应用统计方法实验报告

太阳黑子数量的时序分析和预测

PB22050980 王开元

应用统计方法（015709.01）2025 春

中国科学技术大学

May 2025

1 实验概述

1.1 数据来源

数据来自 Kaggle, Daily Sunspots Dataset (1850 - 2025). 此数据集包含了 1850-01-01–2025-01-31 的太阳黑子数，是最新的关于每日太阳黑子计数的 Kaggle 数据集，没有缺失值。源数据来自比利时布鲁塞尔皇家天文台世界数据中心 SILSO。

1.2 检验与模型选择

利用 ADF 检验测试数据平稳性，利用 Ljung-Box 检验测试纯随机性，利用 SARIMA（季节性自回归移动平均）模型对数据拟合和预测。

2 分析与预测

2.1 数据处理与检验

数据集为日度数据，数据量大（超过 60000 个数据点），波动性高，噪声大且周期短，容易导致噪声干扰趋势，影响模型的解释性，且计算效率低，耗时长。对于太阳活动周期这类天文学问题，我们更关系的是识别长期趋势。因而我们将数据处理为**月平均太阳黑子数**进行建模。

(1) 首先检验数据的平稳性，使用 ADF 检验，得到结果:

```
        Augmented Dickey-Fuller Test

        data:  sunspot_ts
        Dickey-Fuller = -5.3175, Lag order = 12, p-value < 0.01
        alternative hypothesis: stationary
```

Listing 1: ADF 检验结果

$p < 0.05$, 支持拒绝序列非平稳的原假设，说明序列平稳，可以直接拟合。

(2) 计算自相关系数（ACF），偏自相关系数（PACF）。可以观察到:

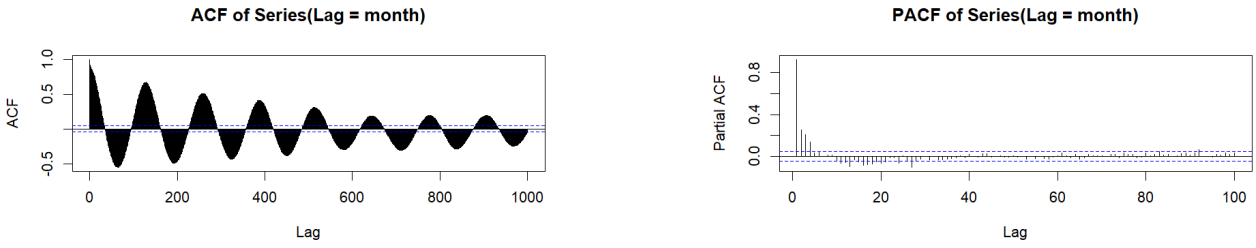

图 1: ACF(左),PACF(右)

