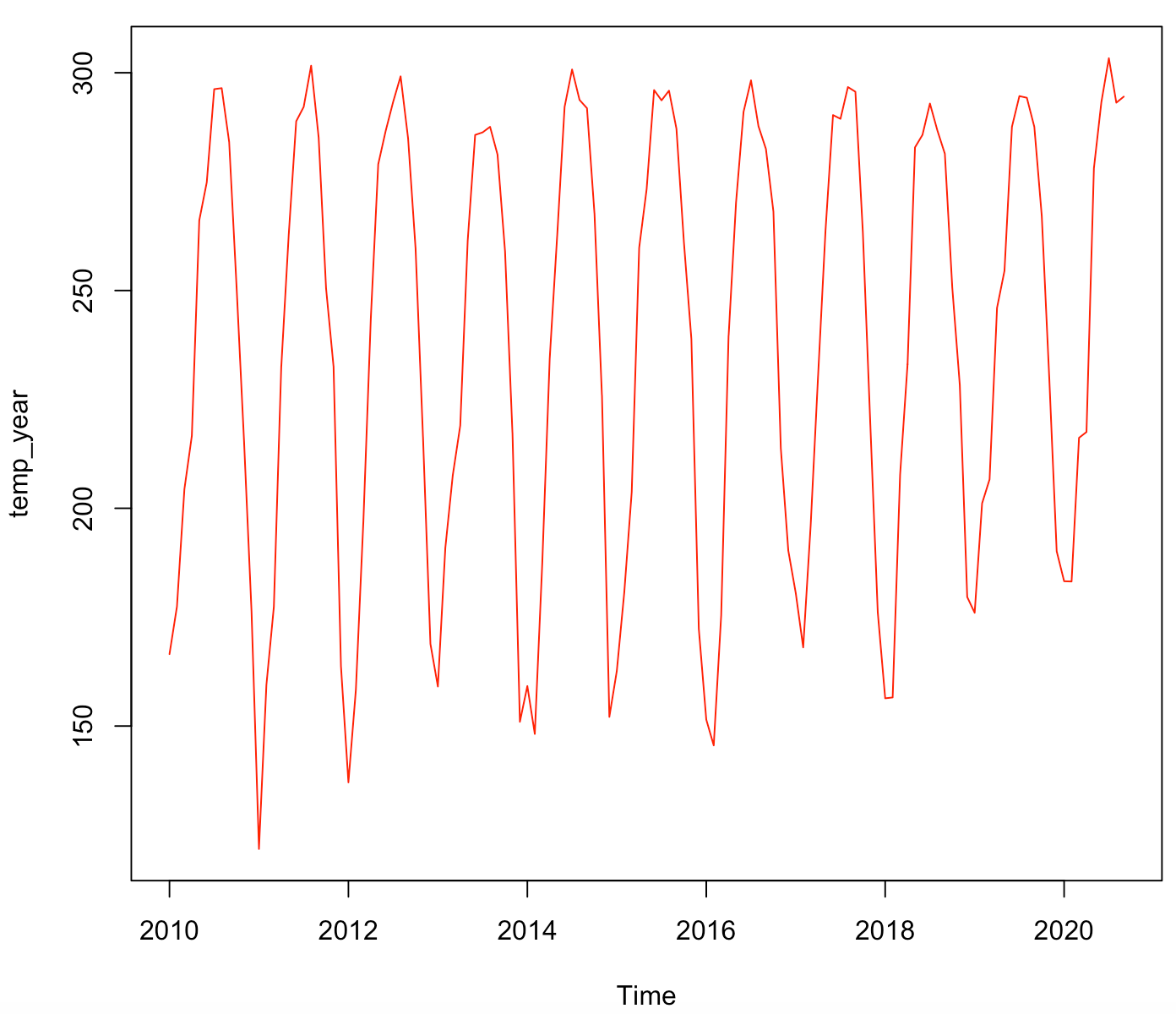


# 5 Image plot

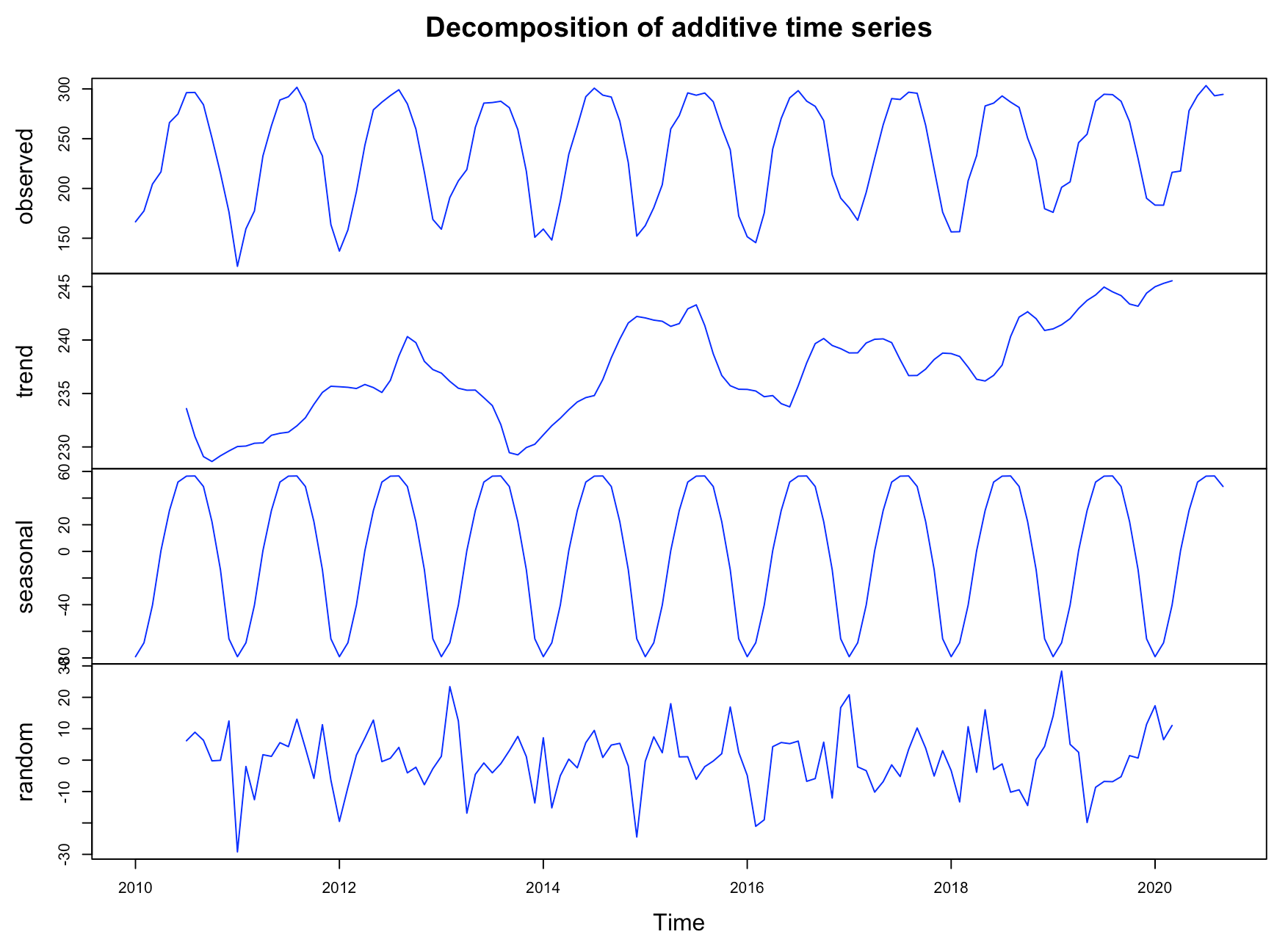


