Forecasting Rainfall in India

Vishal Vincent Joseph

Contents

Motivation

- Floods and droughts are becoming quite common and unpredictable in India with every passing year
- Responsible for causing destruction to property, loss of lives and most importantly, the high farmer suicide rates
- Necessitates the need to be better prepared against these and that requires a knowledge of the expected amount of rainfall in the relevant time period, by means of reliable forecasts
- Being a resident of India and having experienced this destruction first-hand was my primary motivation behind choosing this topic

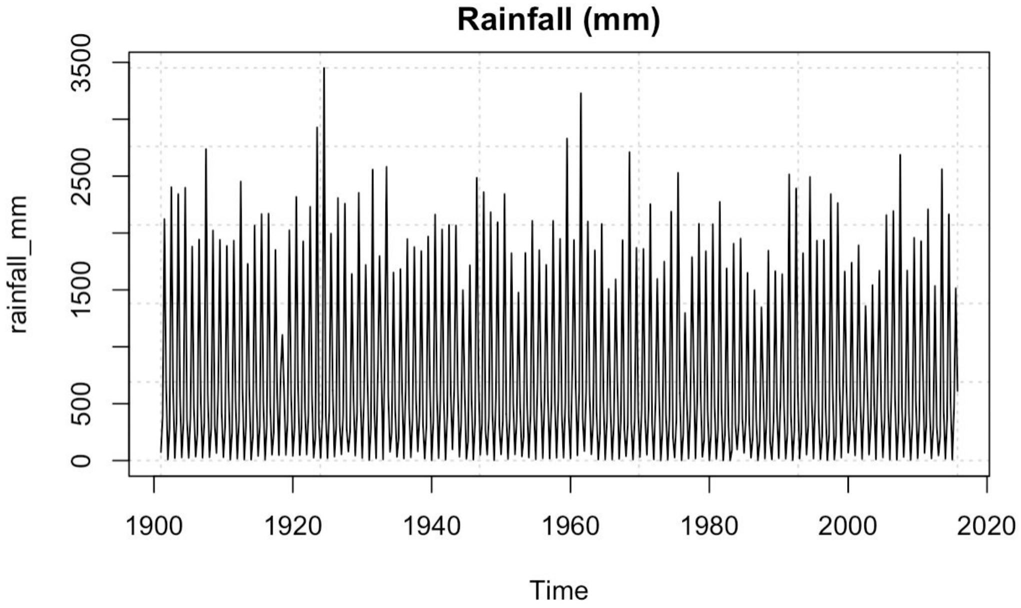
Motivation

Data

- Data acquired from the open source data portal of India and contains annual/quarterly/monthly rainfall data of 36 meteorological subdivisions of India from 1901-2015. Rainfall is measured in mm.
- Data available for around 15 states, but I chose the state of Kerala due to my ancestry and the fact that conditions have been relatively worse of late
- Chose quarterly data since monthly data had zero values and that could complicate analysis and modeling efforts
- Training data from 1901 to 2013; Test data consists of 2014 and 2015