a. ACF 有显著的滞后，这表明序列并非白噪音。

b. 在特定间隔有周期性的峰值，周期约为 133 月，这表明数据具有显著的季节性，且季节周期长。

c. ACF 逐渐衰减，PACF 在某一滞后截断，这表明数据有显著的 AR 特性。

对于 ACF 和 PACF 的描述性观察，我们可以大致确定应该选择的模型，即 SARIMA 。

(3) 数据的时序图 直接观察时序图，我们也可以发现太阳黑子数的明显的周期性（约为 11 年 =132 月）。

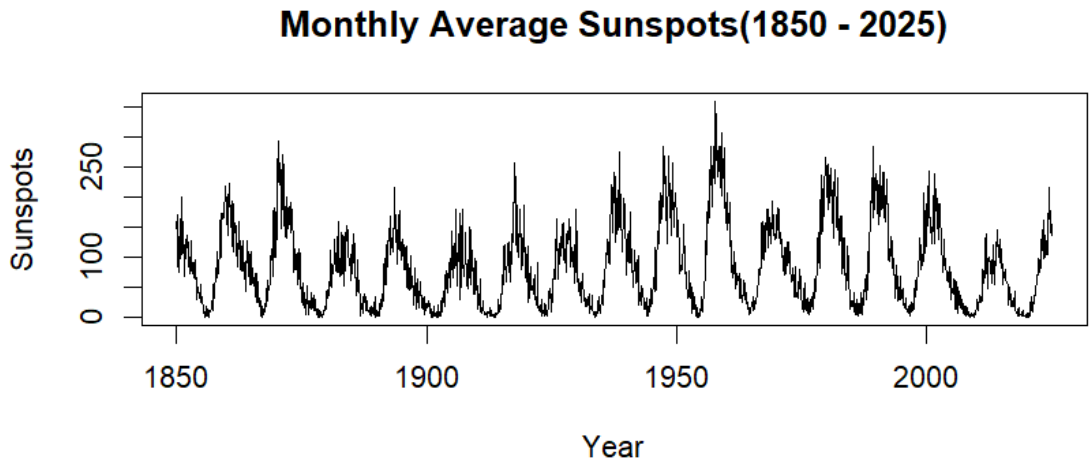

图 2: 太阳黑子月平均数的时序图

2.2 自动拟合 ARIMA 模型

在下面，我们将处理后的数据集（1850-2025 月平均太阳黑子数）分为“训练集”（1850-01–1999-12）和“预测集”（2000-01–2025-01）两部分，利用训练集拟合 SARIMA 模型，并在未来的 25 年进行预测，与预测集的真实数据对比来反应模型的解释性。

```
#数据分段为训练集和预测集
train_ts <- window(sunspot_ts, end = c(1999, 12))
test_ts <- window(sunspot_ts, start = c(2000, 1), end = c(2025, 1))
```

Listing 2: 数据分段

方便起见，我们直接利用 R 内置的函数 *auto.arima()* 拟合模型；然后使用 *forecast()* 函数基于拟合好的时间序列模型进行未来值预测。得到模型和预测如下：

```
        Series: train_ts
ARIMA(2,1,2)(2,0,0)[12]

Coefficients:
          ar1      ar2      ma1     ma2    sar1     sar2
       1.3549  -0.3943  -1.7752  0.8088  0.0306  -0.0163
s.e.   0.0400   0.0370   0.0277  0.0258  0.0252   0.0243

sigma^2 = 612.4:  log likelihood = -8322.48
AIC=16658.97   AICc=16659.03   BIC=16697.43
```