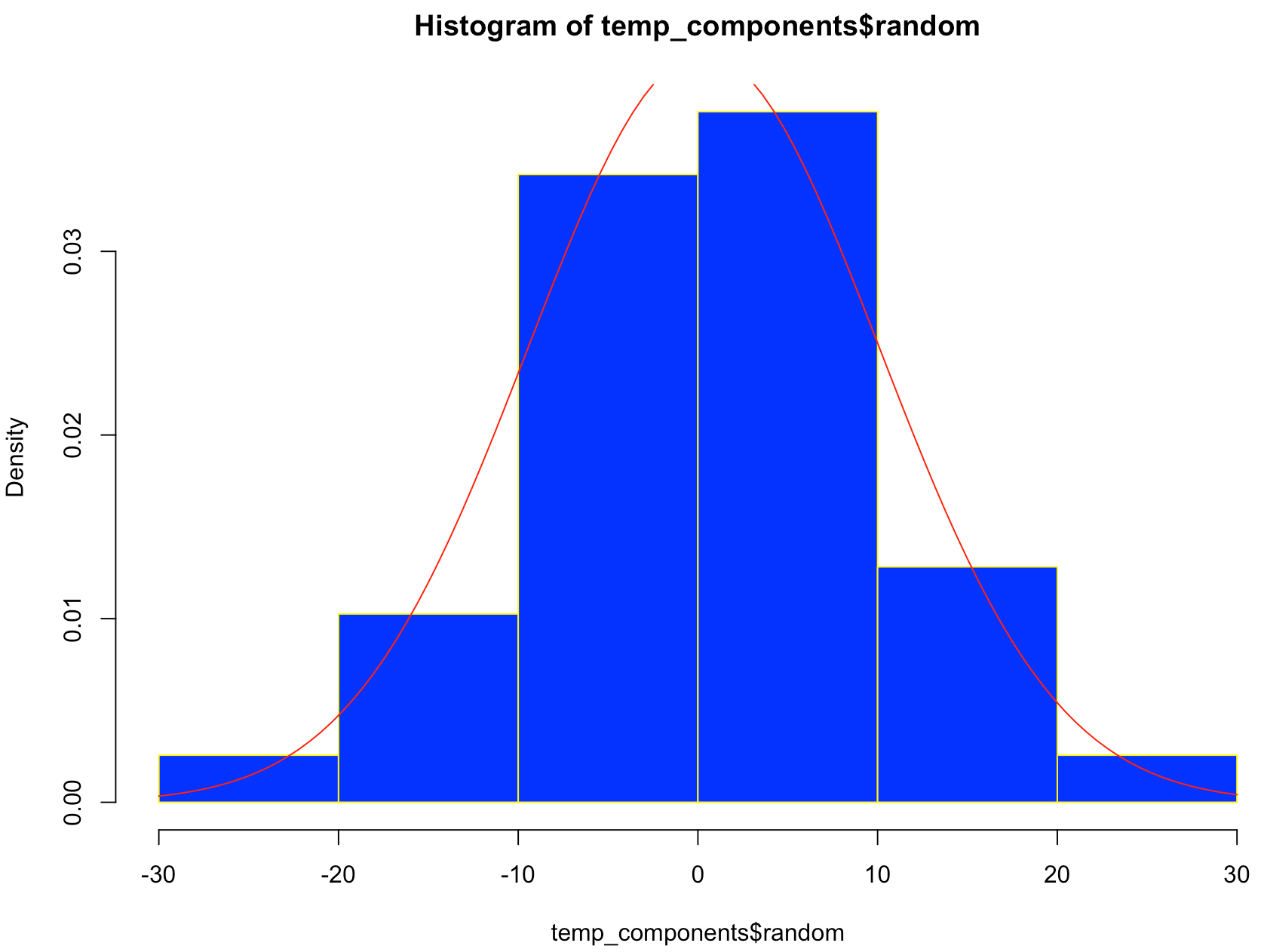
## PS4\_2

2.1 使用 substr获取DATE里面的数据以定义一个由年和月组成的变量year\_month，并且去除TMP中的无效数据。然后group\_by year\_month.再求取2010年一月至2020年八月的月度平均。