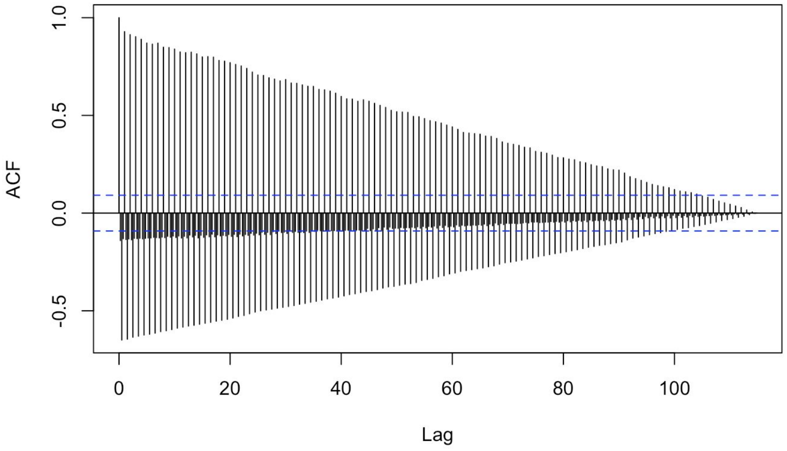
Exploratory Data Analysis

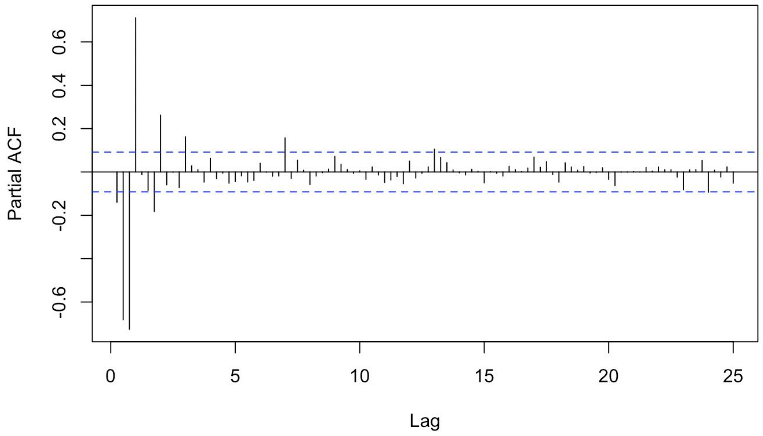
Data characteristics - Raw TS plot



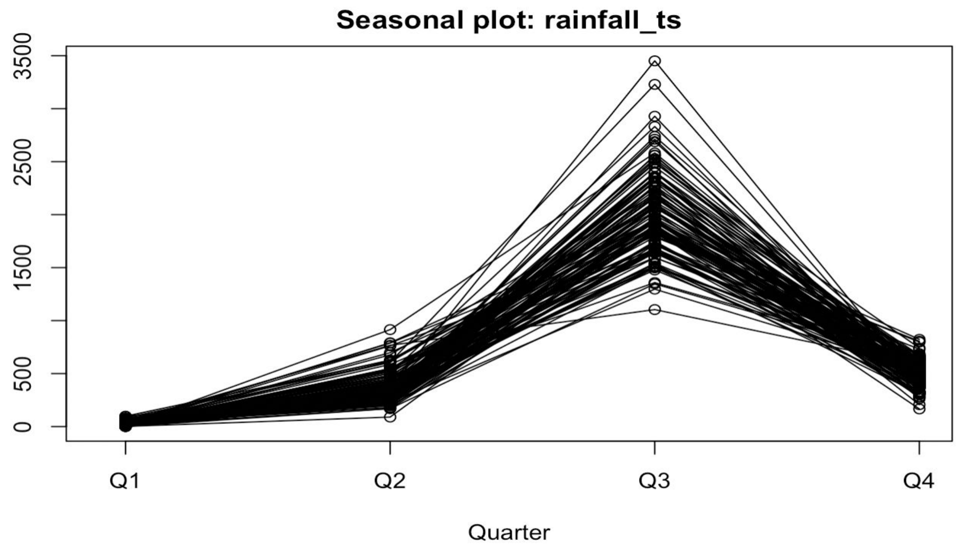
- No visible trend in data, which contradicts increasing levels of destruction caused over the years
- Presence of annual seasonality
- Variance seems to be varying, but need to do further analysis to confirm Box-Cox transformation