Listing 3: 自动拟合 ARIMA 模型

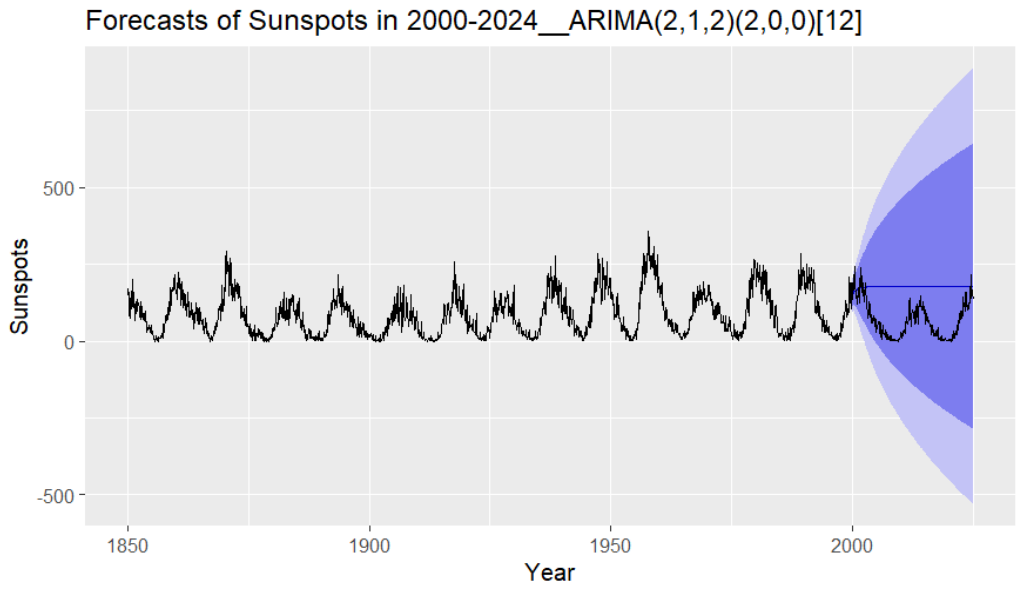

图 3: ARIMA(212)(200)[12] 预测图

显然，*auto.arima* 并没有发现时序数据具有的长周期，且做出了糟糕的预测。为此，我们手动调整参数，寻找最佳的预测模型。

2.3 调节参数

2.3.1 第一次调整

直接设置季节周期为 132，在自动拟合的模型 ARIMA(2,1,2)(2,0,0)[12] 上做调整。因为已经确定了近似的周期，以及观察到模型具有较强的季节性趋势，我们令季节项差分次数 $D = 1$，周期为 132。得到模型和预测结果：

```
Series: train_ts
ARIMA(2,1,2)(2,1,0)[132]

Coefficients:
          ar1      ar2      ma1     ma2     sar1     sar2
      1.1017  -0.2763  -1.5730  0.6425  -0.6400  -0.3080
s.e.  0.0754   0.0364   0.0797  0.0612   0.0239   0.0239

sigma^2 = 671.1:  log likelihood = -7932.5
```

