Title

1 Introduction

The scientific question motivating my work is *can we predict the monthly precipitation of a temperate continental climatic region such as Zhangye using time series methods?* Many factors have an impact on changes of a regional precipitation. The influence not only has long-term trends and seasonal effects, but also has a random disturbance effect. On the basis of monthly precipitation data from January 1951 to December 1990 of a typical temperate continental climatic city - Zhangye, which is located in the northwestern China, we are trying to fit a model and predict the future rainfall, which is very important for agriculture and industry.

Why not updated?

2 Method

Should have this section.

First we will perform exploratory data analysis on the training data(year 1951 - year 1985) to look at our precipitation data, examine the assumptions for our models and make suitable transformation of the data. After confirming a dominant frequency of our data, we decided to use seasonal modeling methods. In this project I will employ three different models:

- Naive Seasonal Mean Model
- Multiplicative Seasonal ARIMA Model
- Spectral Analysis - Harmonic Model

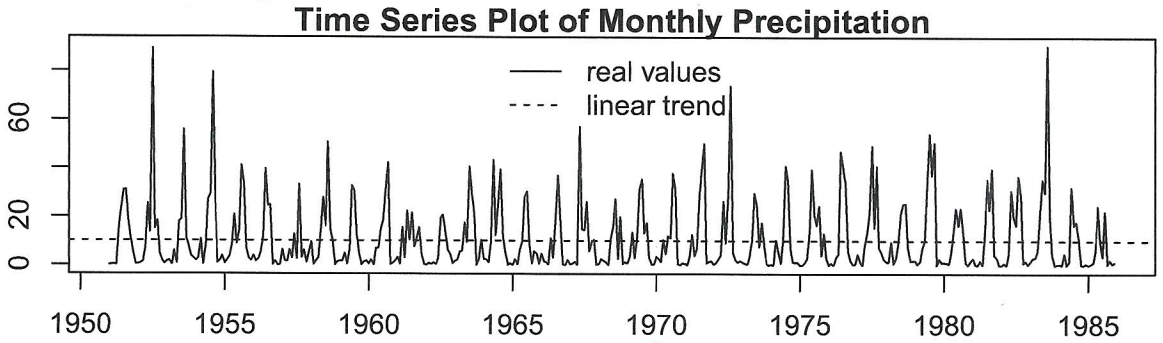
✓

In choosing these models, we use several criteria including AIC, BIC, MSE and adjusted r squared. Then we make forecast for year 1886 - 1990 based on our selected model, and compare to see if our test data fall in the prediction's 95% confidence interval.

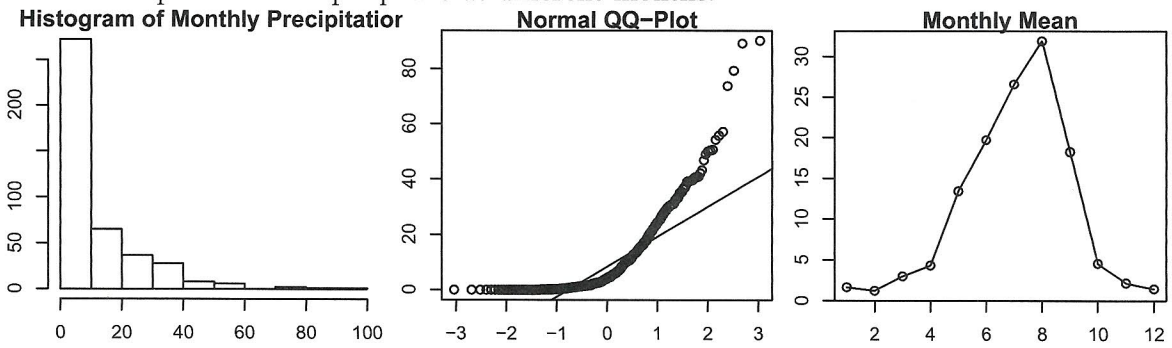
These measure uncertainty.