Data characteristics - ACF & PACF

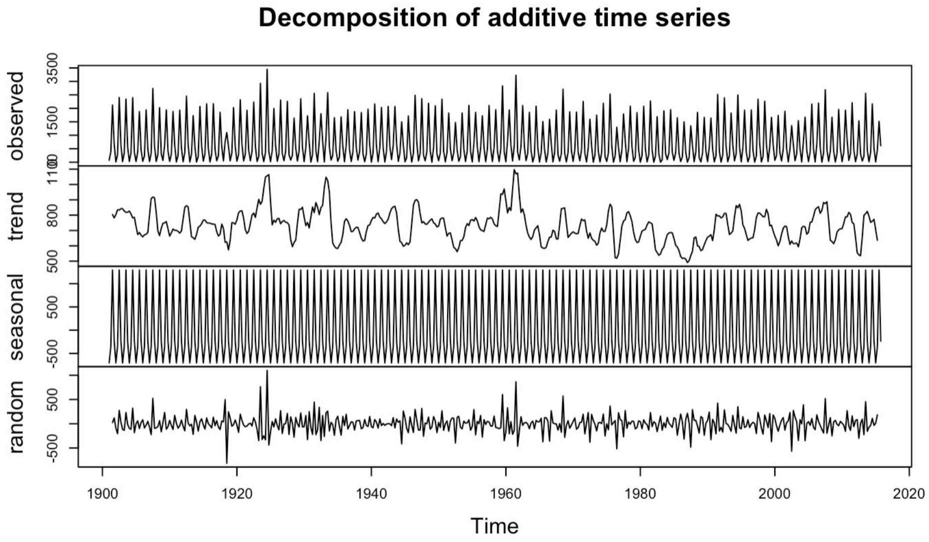




Data characteristics - Seasonal plot

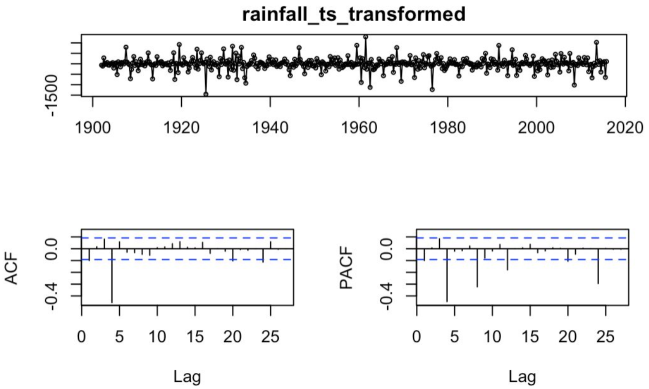


Classical Decomposition

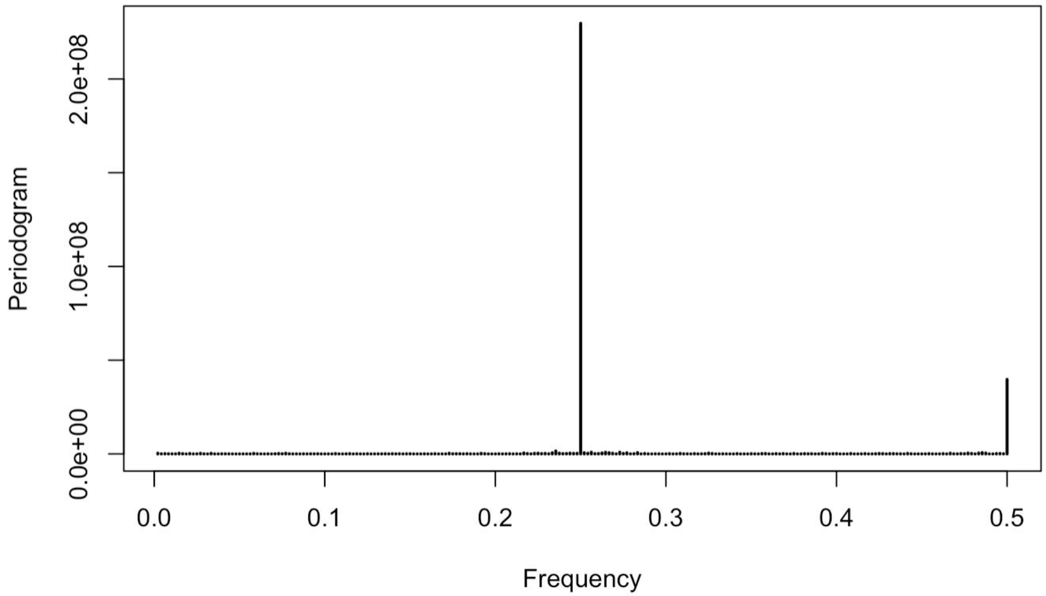


Stationarity analysis

- Both KPSS and ADF test results confirm the presence of stationarity in the raw data, thus deeming the need for any differencing of Box-Cox transformation unnecessary (for level or trend).
- Seasonal differencing at lag 4 gives stationarity. Both KPSS and ADF results confirm stationarity when applied on the final transformed dataset.



Spectral analysis



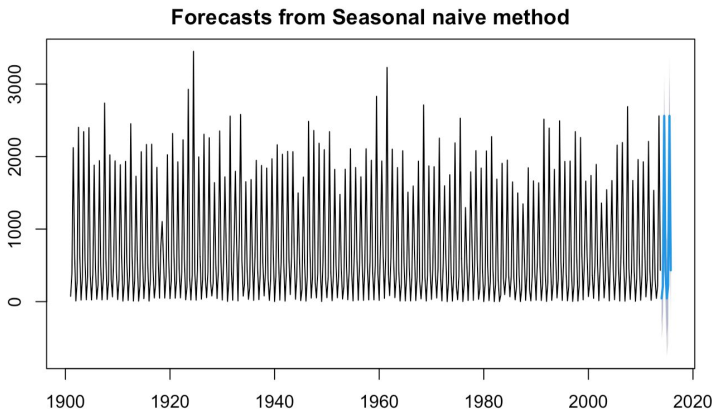
The frequency corresponding to the peak is ~0.25, indicating **annual** seasonality.