Listing 4: 模型]ARIMA(212)(210)[132] 模型

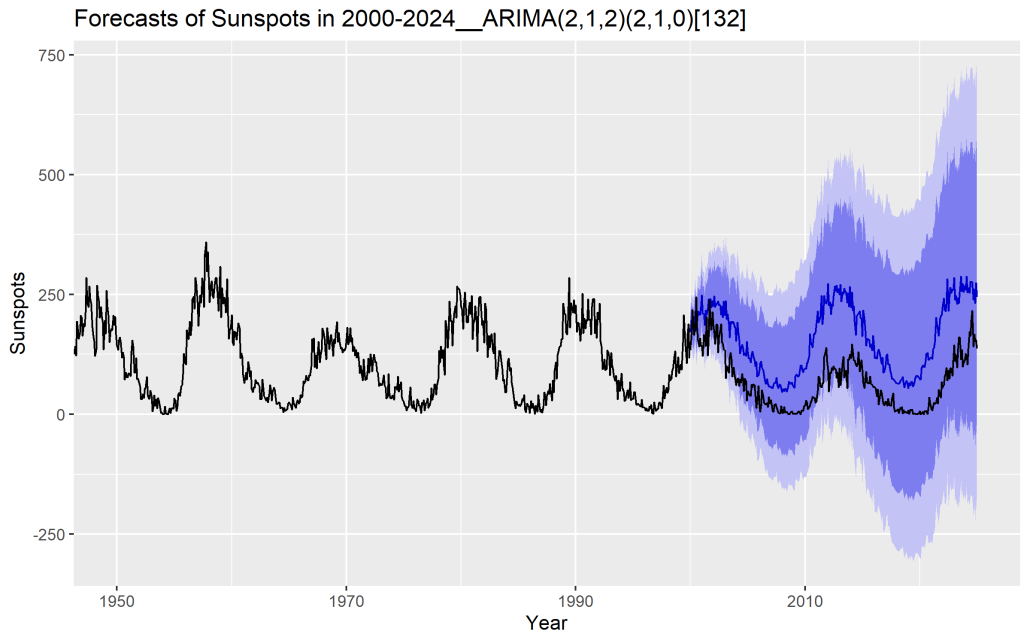

图 4: ARIMA(212)(210)[132] 预测图

我们观察到，调整过的模型已经能够大致识别数据的周期和趋势。但我们还面临准确性不高、置信区间过宽的问题。需要进一步调整。

2.3.2 第二次调整

在先前的检验中，我们已经知道数据是平稳的，由此可以猜测，或许将非季节项的差分调整为 0 更能反应数据的特征，为此我们尝试 ARIMA(2,0,2)(2,1,0)[132]:

```
Series: train_ts
ARIMA(2,0,2)(2,1,0)[132]

Coefficients:
          ar1     ar2      ma1      ma2     sar1     sar2
      0.8989  0.0712  -0.3652  -0.2016  -0.6262  -0.3044
s.e.  0.1092  0.1053   0.1064   0.0581   0.0240   0.0239

sigma^2 = 664.4:  log likelihood = -7928.72
```

Listing 5: 模型]ARIMA(2,0,2)(2,1,0)[132] 模型

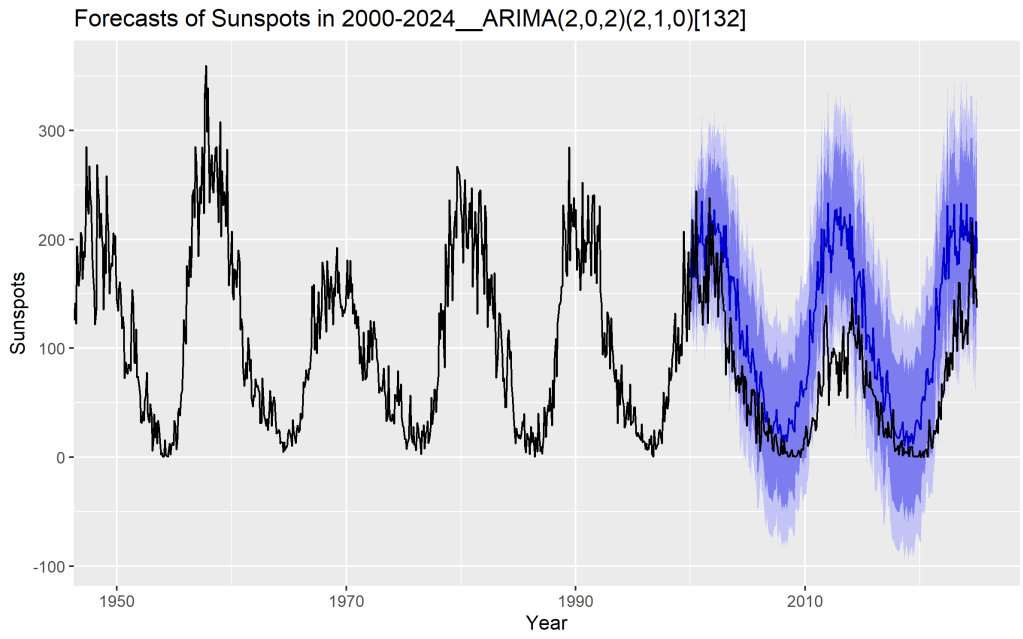

图 5: ARIMA(2,0,2)(2,1,0)[132] 预测图

可以看到，在 2000-2025 年间，模型已经可以比较准确地把握数据的趋势，尤其是对于周期性的拟合程度很好，尤其是在前十年（2000-2010）期间。这表明我们的模型 ARIMA(2,0,2)(2,1,0)[132] 在短期内具有很好的解释性，在更长的时间内也能预测趋势。

