Ziyan Lin

Professor Chan

STAT 6560 Applied Time Series Analysis

05 May 2018

Time Series Analysis on Water Usage

1. Introduction

As a finite natural resource, fresh water accounts for 0.003% of total water available globally. Many areas of the world are already experiencing stress on water availability. Due to the accelerated pace of population growth and an increase in the amount of water a single person uses, it is expected that this situation will continue to get worse. A shortage of water in the future would be detrimental to the human population as it would affect everything from sanitation, to overall health and the production of grain. Therefore, this project focuses on analyzing water usage. Specifically, the purpose is to find the relationship between time and monthly water usage. The series is obtained from the website: https://datamarket.com/data/set/22qu/monthly-water-usage-mlday-london-ontario-1966-1988#!ds=22qu&display=line. It records monthly water usage (mL/day) in London from January 1966 to December 1988. This is a 23-year period, and total 276 observations have been collected.

In this project, the main statistical software is R and its package “TSA” to deal with the series.

2. Model Specification

2.1 The time series plot of monthly water usage level from January 1966 through December 1988

Fig 1 Monthly water usage in London from 1966 to 1988

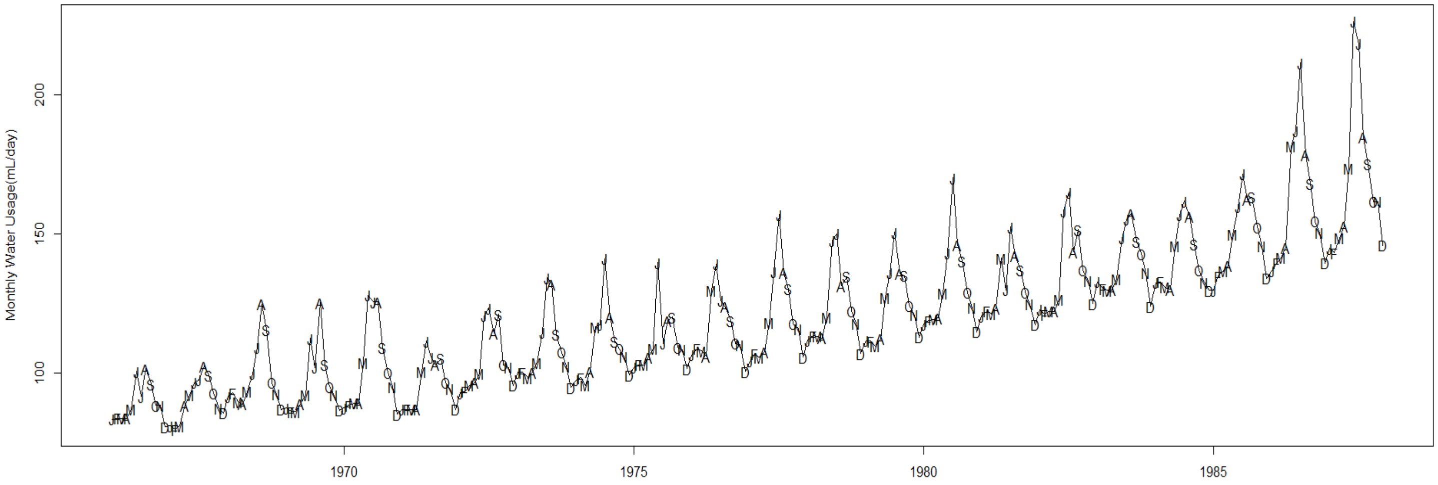
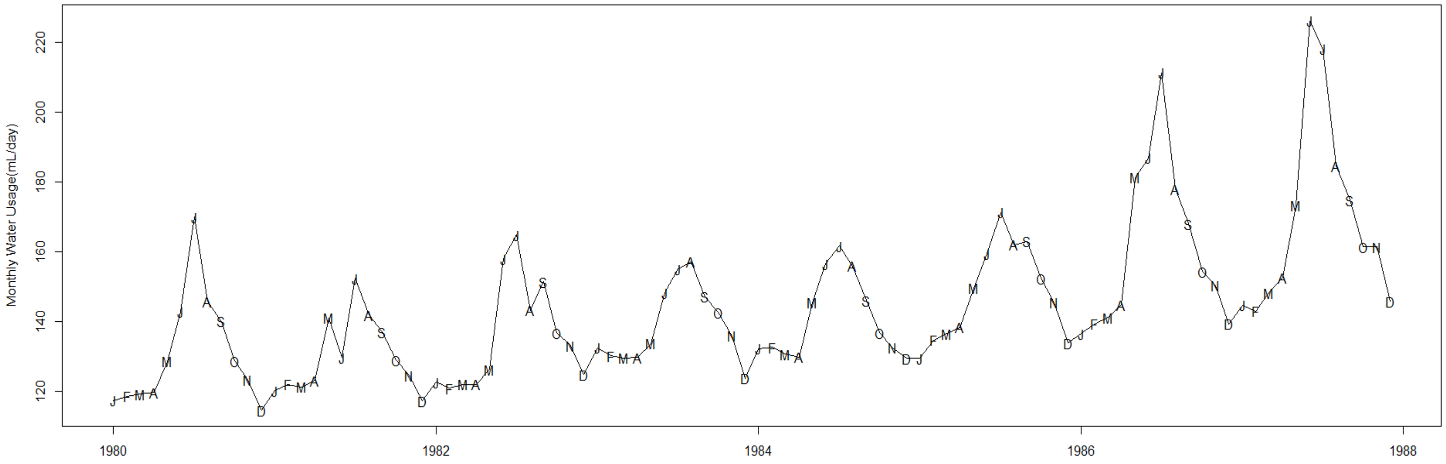


Fig 1 displays the time series plot of monthly water usage level from January 1966 through December 1988 with symbols. There is a strong upward trend but also a seasonality that can be seen better in the mode detailed in Fig 2, which only displays monthly water usage from January 1980 to December 1988.

2.2 The time series plot of monthly water usage level from January 1980 through December 1988

Fig 2 Monthly water usage in London from 1980 to 1988



From Fig 2, the general water usage levels are higher during the summer months and much lower in the winter. Based on the time series pattern, using stochastic seasonal models for this series.

2.3 Power transformation

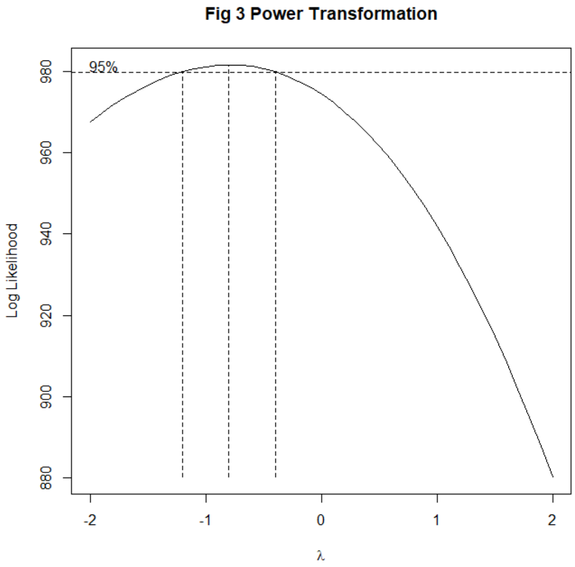


Fig 3 shows that the series needs power transformation. The 95% confidence interval contains 1. Reciprocal transformation may be a good choice.