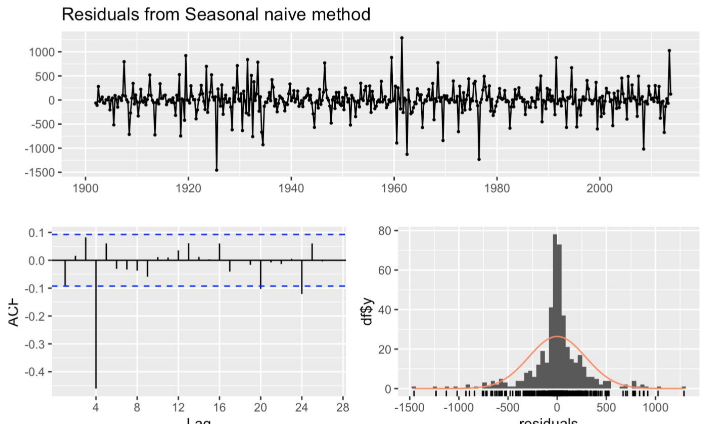
Model building

Models under consideration

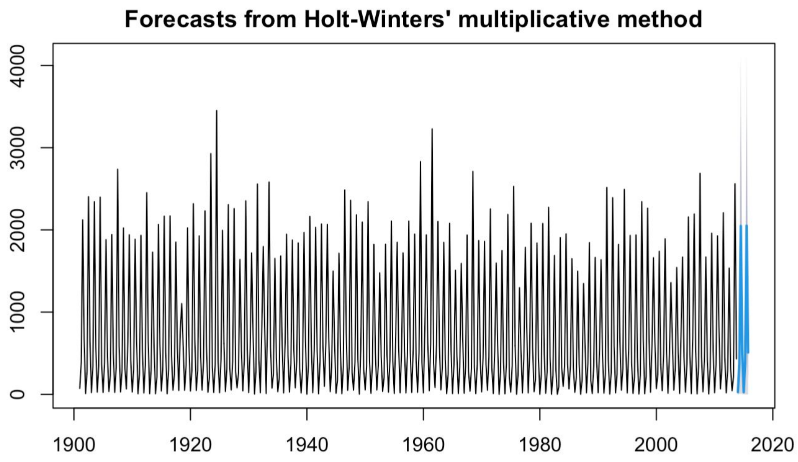
1. Seasonal Naive
2. Holt-Winters seasonal
3. ETS (State Space)
4. ARIMA
5. ARFIMA
6. Neural net