3 Data

The original data consisted of monthly average precipitation record from January 1951 to December 1990 of ZHANGYE (CHINA) [coordinates: 38.93N, 100.43E, 1483m WMO station code: 52652], which were provided by the KNMI Climate Explorer[1]. There are 480 data points, and we use the first 420 data to build our model and the last 60 data to check our model forecast performance.



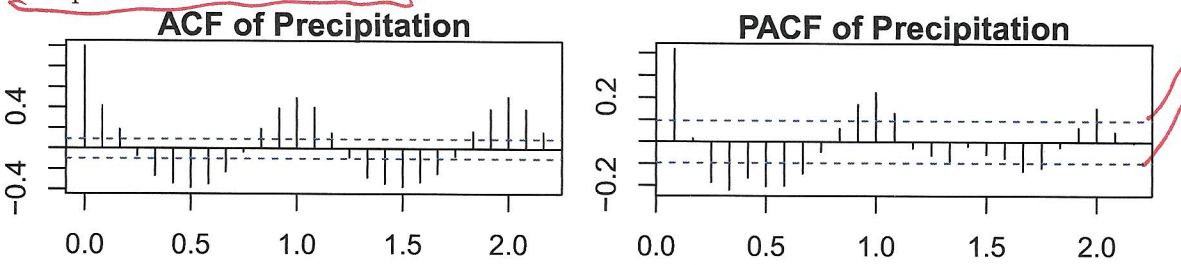
From the time series plot we observed our series is oscillating with spikes and valleys, with an apparent annual cyclic pattern. However each year is very different. For example, we have three peaks with monthly precipitation of 90.3, 89.3 and 90.3, and we also have 32 months that do not have any rainfall. Zoom in, we can see that within a specific year, there can be one peak or multiple peaks at different months.



Look at the histogram, we can see that there are many months with very little precipitation and the distribution of monthly precipitation not normal and has a long right tail.

The quantile-quantile plot confirms this.

The average monthly precipitation shows the differences between each month. Clearly May to Sepetember is the wet season.



Last but not least, the Autocorrelation Function and the Partial Autocorrelation Function show that (1): the ACF does not tail-off quickly, hence the monthly precipitation time series is not stationary; (2) the ACF plot shows a sine(cosine) wave pattern; (3) Many autocorrelations/partial autocorrelation falls outside the 2 standard deviation. This along with the periodogram indicate that we should use a seasonal model.

4 Naive Seasonal Mean Model

The slope of a linear regression on time is close to 0(-0.0004368), as shown by the dashed line above, which means that there is almost no global trend in our data. The analysis done in the previous section leads us to examine the seasonal trend. We take log to improve normality of data and consistency of variance