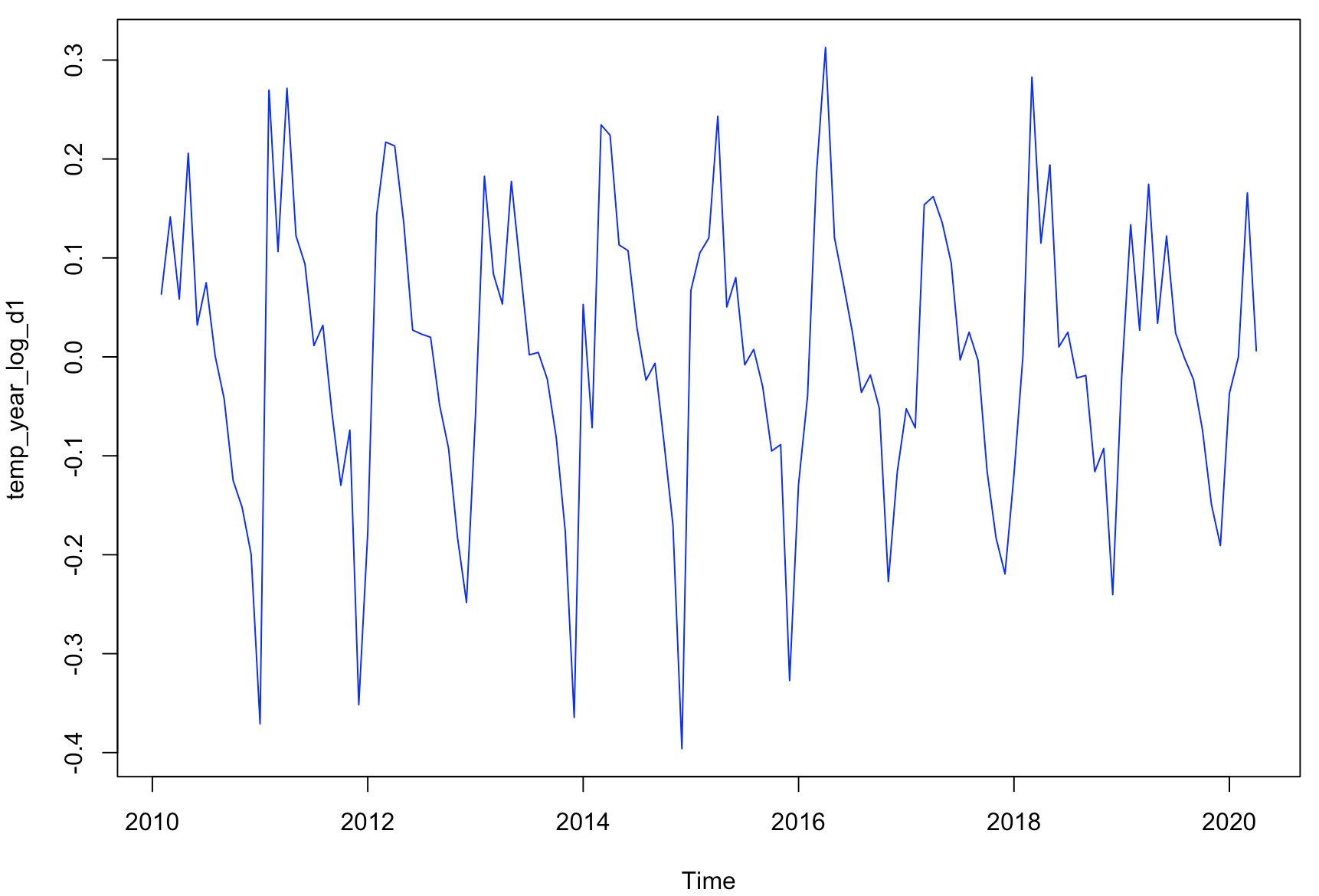
2.2 使用decompose对时间序列分解



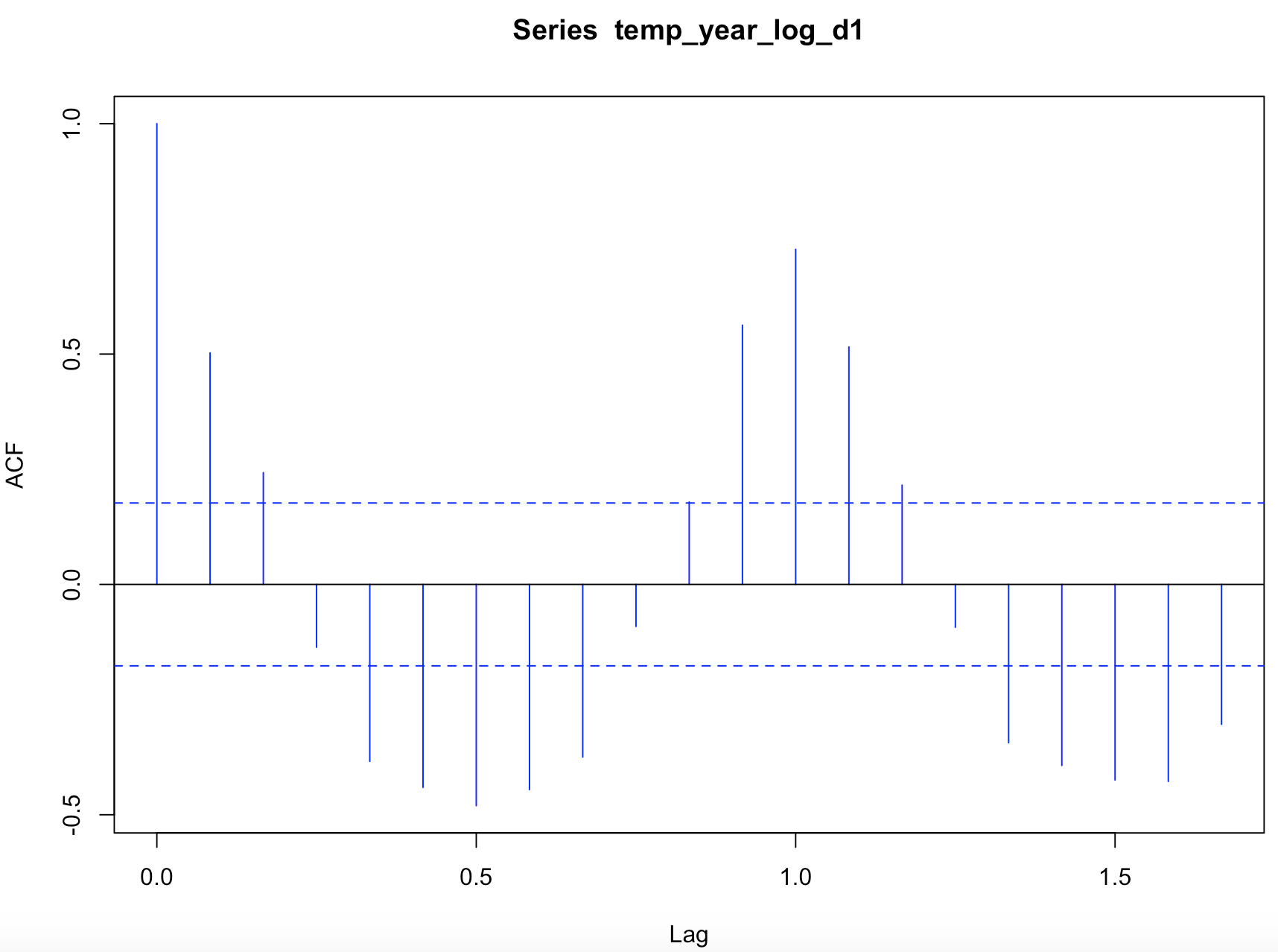