2.4 Sample autocorrelation function for transformed series

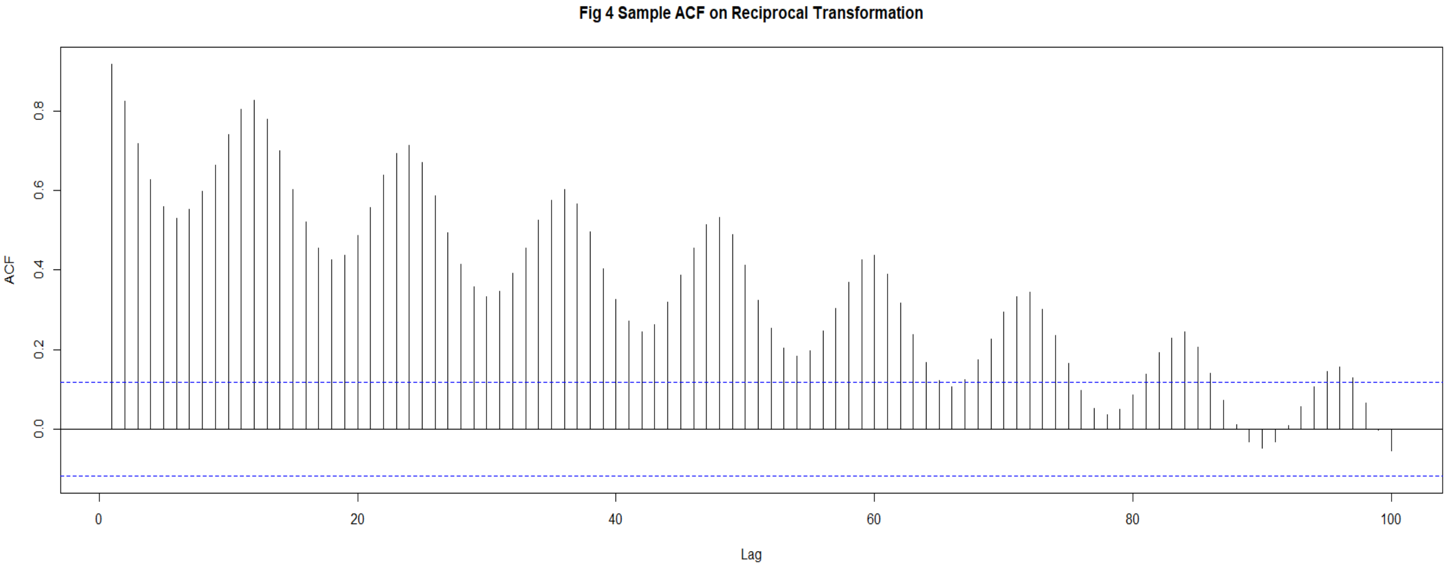


Fig 4 shows the sample autocorrelation function for the series. The seasonal autocorrelation relationships are shown in this plot. And there are strong correlations at lags 12, 24, 36, and so on. Moreover, there is other correlation that needs to be modeled.

2.5 The time Series plot of the first differences of monthly water usage

Fig 5 Time Series Plot of the First Differences of Monthly Water Usage

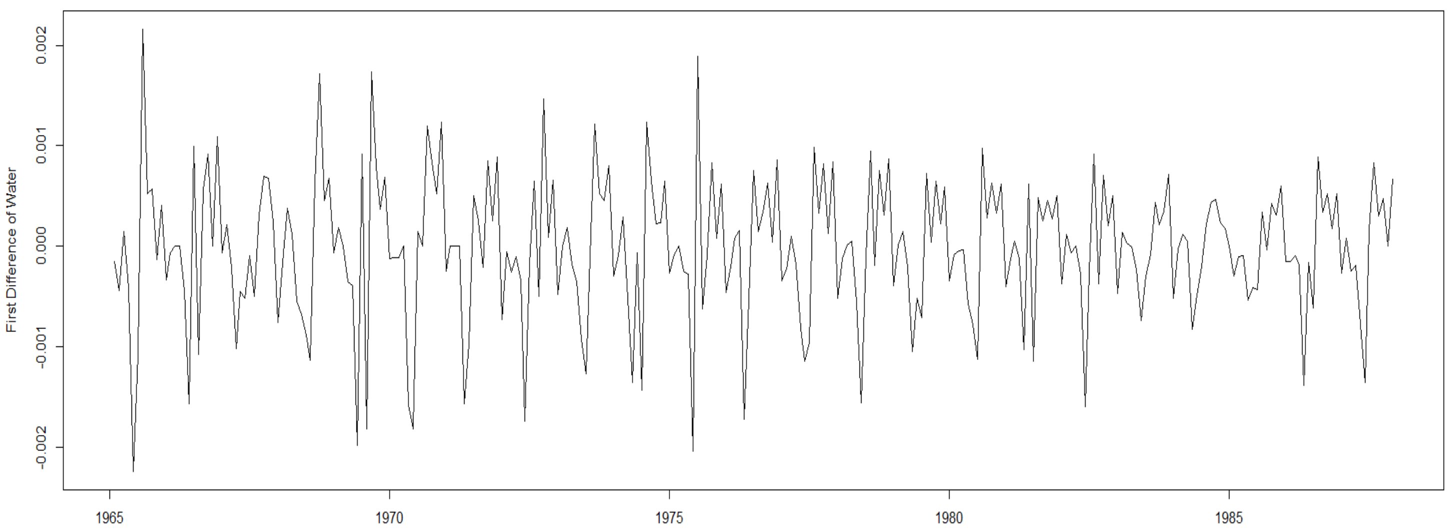


Fig 5 shows the time series plot of the monthly water usage after taking a first difference. The general upward trend has now disappeared but the strong seasonality is still present.