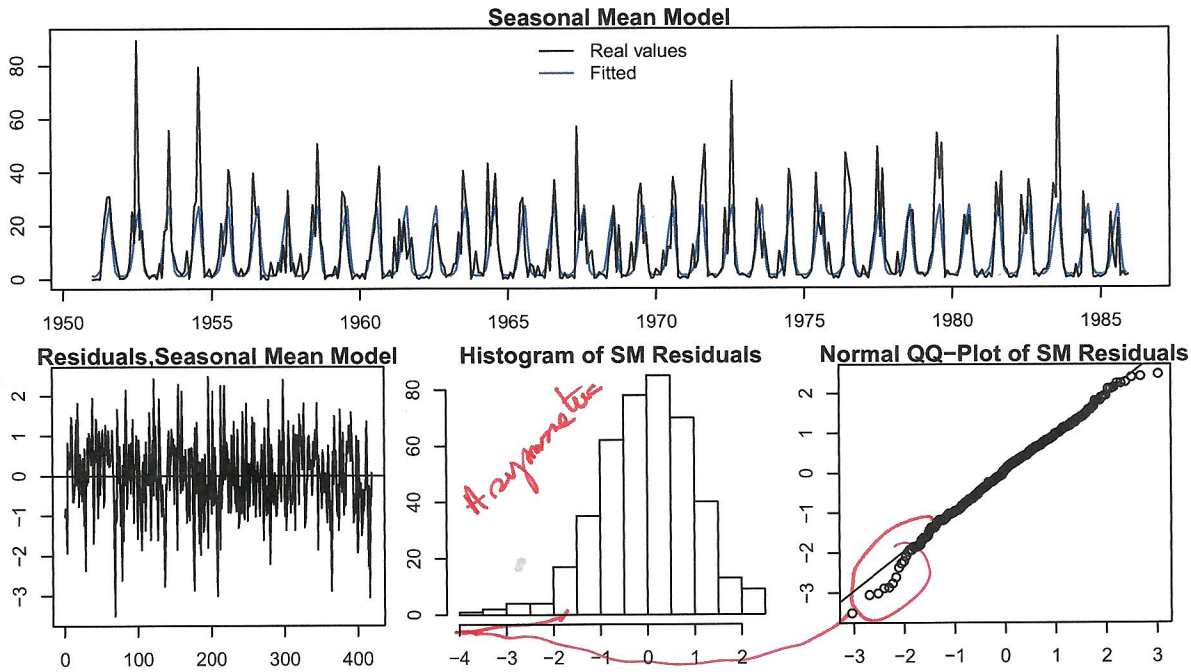
$$ylog_t = ln(x_t + 1)$$

. We use $x_t + 1$ here to avoid NaN because the original precipitation data contains 0.0(when there was no rainfall that month). From the histogram and qq-plot,we can see that our transformed series *logd* looks normal. Shapiro-Wilk test of normality confirmed our result with a p-value of 0.5692. We use the von Mises statistic to compare series z with white noise, and the p-value is 2.140096×10^{-35} which indicates that our series is now a stationary

non-white noise series. Now we can proceed to the next step - model choosing. Denote the log difference as *logd*, the fitted monthly mean are:

Jan	Feb	Mar	Apr	May	Jun
1.127345	1.021492	1.752533	3.183572	8.292285	16.380569
Jul	Aug	Sep	Oct	Nov	Dec
22.566620	27.128956	12.378532	2.954182	1.395050	1.090973

Plot these



Put on same scale as graph above

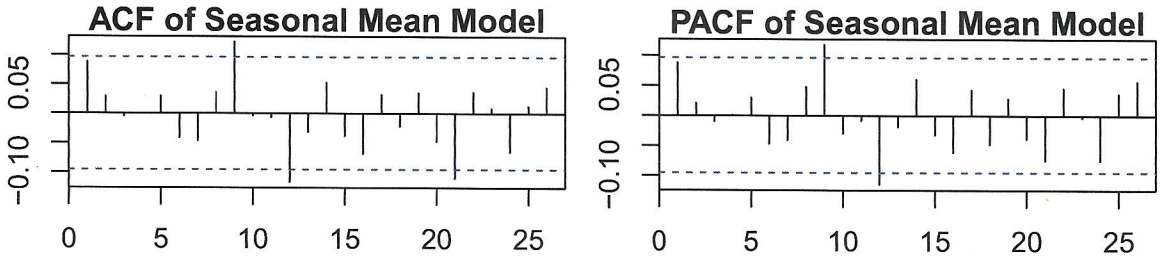
Asymmetric

4.1 Diagnostic

Unclear

Our model fits the general progression of the model, so we examine the residuals.The residual plot follows the straight line pretty closely; The normal-quantile quantile plot and the histogram does not show perfect normality, but they are fairly close to normal. We also performed Shapiro-Wilk Normality Test, which produces a test statistics of W = 0.9916, which corresponds to a p-value of 0.01821. The adjusted r-square gives us 0.8766, and the model mean square error is 0.5480669.