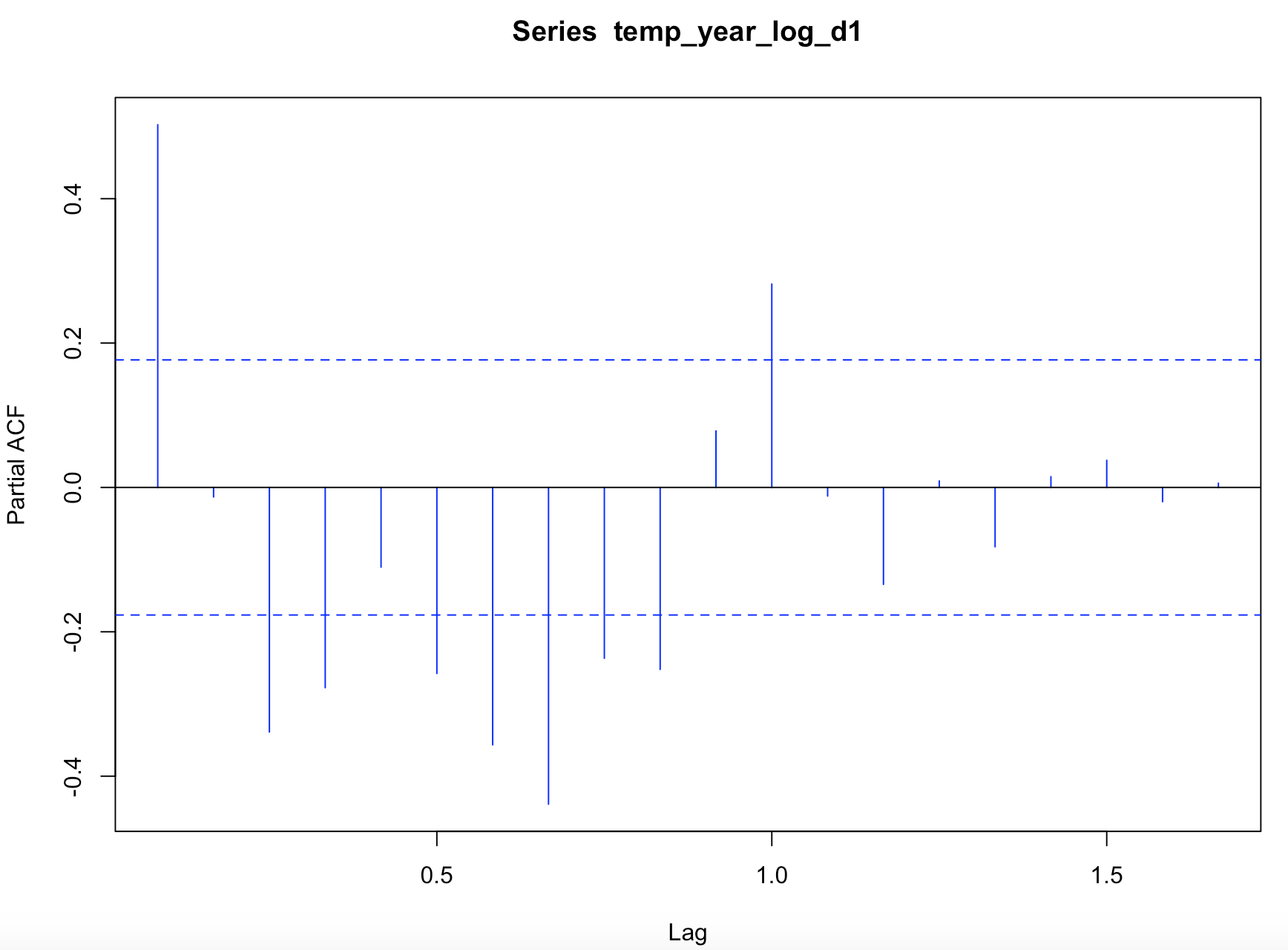
We can see that the error part follows a white noise distribution

2.3 使用ARIMA model 进行拟合，从时间序列分解图可以看出，该时间序列是有明显的趋势的，所以不能直接使用AMIRA model，先使用log进行变换，发现还是有明显趋势，然后再使用差分 diff对数据进行处理