2.6 Sample ACF of the first difference

Fig 6 Sample ACF of First Differences of Monthly Water Usage

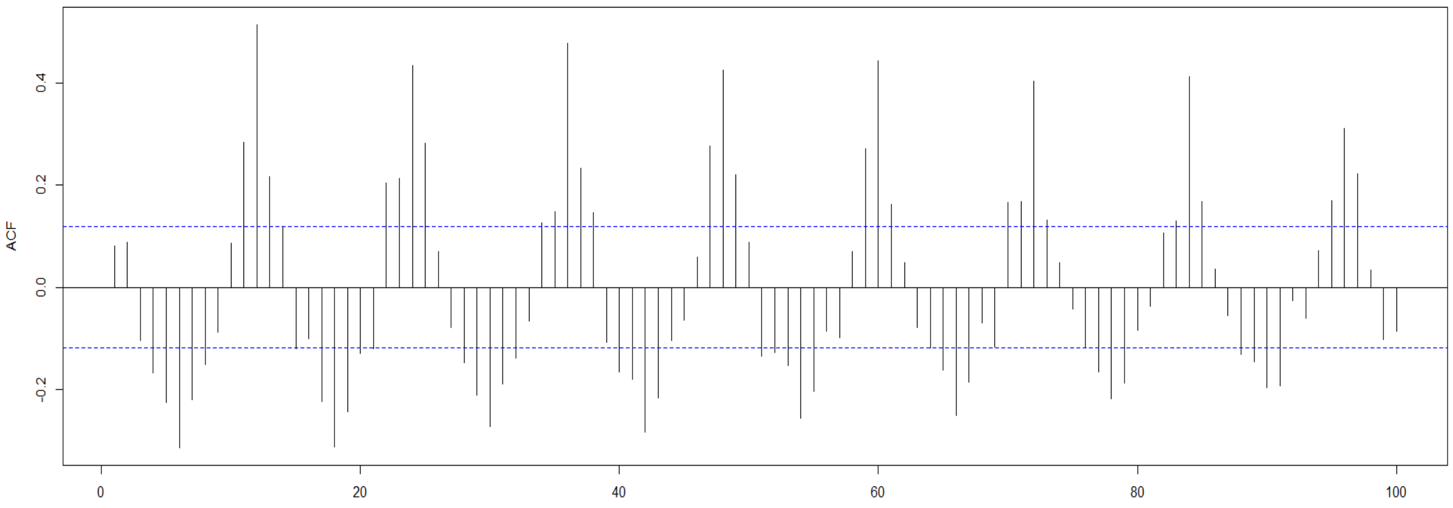


Fig 6 shows the Sample ACF of the first difference. It is obvious that there is a strong seasonality.

2.7 The time series plot of first and seasonal differences of monthly water usage

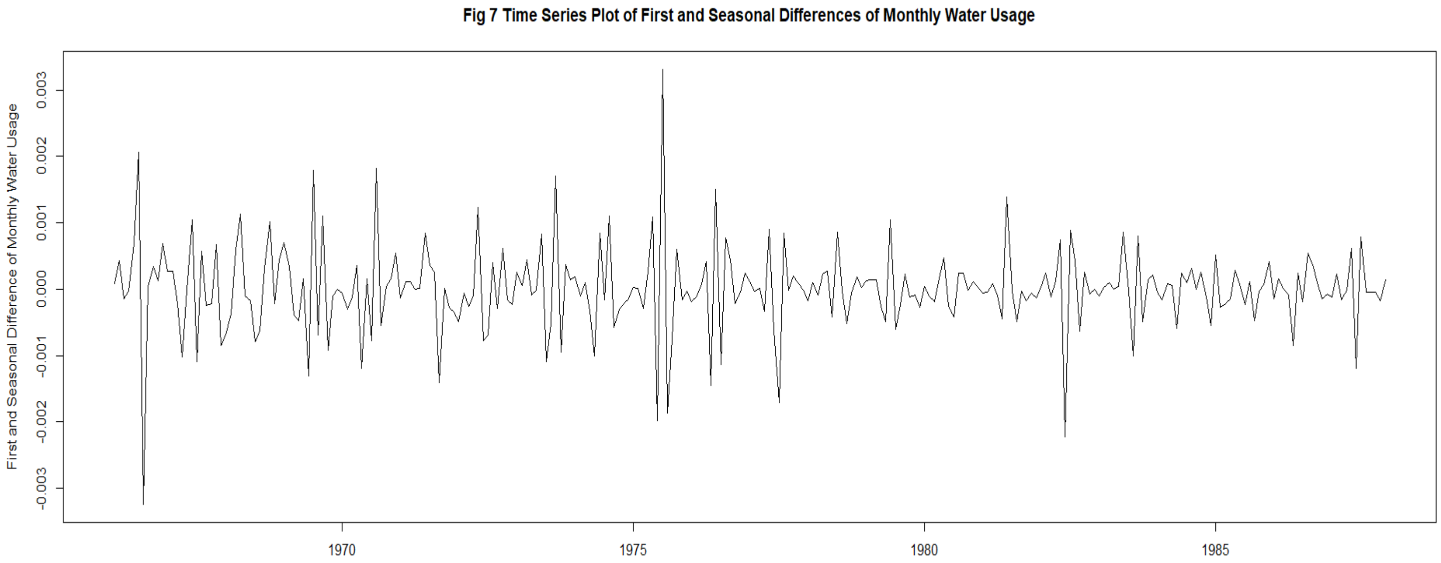


Fig 7 shows the time series plot of the monthly water usage after taking both a first difference and a seasonal difference. The most of the seasonality disappears now.

2.8 Sample ACF of first and seasonal difference of monthly water usage

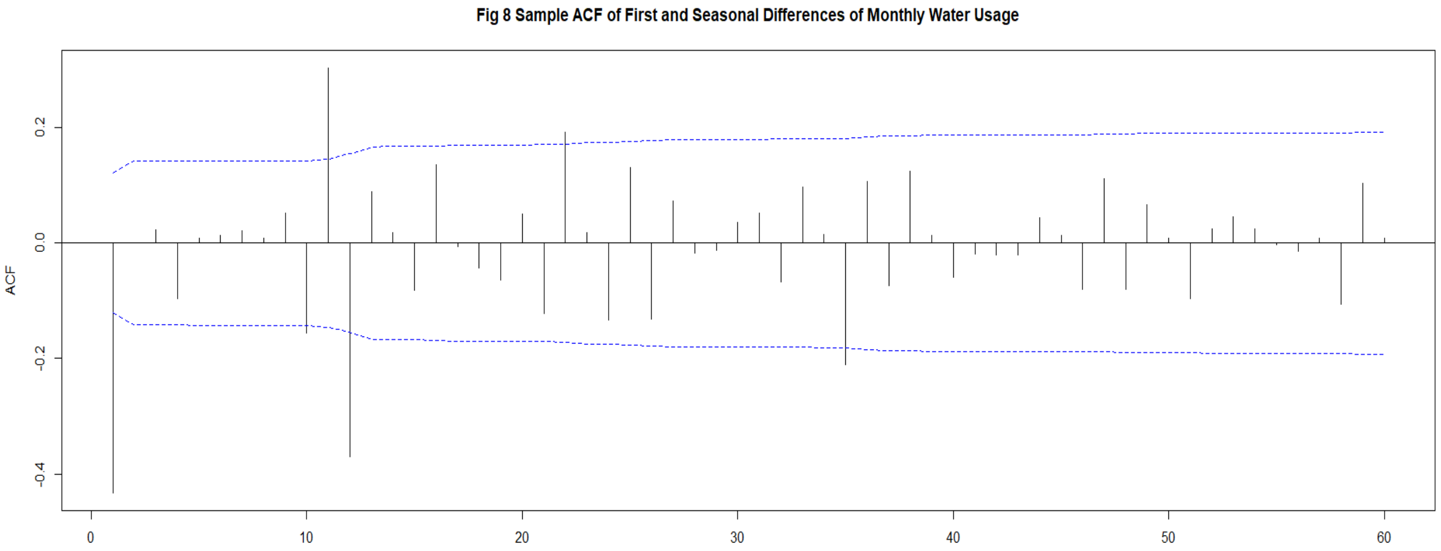


Fig 8 confirms that some autocorrelation remains in the series after taking these two difference.

