**ST 597 Statistical Consulting - Final Report**

Grant Corkren

Department of Information Systems, Statistics, and Management Science

The University of Alabama, AL, USA

April 2023

**1 Data Management and Exploratory Analysis**

The dataset used had water level, precipitation, and date. It was modified to include a lag of rainfall from the previous day’s precipitation to ten days before the precipitation amount. Exploring the data, we can see that there seems to be a seasonal component in the time series, with the water level decreasing in the winter and rising back up in the spring to a consistent level. The amount of precipitation does not seem to correlate with the time series’ seasonal component. The correlation of rainfall and the lags of precipitation showed a strong correlation with water level. The time series plots also show an inconsistent mean and variance which will need to be addressed before an Arima model is performed. The series was then split into a training series and a testing series. The training series starts on January 1st, 2011 and ends on the last day of 2019. The validation series starts from January 1st, 2020 to January 20th, 2023.

**3 Data Modeling**

The models that were applied were the seasonal naïve, Tbats, automatic Arima with rainfall as a regressor, and high-frequency versions of both the exponential smoothing and Arima models. Both the standard exponential smoothing and Arima models cannot manage a seasonality period greater than 24 and 350 respectively. Additionally, hidden Markov models were also tried with various numbers of states and with rainfall as a regressor.

**3.1 Seasonal Naïve Model**

The seasonal naïve model the last season of the time series and the forecast is based only on that. This model seems like a good model comparing the forecast with the actual values, but it has a mean absolute percent error (MAPE) of 5.26, higher than most other valid models.

**3.2 Tbats Model**

The Tbats model stands for trigonometric seasonality(T), box cox transformation(B), Arma error(A), trend components(T), and seasonal components(S). Models are made with a combination of T, B, A, T, and S. The final model produced is chosen by AIC. The mean absolute percent error is 4.25.

**3.3 Arima models**

The Arima models used are the automatic Arima model and a seasonal adjusted Arima model. A non-seasonal Arima model was used to check if the lags could account for the seasonality in the water level since there is a correlation between the two.

**3.3.1 Auto Arima model**

The auto Arima function fits Arima numerous models with different values of autoregressive model component, number of differences, and moving average component along with its seasonal counterparts. This Arima model will not account for seasonality since the period of seasonality is out of range at a frequency of 365 while the automatic Arima model can only manage a frequency of 350. The auto Arima function then uses the Bayesian information criterion to output the best model. The model was run with all components having a set maximum of ten, stationary set as true, and lags set as a regressor. The resulting Arima is poor based on a plot of forecasted and actual water levels. The model does not account for the seasonality in the time series but shows that precipitation is not a great predictor of water level and further shows that seasonal effect is the primary component of the time series.

**3.3.2 seasonal adjusted Arima model**

The seasonally adjusted model subtracts the seasonality for the time series and models an Arima model and then adds the seasonal back. Both with and without regressor were tried with lags one through three. The model without regressors had a lower mean absolute percent error equal to 4.82 and with regressor had a MAPE of 4.92. The model with regressors has components of an autoregressive order of three, a difference equal to one, an order of two for moving average, and three external regressors. The without regressor model has an autoregressive order of three, a difference equal to one, and an order of five for the moving average.