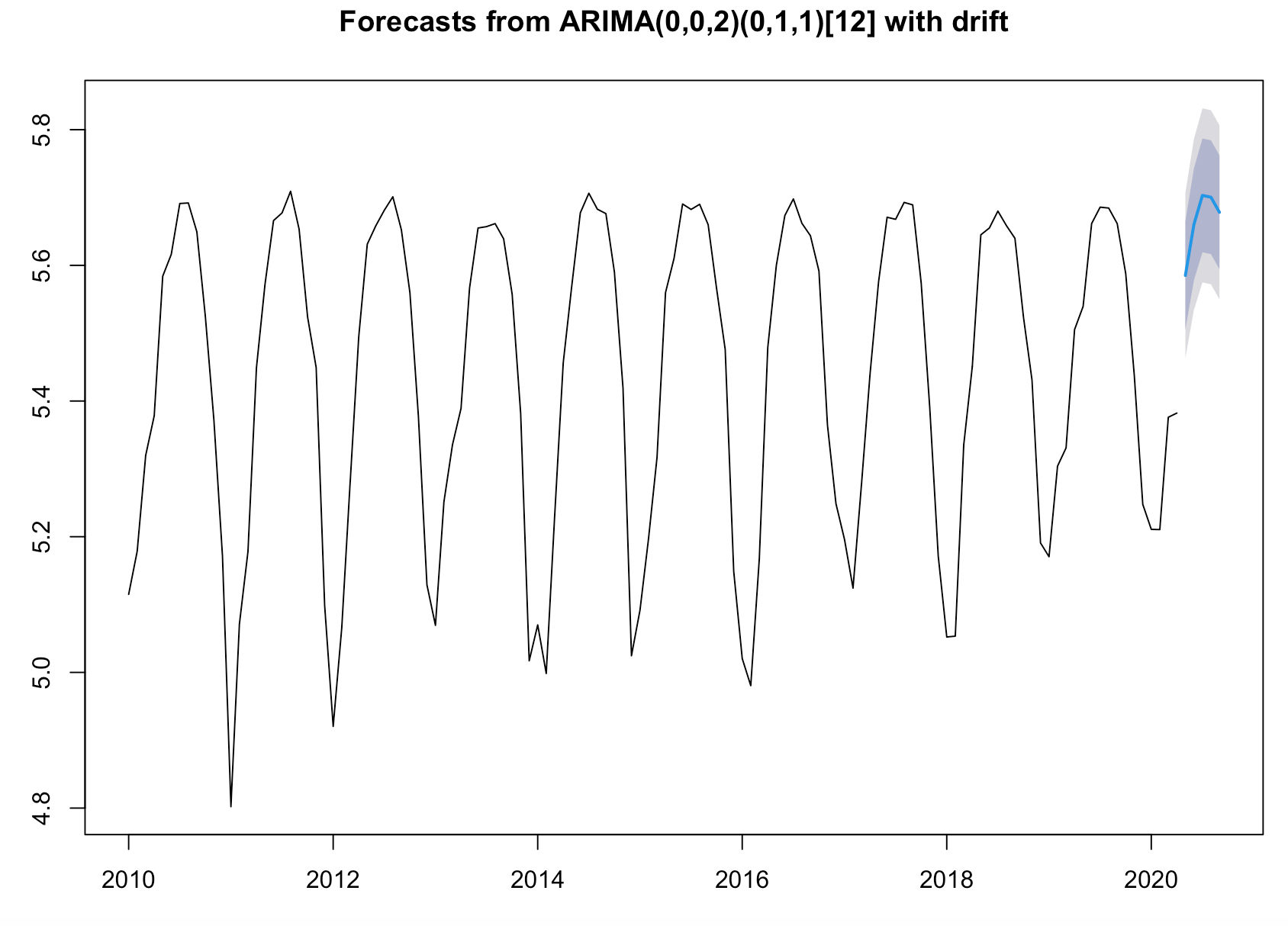
检查 acf 和pacf



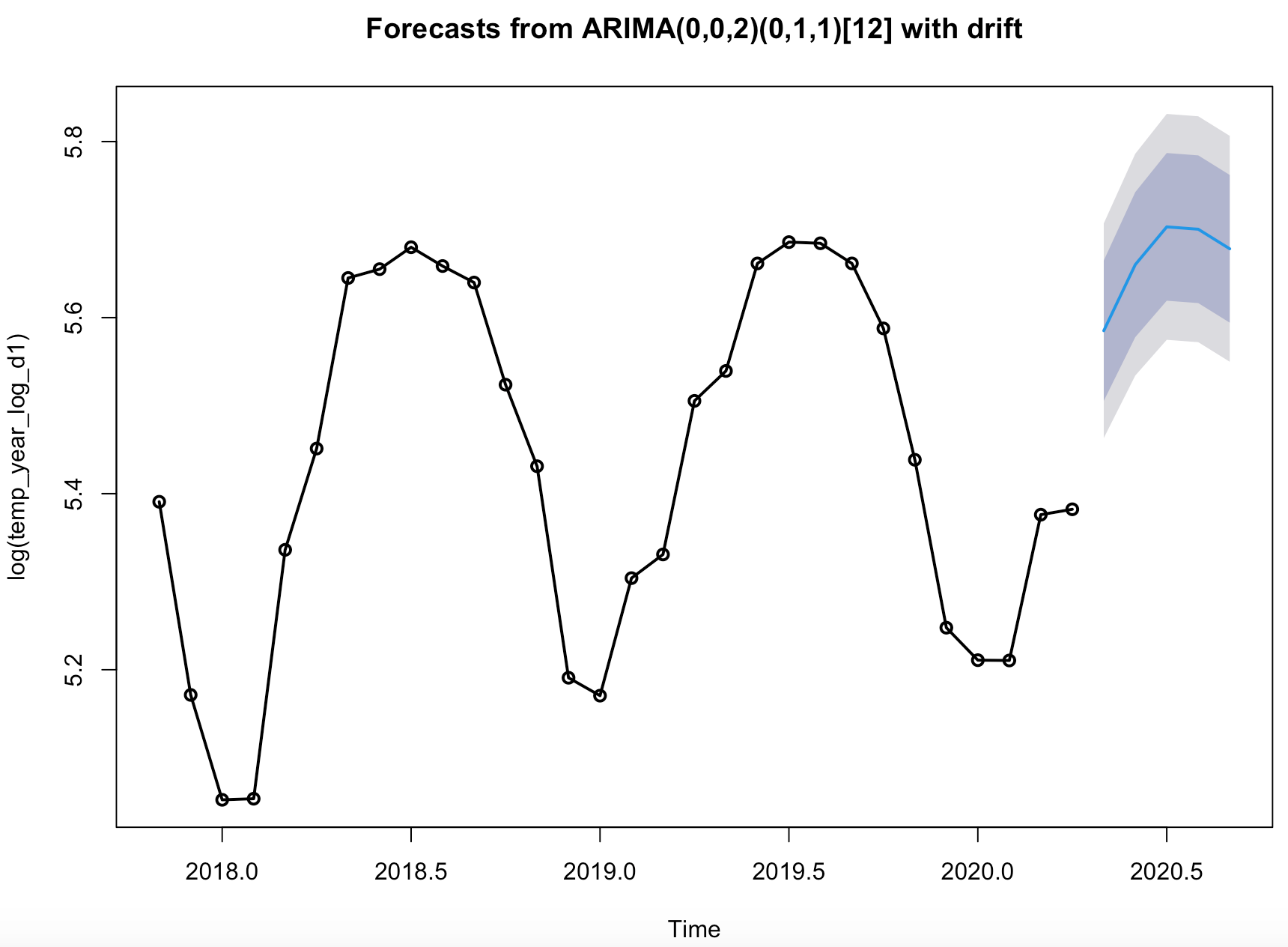


2.4 forecast

预测了5月-9月分的数据



与前30 个真实值一起plot出来



2.4 valutation

如下为预测值和真实值，然后对这五个月份的平均误差和每个月份的误差

#mean\_bias : 0.000226408

#每个月份的误差

# May Jun Jul Aug Sep

# 2020 0.042162860 0.020763919 0.011700228 -0.020018560 0.007128324

如下为：预测值和真实值

# predict values

# forecast\_5months

# Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

# May 2020 5.585117 5.505395 5.664839 5.463193 5.707041

# Jun 2020 5.660167 5.577796 5.742539 5.534191 5.786144

# Jul 2020 5.703099 5.619204 5.786993 5.574793 5.831404

# Aug 2020 5.700424 5.616530 5.784319 5.572118 5.828730

# Sep 2020 5.678195 5.594301 5.762090 5.549890 5.806501

# May Jun Jul Aug Sep

# 2020 266.4315 287.1967 299.7949 298.9942 292.4212

# real\_value

# 1 2020-04 218.

# 2 2020-05 278.

# 3 2020-06 293.

# 4 2020-07 303.

# 5 2020-08 293.

# 6 2020-09 295.