3. Model Fitting

3.1 The ARIMA (0, 1, 1) × (0, 1, 1)$_{12}$ model

Fig 9 Diagnostic Display for ARIMA(0,1,1)*(0,1,1)

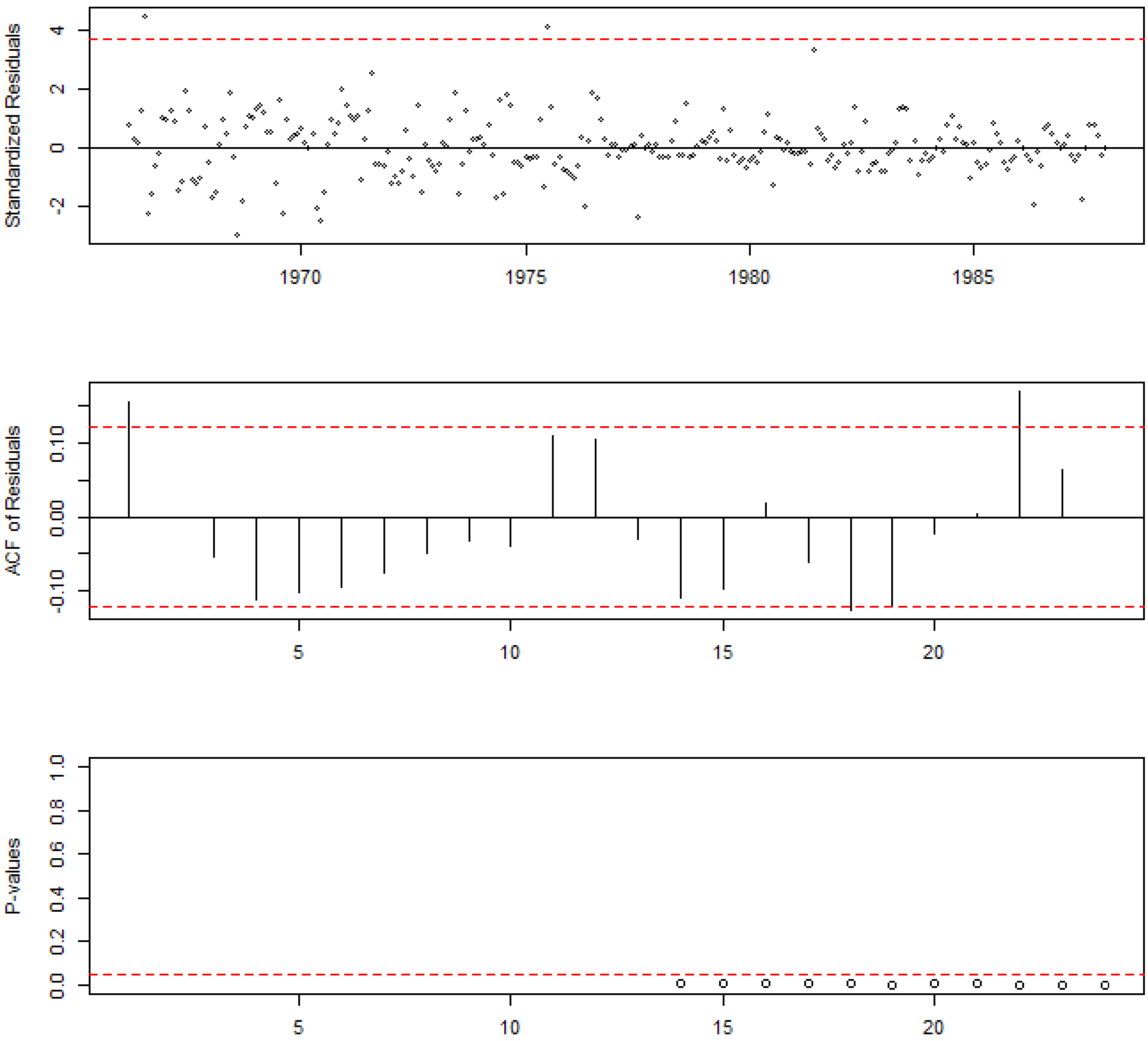


Fig 9 shows three diagnose plots for ARIMA (0, 1, 1) × (0, 1, 1)$_{12}$ model. The time plot of the residuals shows some outliers above the dashed line. The residual autocorrelations are significant at lags 1, 22, and marginally significant at lags 18, 19. The Ljung-Box test statistics have p-values less than 0.05 for all lags, which indicates that the residuals are not white noise. Therefore, ARIMA (0, 1, 1) × (0, 1, 1)$_{12}$ model is not appropriate for this series.