**3.3.3 Hidden Markov Model**

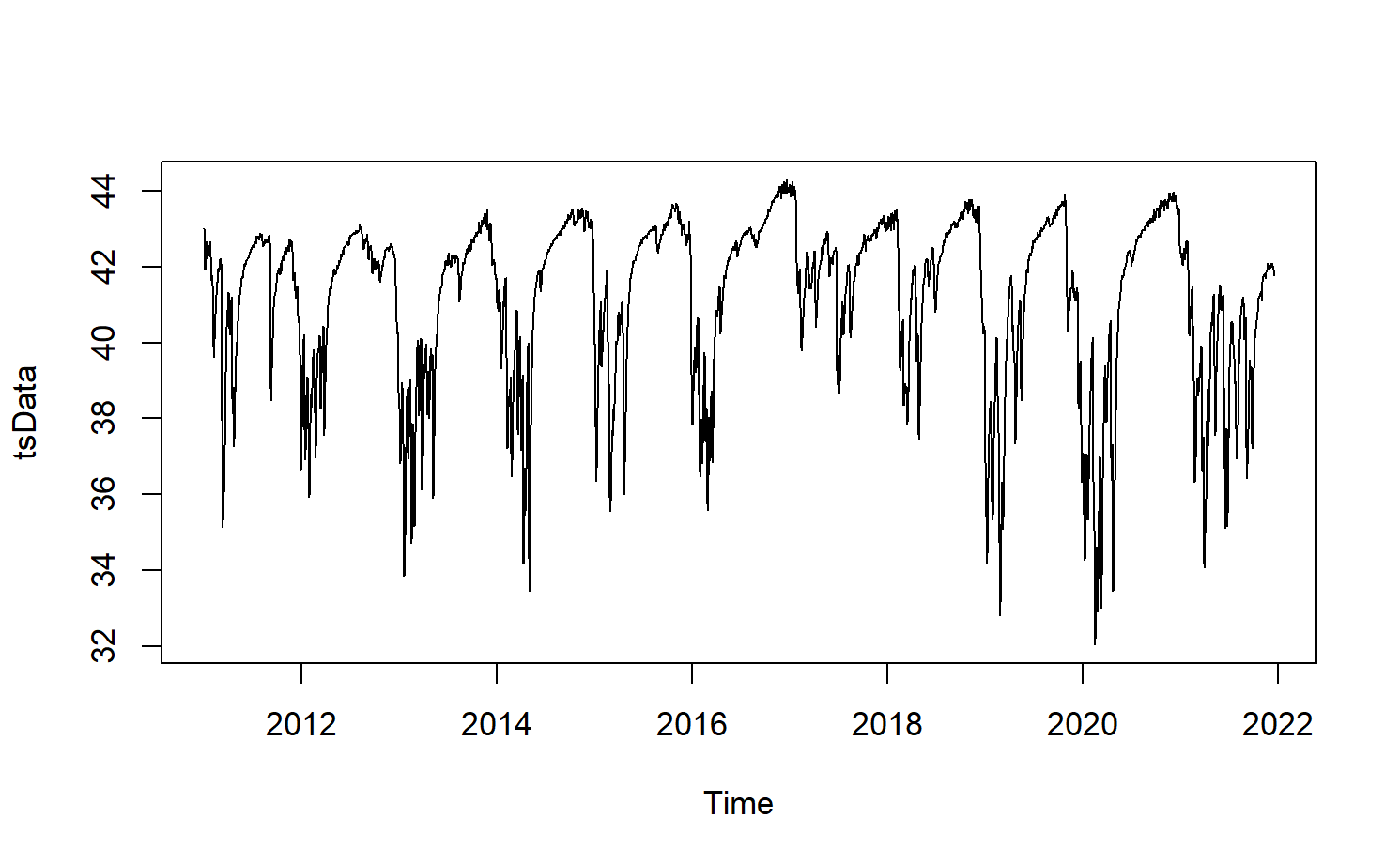
Hidden Markov models were run with two hidden states, lags one through three as a regressor, and an autoregressive order of two and three. The autoregressive order of three had a slightly higher R-squared with an R-square of .98 in state one and an R-squared of .99 in state two. The library package used does not have the capability of forecasting with the hidden Markov models, so the resulting model could be overfitted. Overfitting could be occurring in this model since there is no check of the error of the predicted and validation dataset, but the number of observations, thirty-three hundred observations, should diminish this effect. Another hidden Markov model was used to check if the regressors improve the R-squared of the model and the resulting R-squared shows slight change.