另一方面，可以发现预测的置信区间被控制在了 $[-100, 300]$ 之内，这说明，预测的波动小，可信度高，进一步说明了此模型具有好的解释性。

2.4 残差检验

为了进一步检验模型的性质，我们通过 *checkresiduals()* 函数检验模型的残差，残差图和 Ljung-Box test 如下：

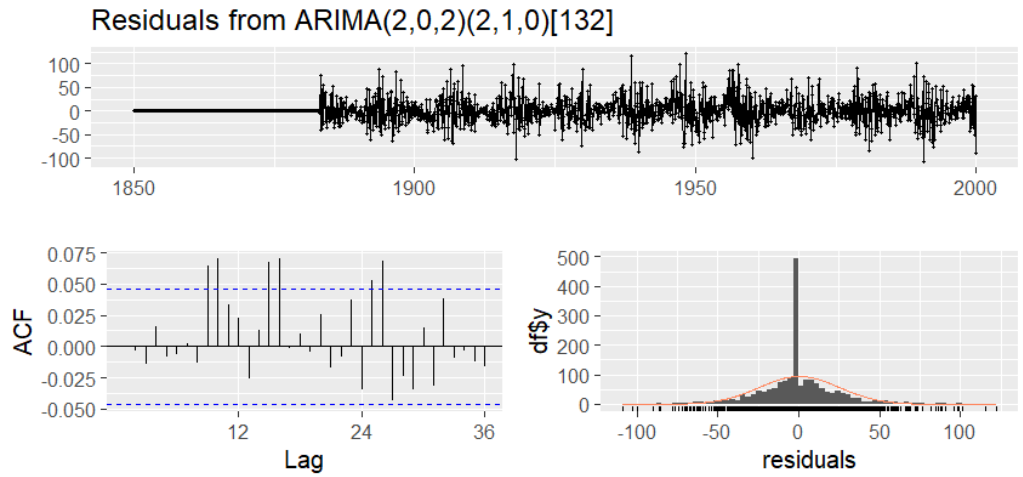

图 6: ARIMA(2,0,2)(2,1,0)[132] 残差时序图、残差自相关图和残差分布图

Ljung–Box test

```
data:  Residuals from ARIMA(2,0,2)(2,1,0)[132]
Q* = 45.887, df = 18, p-value = 0.0003082

Model df: 6.   Total lags used: 24
```

Listing 6: 模型]ARIMA(2,0,2)(2,1,0)[132] 模型

1. 可以观察到残差时序图无明显趋势，说明模型稳定，基本捕捉了数据的信息。
2. 残差自相关系数大部分分布在置信区间（蓝线）以内，少数点稍微越界，但幅度小、模式随机，没有明显周期性。
3. 残差大致服从正态分布，但在中心出现尖峰（待解释）。
4. 在 Ljung-Box test 中，$p-value = 0.0003082$，相比于 ARIMA(2,1,2)(2,1,0)[132] 的 $p-value = 0.004219$ 略小，而 p 值越小越拒绝残差是白噪声。在这个意义上，先前的 ARIMA(2,1,2)(2,1,0)[132] 模型效果更好，这说明我们需要综合多种检验来选择最合适的模型。

3 太阳黑子数和降雨量的相关性分析

根据 [1] 等文献的研究，我们了解到太阳黑子数和地区降水量可能存在联系。[1] 研究了印度 Kerala 地区的降水量和太阳黑子数的关系，得出结论**太阳黑子数量和降雨量之间存在明显的 8 至 12 年的相关性**。

因为太阳活动与季风的相关性较强，所以若要反应太阳活动对降水的影响，研究印度地区会是一个好的选择。以下我们分析年均太阳黑子数和印度孟买 (Mumbai) 地区年均降水量的相关性，数据来源 Kaggle, Mumbai Rainfall Data. 其中包含了 1901-2021 年孟买地区的月降水量。