3.2 The ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model

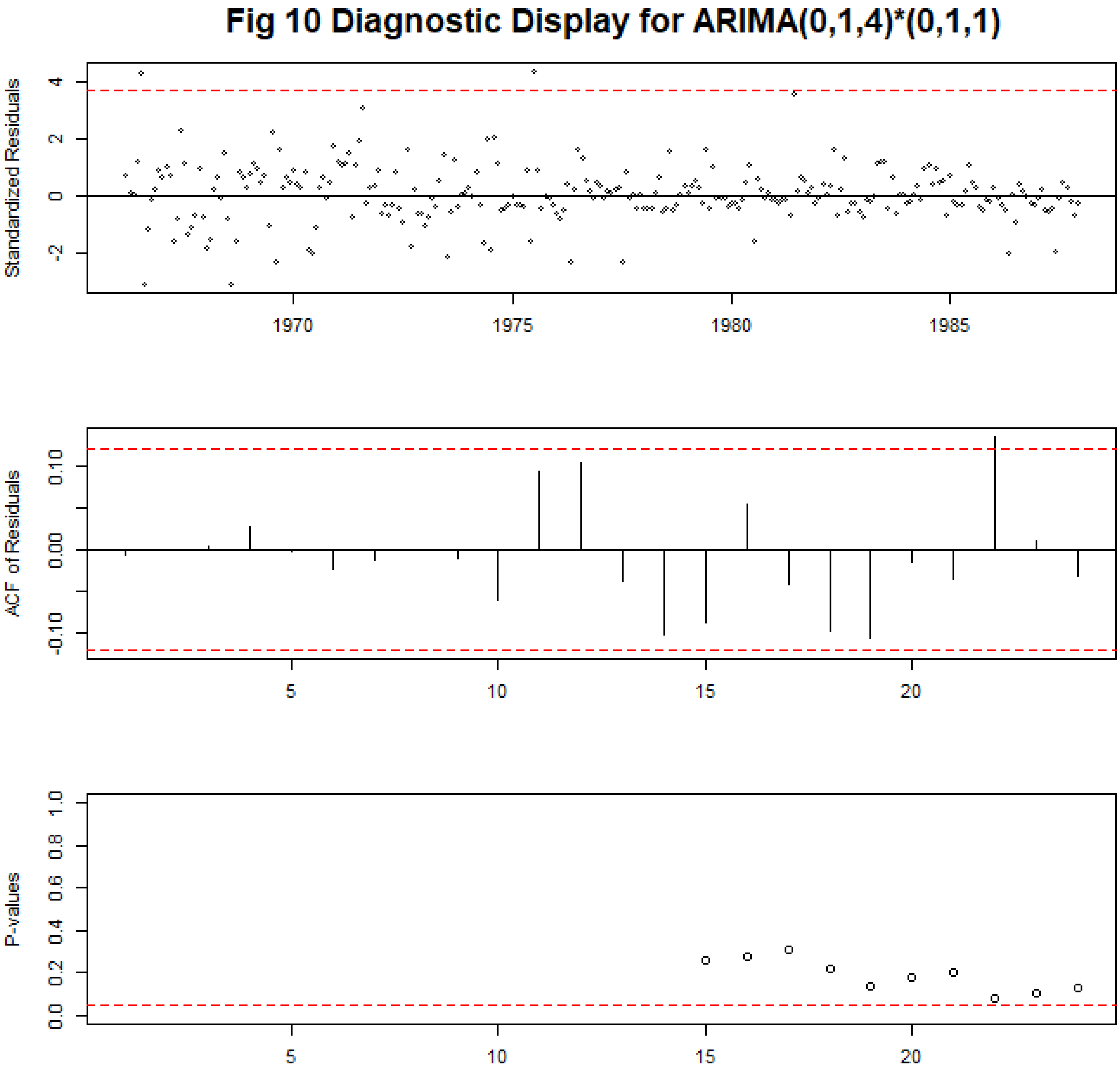
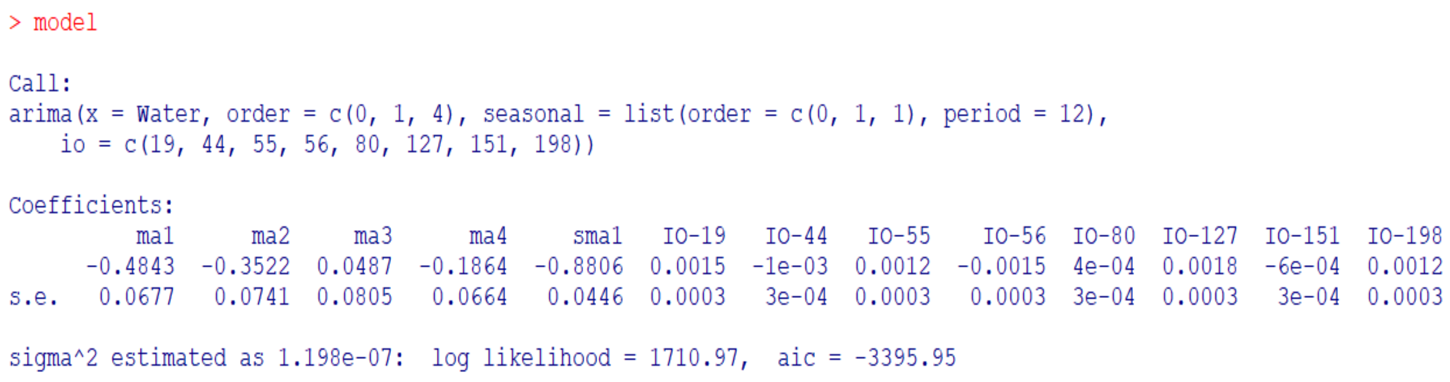


Fig 10 shows three diagnose plots for ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model. The time plot of the residuals still shows some outliers above the dashed line. The residual autocorrelations are significant at lags 22. However, the Ljung-Box test statistics have p-values greater than 0.05 for all lags. Therefore, ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model is better than ARIMA (0, 1, 1) × (0, 1, 1)$_{12}$ model.

3.2.1 Detect Outliers

After detecting AO and IO from the ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model, it shows that there are IO at t = 19, 44, 55, 56, 80, 127, 151, 198.

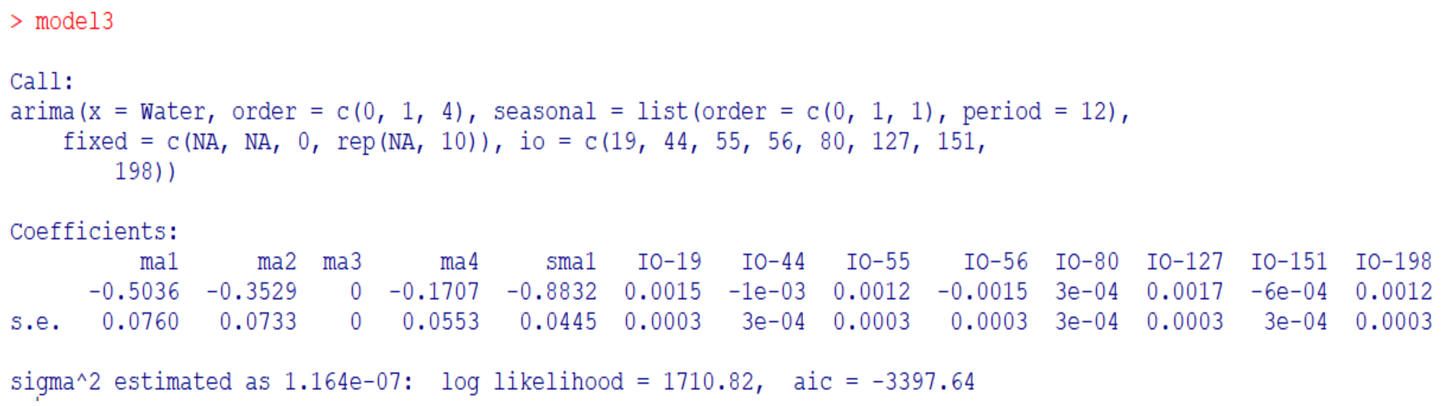
Fig 11 The ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO at t = 19,44,55,56, 80, 127,151,198



```
> model

Call:
arima(x = Water, order = c(0, 1, 4), seasonal = list(order = c(0, 1, 1), period = 12),
    io = c(19, 44, 55, 56, 80, 127, 151, 198))

Coefficients:
          ma1      ma2     ma3      ma4     sma1   IO-19   IO-44   IO-55    IO-56  IO-80  IO-127  IO-151  IO-198
      -0.4843  -0.3522  0.0487  -0.1864  -0.8806  0.0015  -1e-03  0.0012  -0.0015  4e-04  0.0018  -6e-04  0.0012
s.e.   0.0677   0.0741  0.0805   0.0664   0.0446  0.0003   3e-04  0.0003   0.0003  3e-04  0.0003   3e-04  0.0003

sigma^2 estimated as 1.198e-07:  log likelihood = 1710.97,  aic = -3395.95
```

Fig 11 shows the summary output of the ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model. The only insignificant term is “ma3”. The 95% confidence interval of “ma3” is from -0.1093 to 0.2067, which contains zero. All other terms are significant. Hence, only removing “ma3” and keeping all other terms.