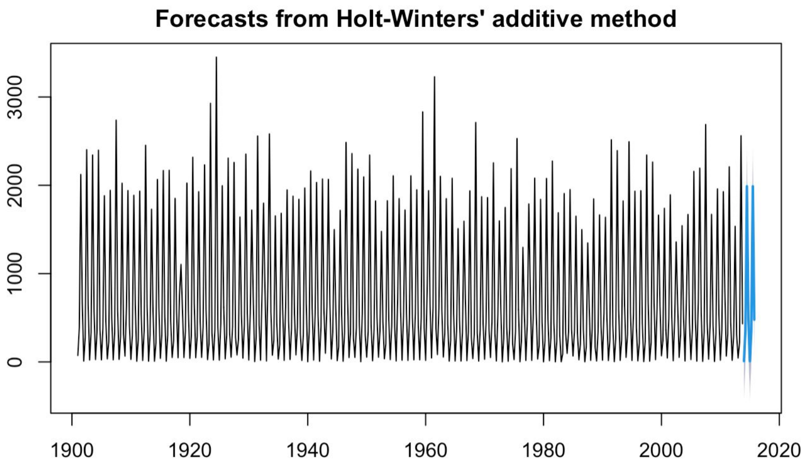
Seasonal Naive





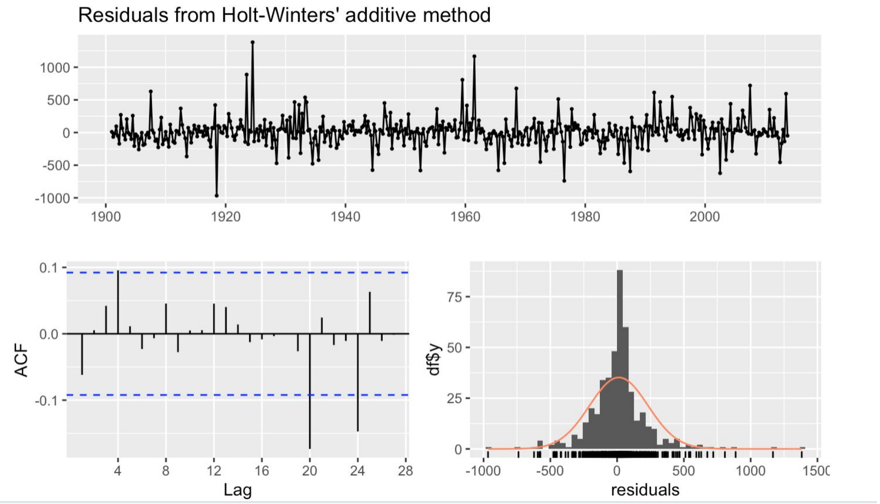
Holt-Winters' - Forecasts

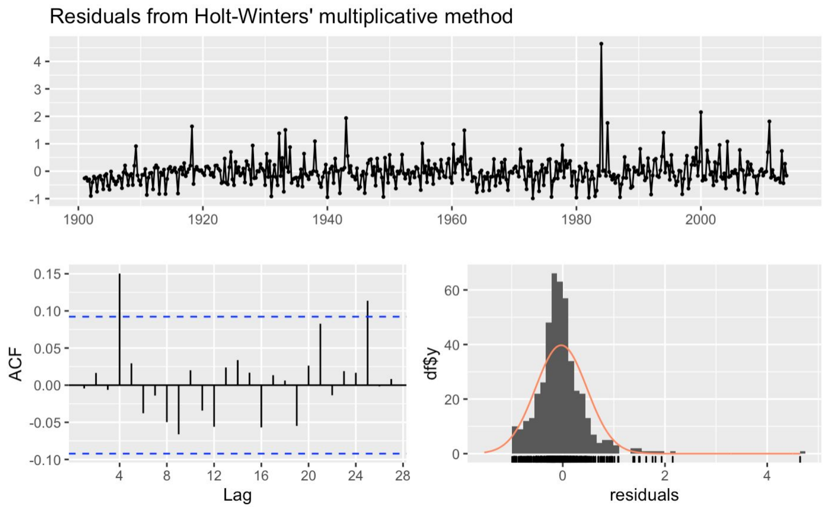




AICc values for multiplicative model lower compared to additive model. This suggests the former would be a better choice but contradicts the raw data plot.

Holt-Winters' - Residuals





Residuals of additive model are closer to white noise compared to multiplicative model. However, this contradicts AICc results but is in agreement with raw data plot.

ETS - Model summary

- The model selected by automated approach is “MNM” while that chosen by manual approach is “ANA”.
- The AICc value of the former is lower than the latter, suggesting that “MNM” is probably the better model (despite no clear multiplicative seasonality).
- However, all other training dataset error measures for the “ANA” are lower than “MNM”.