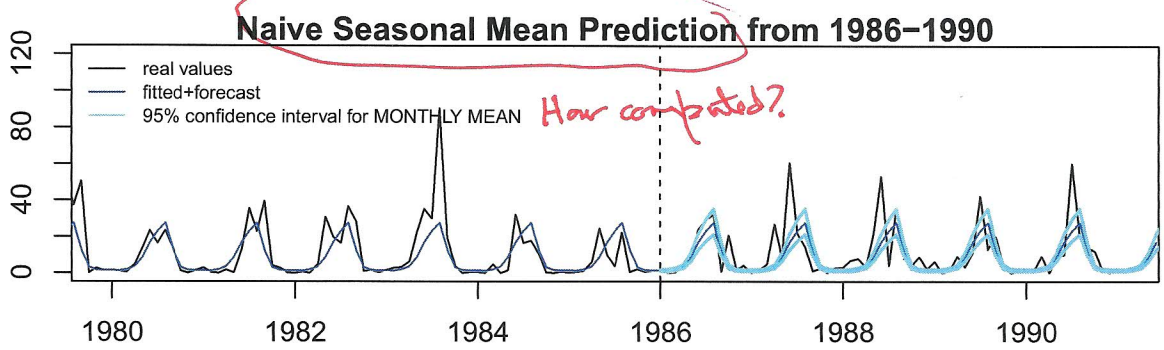
Discuss. Look at histogram and QQ plots



In addition, the ACF and PACF does not show strong correlation, despite some significance showing on lag sadsadasd and asdsada of ACF. Our seasonal mean model gives us an simple prediction and intuitive prediction, but it is not very effectively reflecting the developing from year to year.

How defined?

4.2 Forecast



The Naive seasonal data follows the series pretty well.

5 Seasonal ARIMA Model

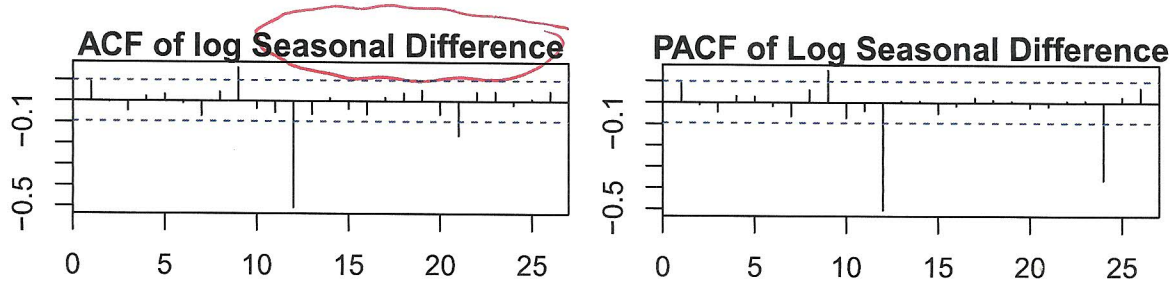
The multiplicative seasonal ARIMA model incorporates both non-seasonal and seasonal factors, therefore we will chose it as our model:

$$ARIMA(p,d,q) \times (P,D,Q)S$$

p = non-seasonal AR order, d = non-seasonal differencing, q=non-seasonal MA order, P = seasonal AR order, D = seasonal differencing , Q = seasonal MA order, and S = time span of repeating seasonal pattern.

The auto.arima returns a fit of SARIMA(1,0,0)(2,0,0)[12], but the AIC and BIC are much larger and the MSE is also larger

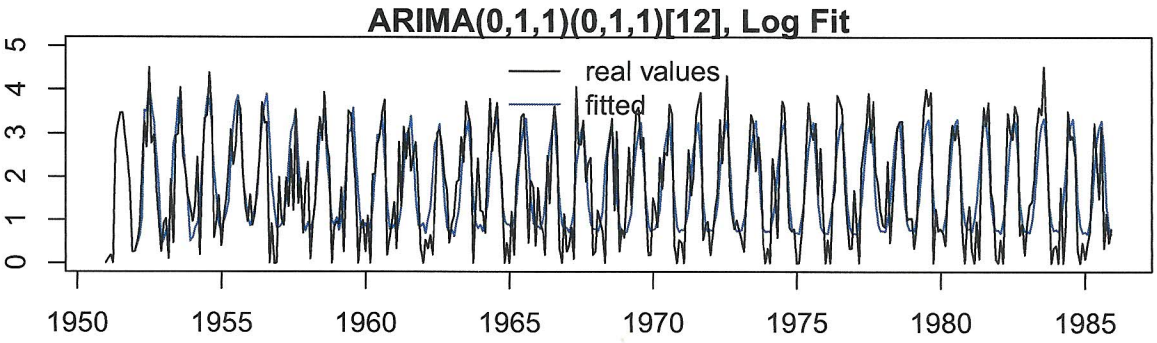
Where are the values?