Fig 12 The ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO at t = 19,44,55,56, 80, 127,151,198 and ma3 = 0



```
> model3

Call:
arima(x = Water, order = c(0, 1, 4), seasonal = list(order = c(0, 1, 1), period = 12),
    fixed = c(NA, NA, 0, rep(NA, 10)), io = c(19, 44, 55, 56, 80, 127, 151,
        198))

Coefficients:
          ma1      ma2  ma3      ma4     sma1   IO-19   IO-44   IO-55    IO-56  IO-80  IO-127  IO-151  IO-198
      -0.5036  -0.3529    0  -0.1707  -0.8832  0.0015  -1e-03  0.0012  -0.0015  3e-04  0.0017  -6e-04  0.0012
s.e.   0.0760   0.0733    0   0.0553   0.0445  0.0003   3e-04  0.0003   0.0003  3e-04  0.0003   3e-04  0.0003

sigma^2 estimated as 1.164e-07:  log likelihood = 1710.82,  aic = -3397.64
```

Fig 12 shows that all terms are significant now. The estimates of all coefficients are close to the estimates from the full ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model. Furthermore, Fig 12 shows relatively smaller value of AIC. Thus, the ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with ma3 = 0 is the best.

Fig 13 Diagnostic Display for ARIMA(0,1,4)*(0,1,1) with io=19,44,55,56,80,127,151,198 and ma3 = 0

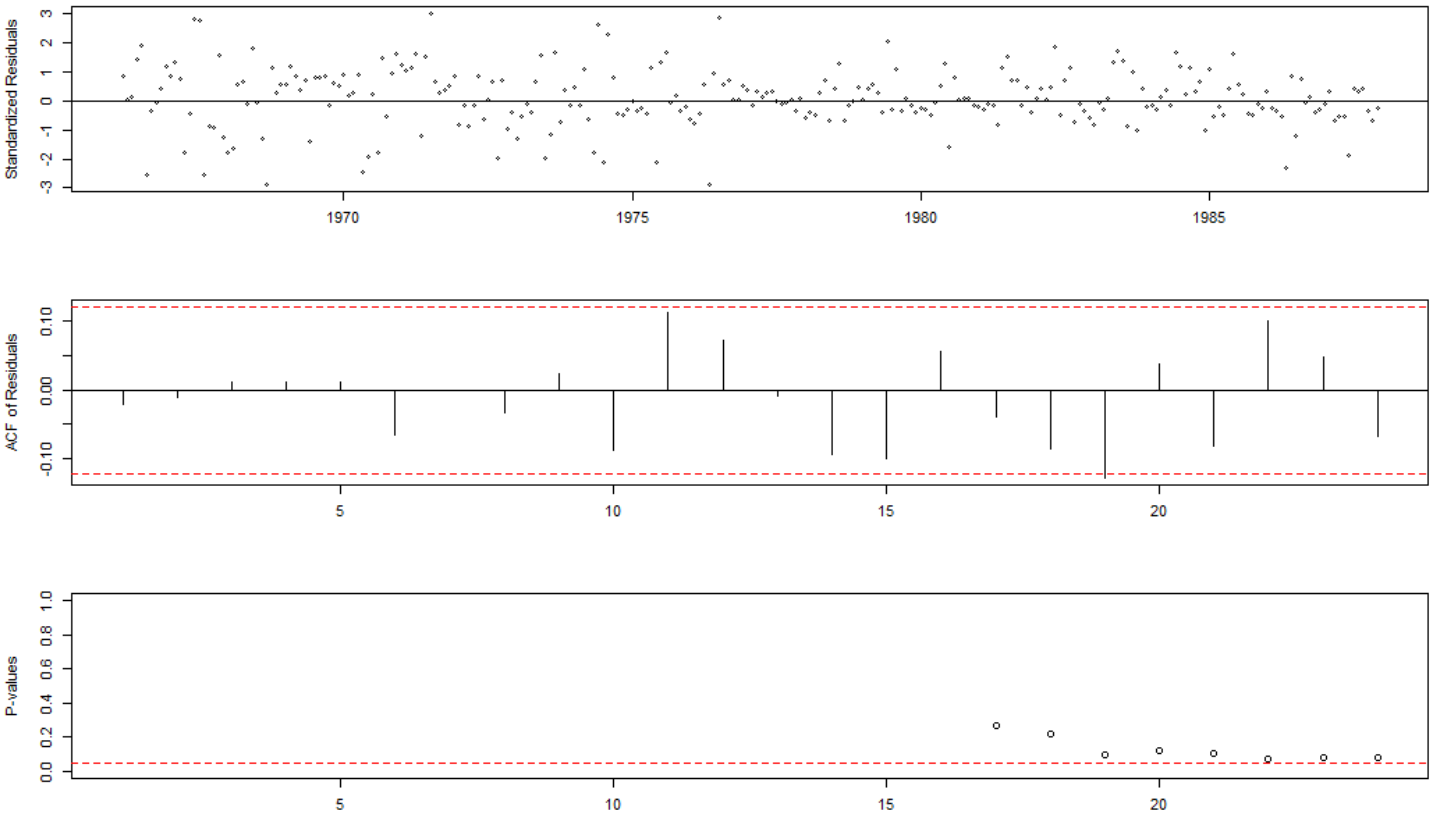


Fig 13 shows three diagnose plots for ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO = 19, 44, 55, 56, 80, 127, 151, 198 and ma3 = 0. The time plot of the residuals still shows no outliers above the dashed line. The residual autocorrelations shows almost no significant lags. The Ljung-Box test statistics have p-values greater than 0.05 for all lags. Therefore, the most appropriate model is the ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO = 19, 44, 55, 56, 80, 127, 151, 198 and ma3 = 0.

4. Diagnostic Checking

4.1 Residuals from ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO = 19, 44, 55, 56, 80, 127, 151, 198 and ma3 = 0