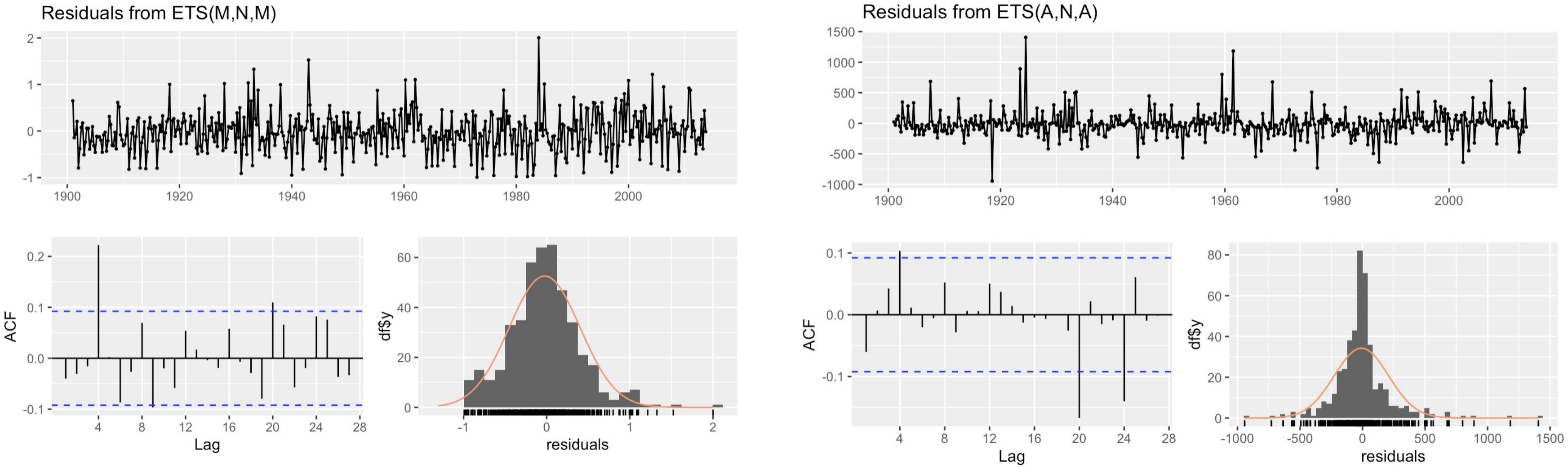
```
     AIC     AICc      BIC
7259.597 7259.849 7288.392

Training set error measures:
                  ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 38.0828 229.6326 137.7357 -98.72692 118.0903 0.7827869 -0.05013145
```

“MNM” model

```
     AIC     AICc      BIC
7641.069 7641.321 7669.864

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set -9.004383 217.0617 136.9752 -47.25703 97.71366 0.7784644 -0.06022874
```

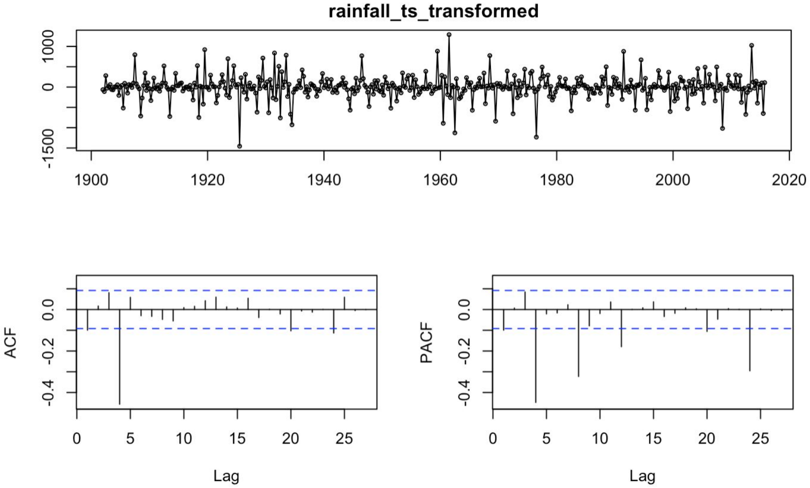
“ANA” model

ETS - Residual analysis



Residuals for both are autocorrelated as per Ljung-Box test but it looks like those from the “ANA” model has a closer resemblance to white noise.