**4 Results**

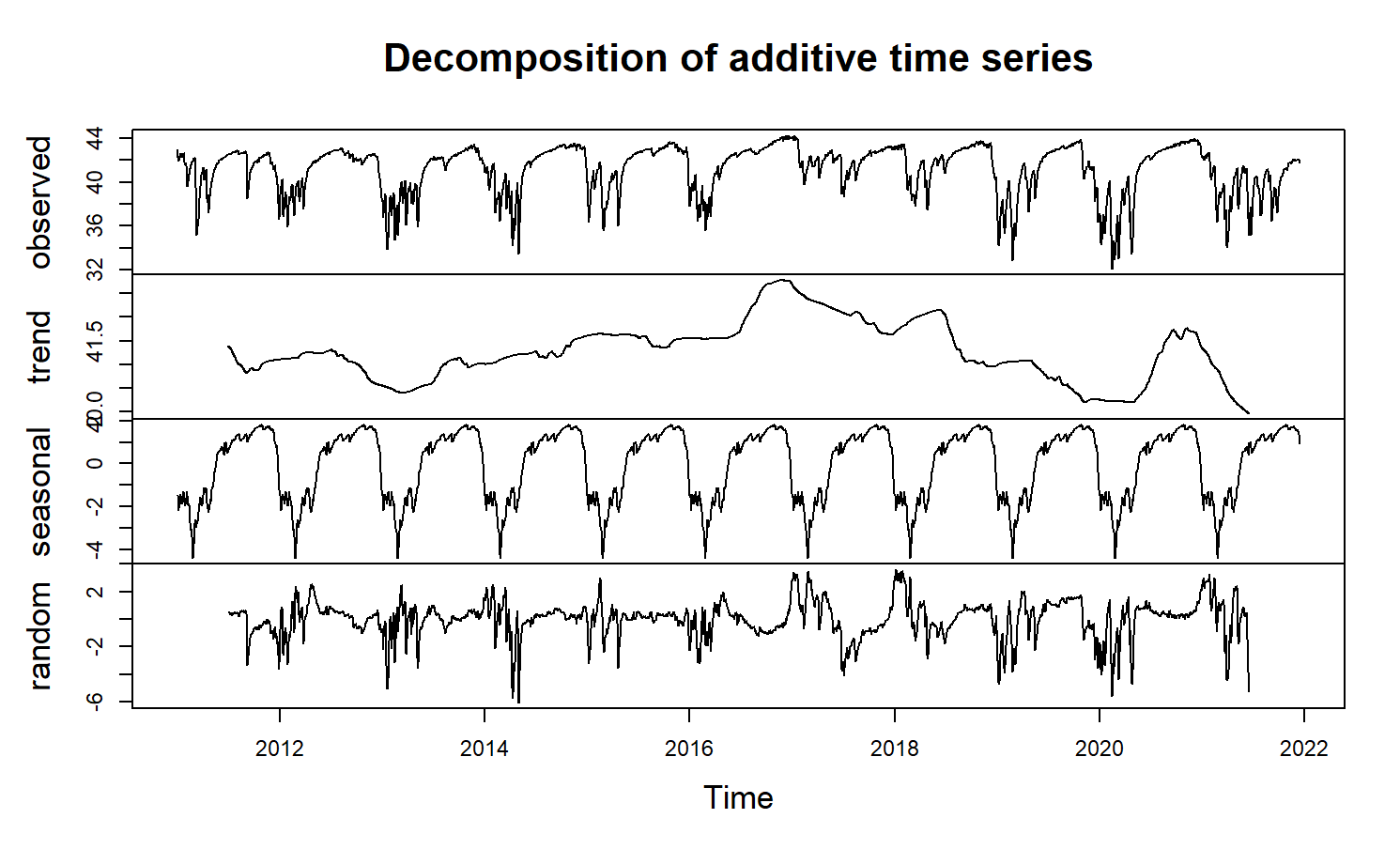
The time series is primarily seasonal with the water level decreasing in the winter and rising to a consistent level in the spring remaining level to the fall. Results from the automatic Arima model show that precipitation does not explain the variation in the data very well. The results from the hidden Markov models also support this claim. These models show that the main component of variation comes from the seasonal effect. There is no trend, and the time series can be forecasted with a good degree of accuracy well into the future by accounting for the seasonal effect.