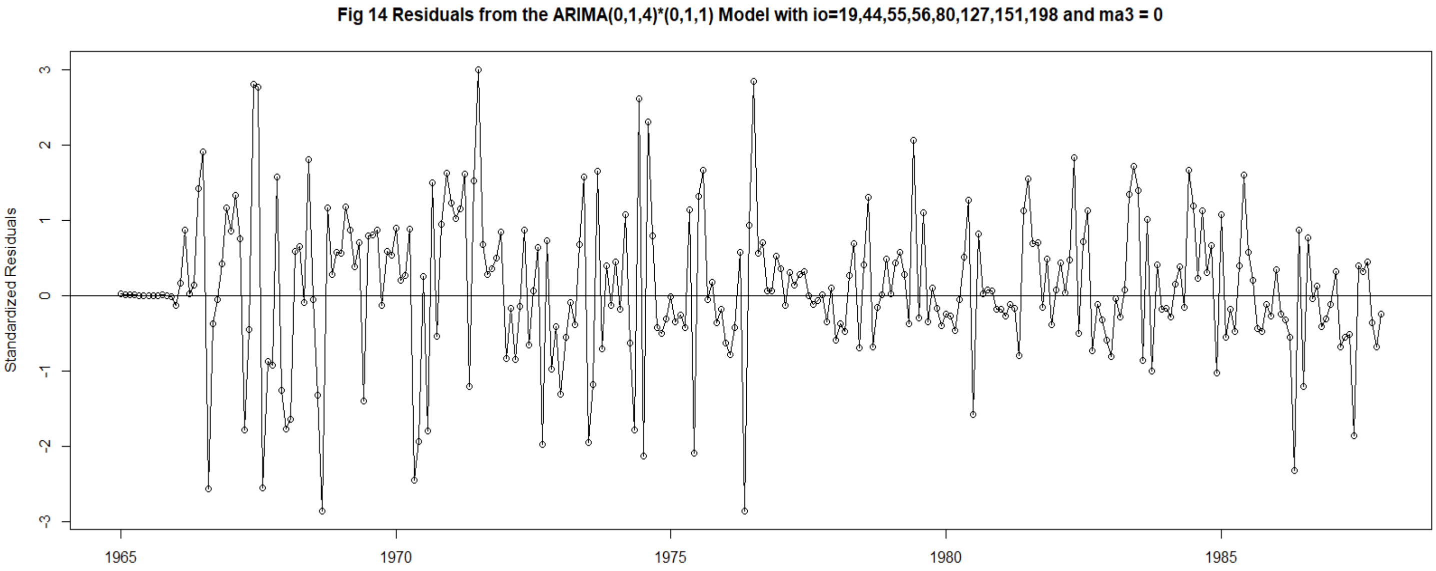


Fig 14 gives the plot for standardized residual. This plot does not suggest any major irregularities with the model.

4.2 Sample ACF of Residuals from the ARIMA(0,1,4)(0,1,1) Model with IO=19,44,55,56,80,127,151,198 and ma3 = 0*

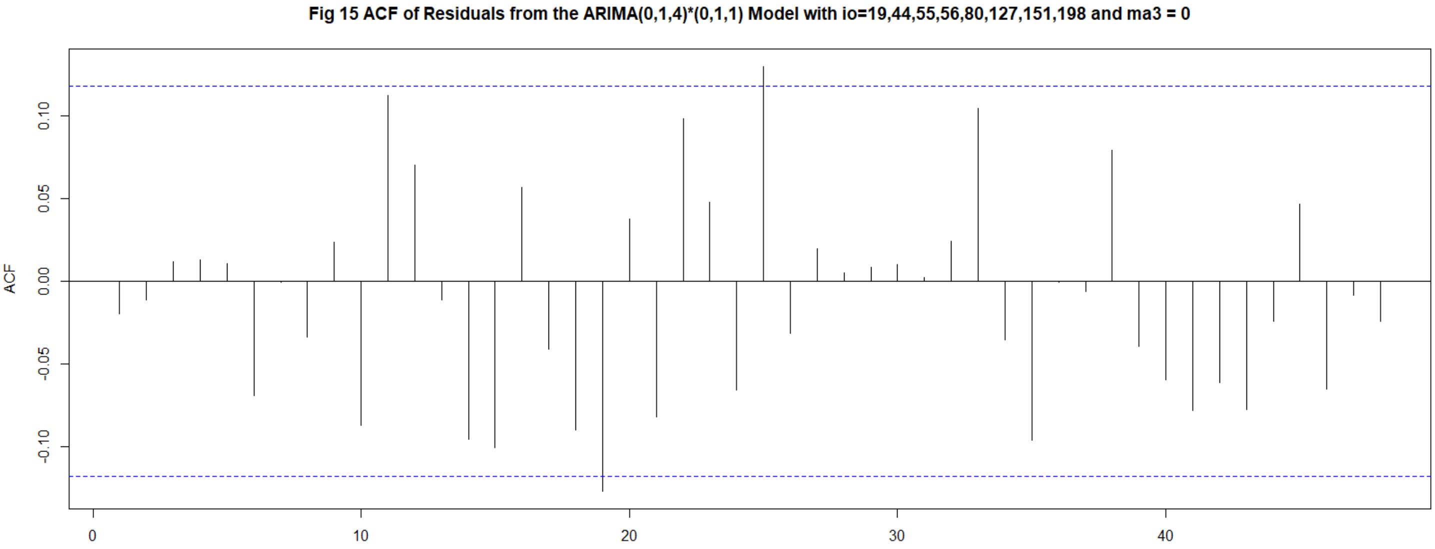
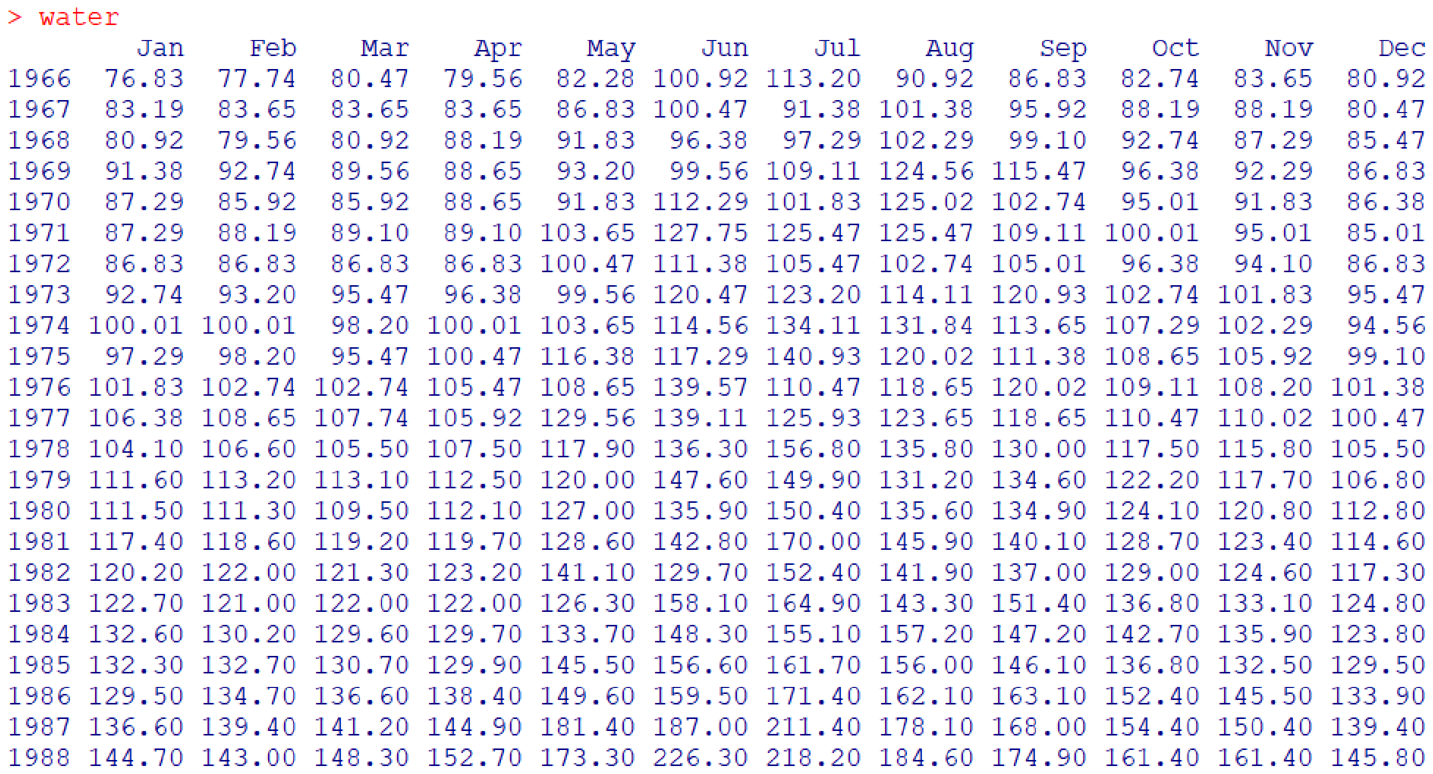