ARIMA



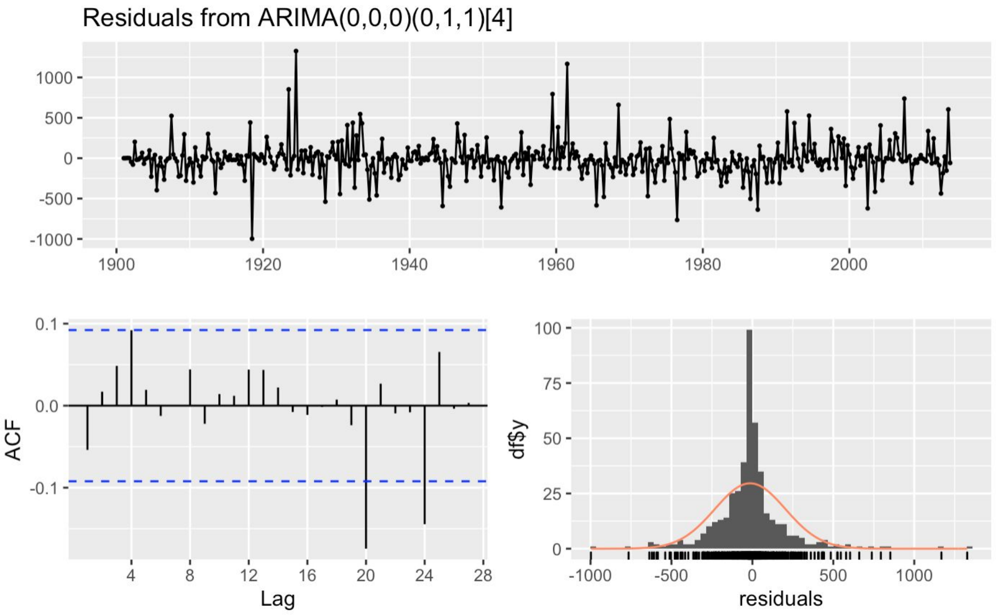
- The PACF decays exponentially with most significant lags at the seasonal lags of 4, 8, etc while the ACF drops off abruptly post the seasonal lag of 4.
- This points to the fact that the stationary process is most probably an ARIMA(0,0,0)(0,1,1).

ARIMA

- Non-seasonal auto.arima (though irrelevant here) gave ARIMA(4,0,0) as the best model but residuals don't resemble white noise. (AICc = 6357.66)
- SARIMA model via auto.arima gave ARIMA(1,0,0)(2,1,0)[4]. From Ljung-Box test, the residuals seem to not be autocorrelated. (AICc = 6194.13)
- Based on visual analysis, model was run for ARIMA(0,0,0)(0,1,1)[4]. Residuals are uncorrelated. (AICc = 6115.05)
- From the AICc values of the above models, it can be observed that the model by visual analysis i.e ARIMA(0,0,0)(0,1,1)[4] performs the best.
- Experimented with other combinations but none performed as well as the best model above.

ARIMA - Best model

```
Series: train_ts
ARIMA(0,0,0)(0,1,1)[4]

Coefficients:
         sma1
      -0.9652
s.e.   0.0156

sigma^2 = 48102:  log likelihood = -3055.51
AIC=6115.02   AICc=6115.05   BIC=6123.23

        Ljung-Box test

data:  Residuals from ARIMA(0,0,0)(0,1,1)[4]
Q* = 7.5586, df = 7, p-value = 0.3731

Model df: 1.   Total lags used: 8
```