以下是 1901-2021 年 Mumbai 年均降水量的时序图（左）。标准化后和年均太阳黑子数比较（右），直观上看，没有发现显著的相关性。

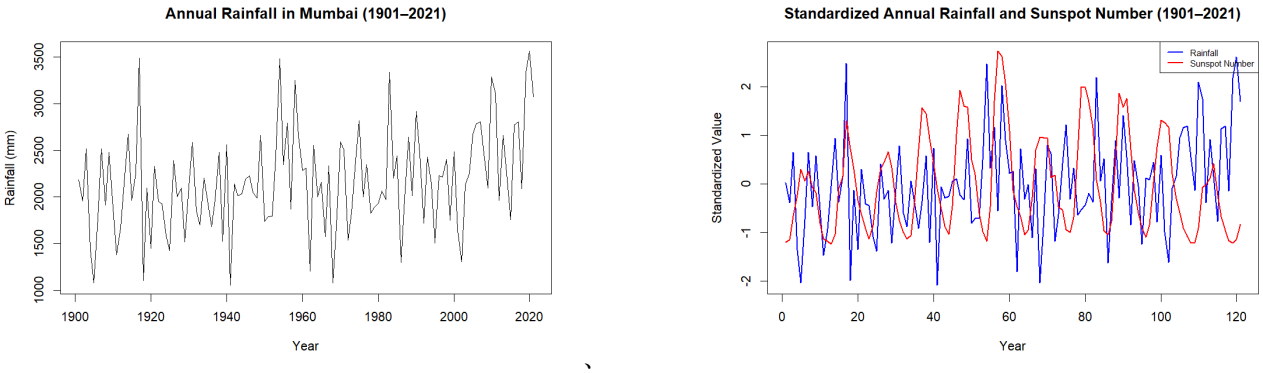

图 7: Mumbai 年均降水量 (左), 标准化的 Mumbai 年均降水量和年均太阳黑子数 (右)

直接做 Pearson 相关性检验，得到 $\rho = -0.04770362$，说明相关性弱。

```
Pearson's product-moment correlation

data:  rain_trim and sunspot_trim
t = -0.52098, df = 119, p-value = 0.6034
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 −0.2242903  0.1319162
sample estimates:
cor
−0.04770362
```

Listing 7: Pearson 相关性检验

可能太阳黑子数和降水量不存在直接到相关性，下面考虑滞后相关性。因为样本量为 120 ，所以我们考虑 30 年内的滞后相关性 $(-30 < lag < 30)$。如图所示：

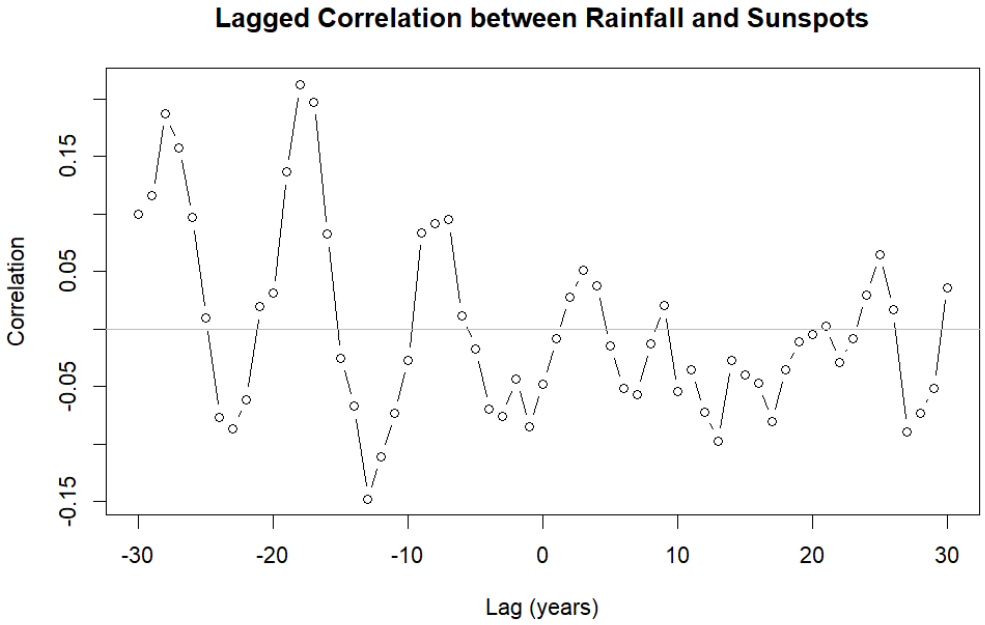

图 8: 滞后 30 年内太阳黑子数和降雨量的相关系数

发现在 $lag = -18$ 处取到峰值，约为 0.2。这意味着**太阳黑子数变化对降水量的影响具有一定的周期性延迟响应**。尽管相关系数不大，但仍然说明这不是纯随机的噪声，而可能存在某种弱但系统性的联系。