Fig 12 shows the statistically significant correlations are at lag 19 and 25. Both of them have a small value of correlation, not exceed 0.15. Moreover, for all 48 lags displayed on the plot, there is only two statistically significant, which could be caused by chance alone. Therefore, except for marginal significance at lag 19 and 25, the model seems to have captured the essence of the dependence in the series.

5. Conclusion

Time series analysis of 23 years of monthly water usage shows a strong upward trend and a seasonal pattern. After taking power transformation, first difference, seasonal difference, outlier detecting, and model fitting, the most appropriate model is an ARIMA (0, 1, 4) × (0, 1, 1)$_{12}$ model with IO = 19, 44, 55, 56, 80, 127, 151, 198 and ma3 = 0. This model describes most featured of the series and expresses the seasonal pattern. Although the model cannot make predictions, it already offers sufficient messages.

6. The Data



```
> water
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
1966  76.83  77.74  80.47  79.56  82.28 100.92 113.20  90.92  86.83  82.74  83.65  80.92
1967  83.19  83.65  83.65  83.65  86.83 100.47  91.38 101.38  95.92  88.19  88.19  80.47
1968  80.92  79.56  80.92  88.19  91.83  96.38  97.29 102.29  99.10  92.74  87.29  85.47
1969  91.38  92.74  89.56  88.65  93.20  99.56 109.11 124.56 115.47  96.38  92.29  86.83
1970  87.29  85.92  85.92  88.65  91.83 112.29 101.83 125.02 102.74  95.01  91.83  86.38
1971  87.29  88.19  89.10  89.10 103.65 127.75 125.47 125.47 109.11 100.01  95.01  85.01
1972  86.83  86.83  86.83  86.83 100.47 111.38 105.47 102.74 105.01  96.38  94.10  86.83
1973  92.74  93.20  95.47  96.38  99.56 120.47 123.20 114.11 120.93 102.74 101.83  95.47
1974 100.01 100.01  98.20 100.01 103.65 114.56 134.11 131.84 113.65 107.29 102.29  94.56
1975  97.29  98.20  95.47 100.47 116.38 117.29 140.93 120.02 111.38 108.65 105.92  99.10
1976 101.83 102.74 102.74 105.47 108.65 139.57 110.47 118.65 120.02 109.11 108.20 101.38
1977 106.38 108.65 107.74 105.92 129.56 139.11 125.93 123.65 118.65 110.47 110.02 100.47
1978 104.10 106.60 105.50 107.50 117.90 136.30 156.80 135.80 130.00 117.50 115.80 105.50
1979 111.60 113.20 113.10 112.50 120.00 147.60 149.90 131.20 134.60 122.20 117.70 106.80
1980 111.50 111.30 109.50 112.10 127.00 135.90 150.40 135.60 134.90 124.10 120.80 112.80
1981 117.40 118.60 119.20 119.70 128.60 142.80 170.00 145.90 140.10 128.70 123.40 114.60
1982 120.20 122.00 121.30 123.20 141.10 129.70 152.40 141.90 137.00 129.00 124.60 117.30
1983 122.70 121.00 122.00 122.00 126.30 158.10 164.90 143.30 151.40 136.80 133.10 124.80
1984 132.60 130.20 129.60 129.70 133.70 148.30 155.10 157.20 147.20 142.70 135.90 123.80
1985 132.30 132.70 130.70 129.90 145.50 156.60 161.70 156.00 146.10 136.80 132.50 129.50
1986 129.50 134.70 136.60 138.40 149.60 159.50 171.40 162.10 163.10 152.40 145.50 133.90
1987 136.60 139.40 141.20 144.90 181.40 187.00 211.40 178.10 168.00 154.40 150.40 139.40
1988 144.70 143.00 148.30 152.70 173.30 226.30 218.20 184.60 174.90 161.40 161.40 145.80
```

7. Summary

Based on the improvement of life quality, the demand for water usage in city rapidly increases. Therefore, the strong upward trend is corresponded with the increasing water usage. Compared with winter months, people usually take more showers and play more water games in summer, as a result, the demand for water is highest during the warmer summer months and lowest during the winter time. This physical understanding of water usage phenomenon helps to account for seasonality and trends.

The seasonal autocorrelation from the sample ACF plot confirms that there is a periodical pattern on water usage. The disappearance of seasonality in sample ACF plot when taking first and seasonal difference confirms that seasonal ARIMA model is necessary. The EACF plot of this series is so poor and hard to read, so this paper does not include the EACF plot, since it does not provide any useful information.

There are some troubles in model fitting. The basic ARIMA (0, 1, 1) × $(0, 1, 1)_{12}$ model shows very poor diagnose condition. After changing the values of p, q, P, and Q in the model, the diagnose plots still report many problems. And then detecting outliers is essential for this series.

No additive outlier but around ten innovative outlier means that the next key step for this series is to handle with these outliers. After dealing with these outliers and removing the insignificant terms, the model looks good, and the diagnose plots do not display any series issues.

The final model for this series is an ARIMA (0, 1, 4) × $(0, 1, 1)_{12}$ model with IO = 19, 44, 55, 56, 80, 127, 151, 198 and ma3 = 0. After fitting model, checking model is necessary as well. Residual plots shows no series problems for this model. Hence, this ARIMA model is acceptable based on the all statistical methods included in.

Since there are many innovation outliers in the series, future prediction cannot be done. In statistic study, it is important to make prediction for the future. No prediction really discounts the effects of model. But people can obtain enough information from this seasonal model, because seasonal ARIMA model self has enough properties.

Increasing demand of water usage truly warns people to protect water sources, especially in the situation that water scarcity and water contamination are major issues all around world. Seasonal variation gives a good suggestion for government to fully prepare for water demand. Therefore, ARIMA model really provides enough information for people to deal with water usage problem.