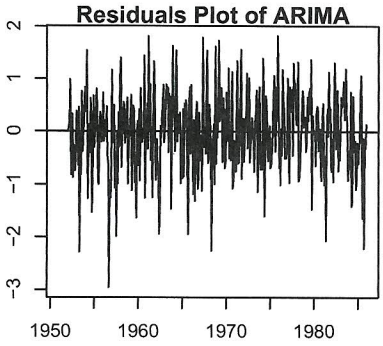
In terms of choosing the model, from the ACF and PACF, we see that the model follows a cyclic pattern of S=12. All lags are very highly correlated. We plot the ACF and PACF of a seasonal difference, and that removes most of the correlation. Therefore we have a D=1 in our model. We still have some seasonality left in the data, as shown at lag 12 and 24 of both ACF and PACF plots.

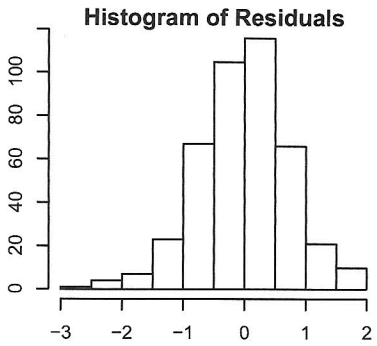
From the armasubsets and eacf function form R, and comparing their AIC and BIC, we choose our model to be SARIMA(0,1,1)×(0,1,1)[12].On the log scale, the model fits well with a residual mean squared error of 0,73868. The coefficients of the random component and the seasonal component, in our case the moving average 1 and seasonal moving average term are:

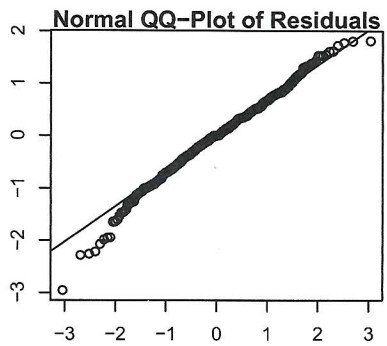
MA(θ_1)	-0.9999922	SMA(Θ_1)	-0.9999639



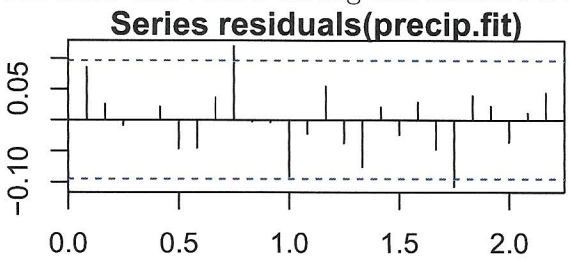
5.1 Diagnostic

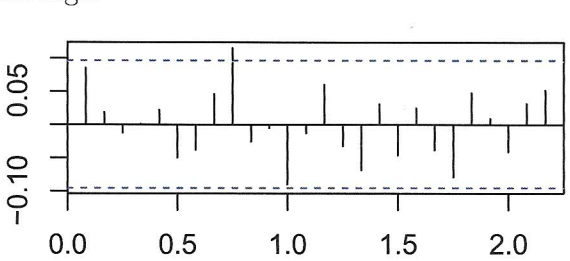




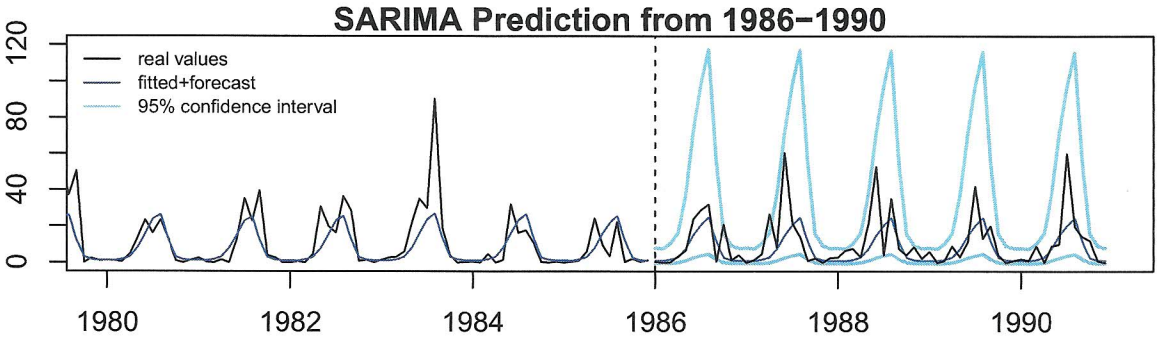


The residuals give a fairly good result, as compared with the naive seasonal mean model, the residual has less trend. The histogram and qq-plot of the residuals are pretty normal. The p values for Ljung-Box statistics does not show significance. ACF and PACF of the residuals don't show strong correlation between lags.



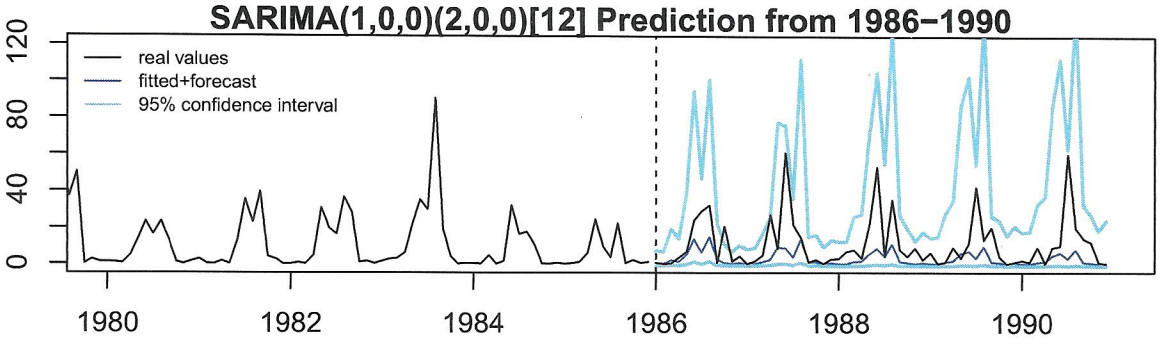


5.2 Forecast



As shown in the plot, our prediction in general matches the precipitation data from 1986 to 1990. The 95% confidence interval also covers almost all data points, despite one data point at the early month of year 1987.