下面对 $lag = -18$ 的情况做相关性检验，结果如下：

```
Pearson's product-moment correlation

data:  rain_lagged and sunspot_lagged
t = 2.1792, df = 101, p-value = 0.03164
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0191740 0.3894649
sample estimates:
      cor
0.2119123
```

Listing 8: Pearson 相关性检验（lag = 18）

分析新的结果：

1. Pearson 相关系数 $cor = 0.2119123$，表示太阳黑子数与降雨量呈轻度正相关。

2. p 值 $0.03164 < 0.05$，表面相关性显著。且 95% 置信区间不包含 0。

延迟后的太阳黑子数和降水量对比图如下（标准化后）：

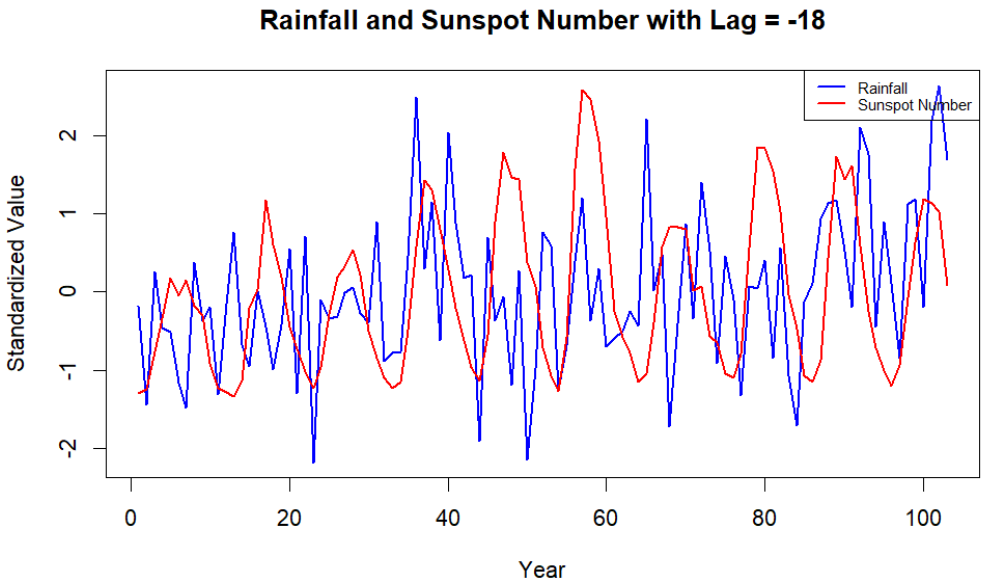

图 9: 滞后 18 年的年均降水量和年均太阳黑子数对比图

4 总结

我们分析了 1985-2025 年的太阳黑子数时序数据，发现其存在显著的周期性，并用 ARIMA 模型对数据做了拟合和预测，得到了具有较好解释性的模型。另一方面，我们对太阳黑子数和孟买地区的降水量进行了相关性分析，发现在太阳黑子数领先 18 年时，相比于其他延迟，太阳黑子数和降水量存在最强的正相关性。尽管相关系数较小，也能在一定程度上反应太阳黑子数对降水量的延迟性影响，为进一步的猜想和研究提供了依据。

参考文献

[1] Elizabeth Thomas and Noble P. Abraham, “Relationship between sunspot number and seasonal rainfall over Kerala using wavelet analysis,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 240, pp. 105943, 2022.