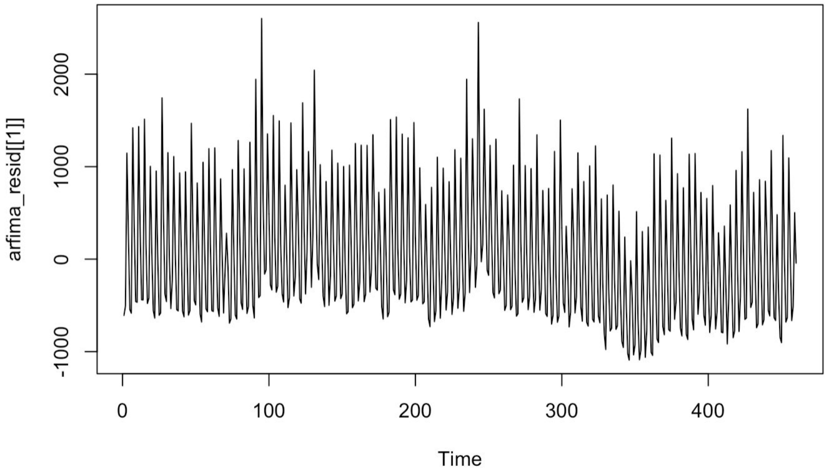
ARFIMA - Model summary

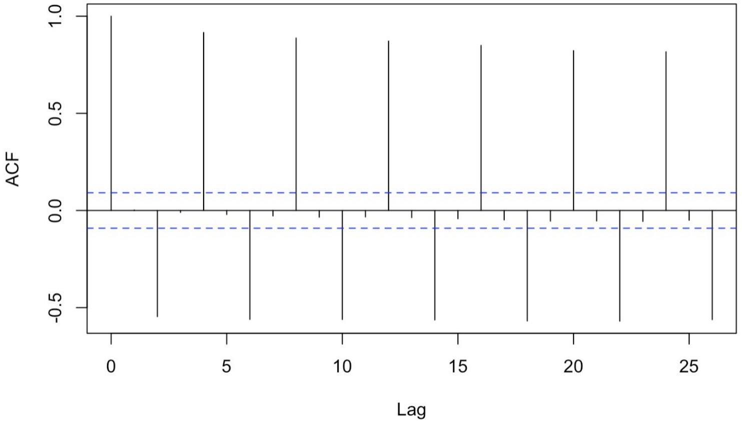
```
arfima(z = rainfall_ts)


Mode 1 Coefficients:
               Estimate  Std. Error Th. Std. Err.  z-value   Pr(>|z|)
d.f          -0.3814752   0.0284851     0.0363593 -13.3921 < 2.22e-16 ***
Fitted mean 731.2155040   4.4821870            NA 163.1381 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
sigma^2 estimated as 522793; Log-likelihood = -3027.89; AIC = 6061.78; BIC = 6074.18
```

The above FARIMA(0,-0.381,0) process exhibits intermediate memory (anti-persistence).

Fractional difference value, **d = -0.38**

ARFIMA - Residual analysis





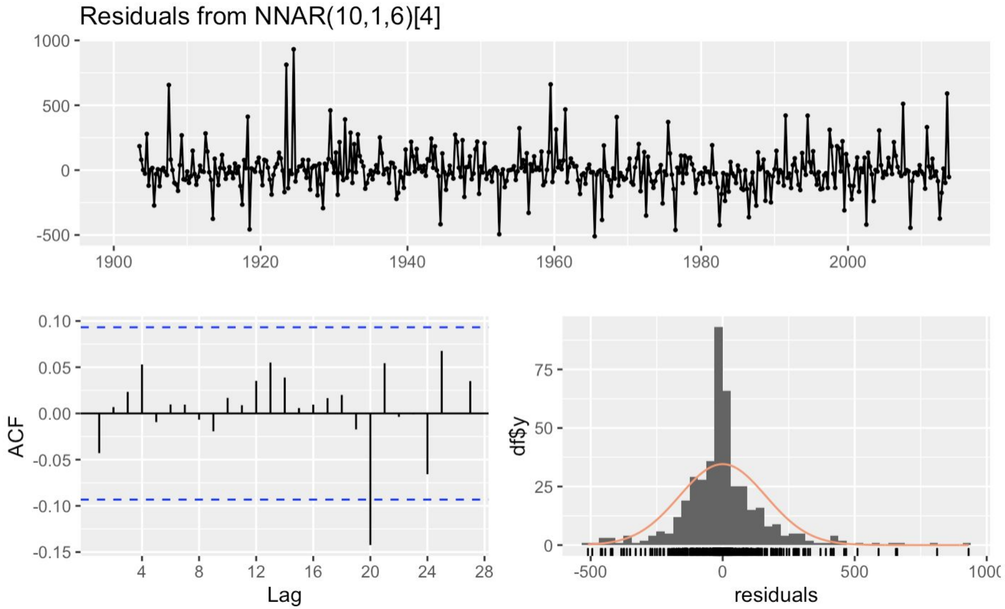
ACF of residuals DO NOT resemble white noise and ARFIMA is probably not the best option here.

Neural Net - Model summary

Meteorological data like rainfall tends to exhibit non-linearity and Neural Networks can capture this.

```
Series: train_ts
Model:  NNAR(10,1,6)[4]
Call:   nnetar(y = train_ts, p = 10, repeats = 30)

Average of 30 networks, each of which is
a 10-6-1 network with 73 weights
options were - linear output units
```

Neural Net - Residual analysis



Residuals are uncorrelated until lag 20, which is a good enough resemblance to white noise.

Model comparison and selection

Evaluation on MSE and MAPE

model_type <chr>	mse <dbl>	mape <dbl>
snaive	171746.82	101.47352
HW (additive)	35658.62	$\boxed{17.42961}$
ETS (ANA)	36275.01	40.09726
SARIMA	$\boxed{34411.94}$	42.86603
Neural Net	35088.21	45.45511

MSE was chosen as a metric due to its ability to penalise outliers, while MAPE has no such penalisation.

SARIMA model performs best on MSE while Holt-Winters’ performs best on MAPE.

Concluding remarks

- The SARIMA model performs best on MSE while Holt-Winters’ does best on MAPE.
- Next steps
  - Model validation using Cross-validation
  - ARMAX and VAR analysis using a suitable predictor

Thank you!