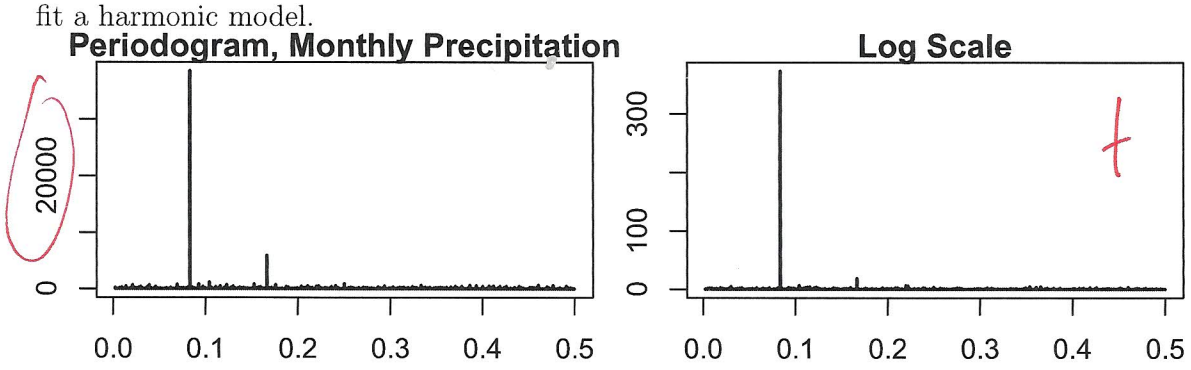
5.3 Auto Fit Seasonal ARIMA using R



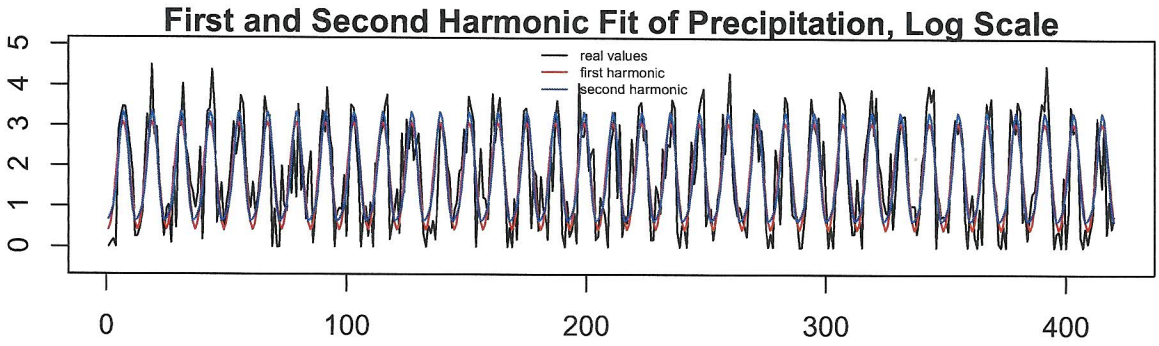
The auto-fit from the R packageforecast does not do a very good job. The AIC and BIC are large, and the fit is not as good.

6 Spectral Analysis

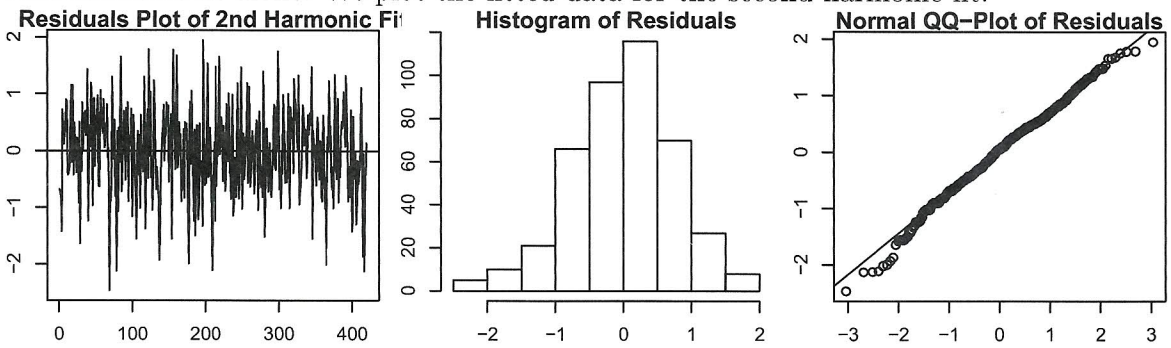
Some time series have a seasonal component difficult to spot, especially if we do not know the period in advance: a periodogram, also known as "sample spectrum" (simply a discrete Fourrier transform) can help us find the period[4]. For us, as we analyzed before, the period is fairly easy to spot - 12 months per cycle. We will examine the harmonics for our data and fit a harmonic model.