**Appendix**

ts.plot()



decomposition



Time series plot of precipitation A picture containing diagram, line, screenshot, sketch

Description automatically generated

Seasonal naïve model

## ME RMSE MAE MPE MAPE MASE  
## Training set -0.0860974 1.935309 1.299346 -0.3486584 3.253753 1.000000  
## Test set -0.3527842 2.734553 2.064369 -1.1032300 5.260840 1.588775  
## ACF1 Theil's U  
## Training set 0.9772894 NA  
## Test set 0.9845821 7.140577

A graph with red and blue lines

Description automatically generated with low confidence

exponential smoothing

## ME RMSE MAE MPE MAPE MASE  
## Training set -0.0002320215 0.2042334 0.1097034 0.002086648 0.2770253 0.08442973  
## Test set 2.6963604665 3.7152937 3.2751505 6.294195738 7.9603286 2.52061415  
## ACF1 Theil's U  
## Training set 0.2458509 NA  
## Test set 0.9888071 9.012018

A graph with blue lines

Description automatically generated with low confidence

Tbats

## ME RMSE MAE MPE MAPE MASE  
## Training set -0.001714109 0.179477 0.09859464 -0.005413387 0.2465479 0.07588019  
## Test set 0.340072898 2.033184 1.68532644 0.570053603 4.2541349 1.29705723  
## ACF1 Theil's U  
## Training set 0.008342004 NA  
## Test set 0.982653132 5.304137

A graph with blue and red lines

Description automatically generated with low confidence

stl exponential smoothing

ME RMSE MAE MPE MAPE MASE ACF1

Training set 0.001165068 0.1910643 0.1096602 0.005499473 0.2759329 0.08439644 0.2417248

Test set -0.125676852 1.8939248 1.4395844 -0.568690174 3.6956410 1.10792981 0.9794079

Theil's U

Training set NA

Test set 5.076345

ETS(A,Ad,N)

Call:

ets(y = x, model = etsmodel, allow.multiplicative.trend = allow.multiplicative.trend)

Smoothing parameters:

alpha = 0.9999

beta = 0.9999

phi = 0.8

Initial states:

l = 44.2644

b = -0.4734

sigma: 0.1906

stl arima model with regressor

## ME RMSE MAE MPE MAPE  
## Training set -0.001614946 0.1603521 0.09835716 -0.004982487 0.2447377  
## Test set 1.324088799 2.3210417 2.01256405 3.042692736 4.9692305  
## MASE ACF1 Theil's U  
## Training set 0.07569742 0.0006090183 NA  
## Test set 1.54890512 0.9795102863 5.779347

Series: x

Regression with ARIMA(3,1,2) errors

Coefficients:

ar1 ar2 ar3 ma1 ma2 lagged1 lagged2 lagged3

1.4294 -0.4736 -0.0393 -0.3925 -0.5264 -0.0715 -0.1163 -0.0541

s.e. 0.0420 0.0687 0.0349 0.0375 0.0308 0.0052 0.0079 0.0053

sigma^2 = 0.02578: log likelihood = 1349.54

AIC=-2681.07 AICc=-2681.02 BIC=-2626.2

A picture containing text, font, screenshot, plot

Description automatically generated

Stl arima model without regressor

## ME RMSE MAE MPE MAPE  
## Training set -0.001760413 0.1652415 0.09745356 -0.005264304 0.2430827  
## Test set 1.217379679 2.2588815 1.95023735 2.776492046 4.8240988  
## MASE ACF1 Theil's U  
## Training set 0.07500199 0.0001904716 NA  
## Test set 1.50093738 0.9794342206 5.645241

Series: x

ARIMA(3,1,5)

Coefficients:

ar1 ar2 ar3 ma1 ma2 ma3 ma4 ma5

0.4645 0.6929 -0.3304 0.5898 -0.7078 -0.4763 -0.1371 -0.1171

s.e. 0.0732 0.0767 0.0615 0.0736 0.0664 0.0502 0.0598 0.0286

sigma^2 = 0.02738: log likelihood = 1250.87

AIC=-2483.73 AICc=-2483.68 BIC=-2428.86

A picture containing text, font, screenshot, plot

Description automatically generated

Stl ets model

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.001165068 0.1910643 0.1096602 0.005499473 0.2759329 0.08439644  
## Test set -0.125676852 1.8939248 1.4395844 -0.568690174 3.6956410 1.10792981  
## ACF1 Theil's U  
## Training set 0.2417248 NA  
## Test set 0.9794079 5.076345

A graph with blue and red lines

Description automatically generated with low confidencef