We plot the periodogram to confirm the dominant cycles in this series. There are two obvious peaks at about $\frac{1}{12}$ and $\frac{1}{24}$, which indicates the length of a cycle is 12 months or 6 months. The log transformed series shows an even more dominant peak at $\frac{1}{12}$

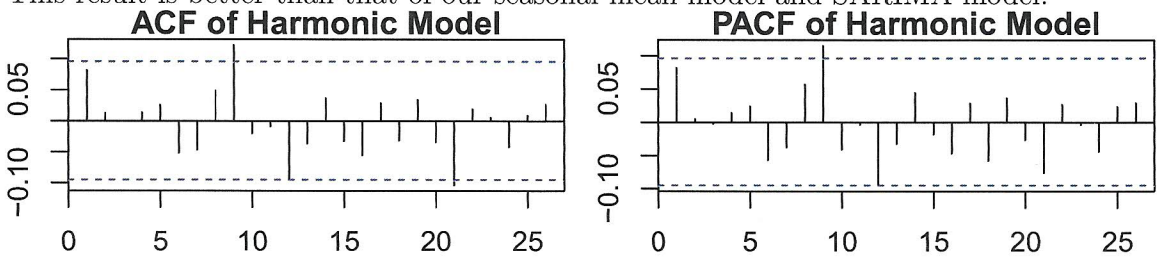


We examine each order and test wehther each term is significant. It turns out the second harmonic is significant, but not for the third harmonic and above. We compare the two fits on log scale in the plot above, and it does not show much difference from the simple fit of first harmonic fit alone. We plot the fitted data for the second harmonic fit.

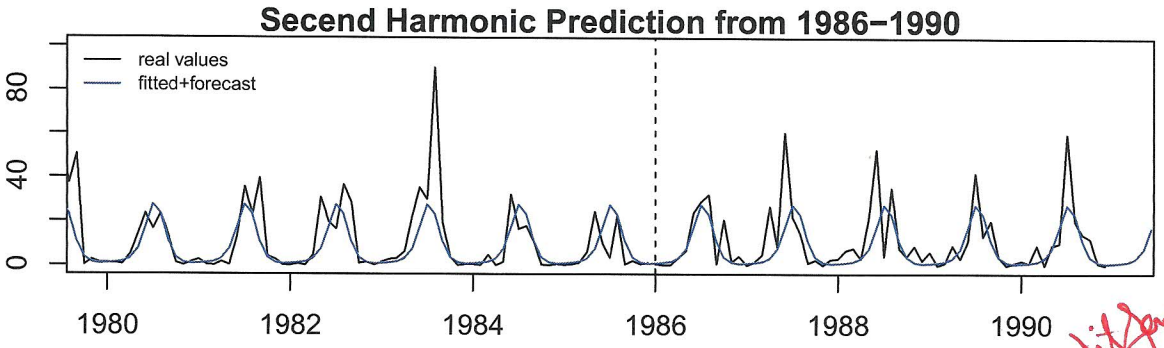


6.1 Diagnostic

The histogram and qq-plot shows that our data is close to normal. We performed Shapiro-Wilk normality test, with a test statistics of W = 0.9936 and according p-value = 0.07144. This result is better than that of our seasonal mean model and SARIMA model.



6.2 Forcast



We can see that the prediction follows the real data well.

7 Discussion

7.1 Comparing Models

Among the three methods, the naive seasonal model is very straightforward, and overall gives us a good fit. The Seasonal ARIMA fit and sine fit at the second harmonic also gives a fairly good fit. All three model give pretty similar results, therefore we can choose any one. The harmonic fit gives the most normal residuals and the smallest AIC/BIC.

7.2 Limitations

Unfortunately, our model is not sophisticate enough to predict the exceptionally high precipitation months. The effects are random according to our model. If we have other data that could relate to rainfall, for example temperature and humidity, we might be able to explain the precipitation of Zhangye city better. Moreover, the dataset does not contain the most recent data, therefore it can be not very effective predicting today's precipitation.

8 Conclusion

The answer to my question is *we can use any one of the seasonal mean model, Seasonal ARIMA model and Harmonic sine wave model to predict the monthly precipitation of Zhangye(a temperate continental climatic region) using time series methods. They give similar results.*