Dataset 3: Executive Report

The goal of the project is to fit a model to a simulated data set, and to use the model to make forecasts for the next 13 values. There is an apparent downward trend in the data, which required a transformation of the data in order to conduct the analysis. The fitted model equation in terms of the original data is:

$$X_t = 1.6938X_{t-1} + 0.1547X_{t-2} - 1.6903X_{t-3} + 0.8418X_{t-4} + Z_t + 0.9007Z_{t-1}$$

The forecasts and associated prediction intervals can be found in Table 3.1 and are graphically represented in Figure 3.1.

Time	Prediction	95% Prediction Interval: Lower Bound	95% Prediction Interval: Upper Bound
585	-10408.71	-10503.5	-10313.921
586	-10407.08	-10670.66	-10143.497
587	-10479.77	-10985.19	-9974.348
588	-10440.94	-11231.79	-9650.084
589	-10477.05	-11574.65	-9379.457
590	-10407.97	-11803.2	-9012.751
591	-10423.38	-12089.99	-8756.765
592	-10345.05	-12240.48	-8449.609
593	-10361.92	-12441.26	-8282.578
594	-10294.19	-12512.25	-8076.139
595	-10327.46	-12647.8	-8007.117
596	-10278.87	-12672.64	-7885.108
597	-10330.4	-12779.26	-7881.532

Table 3.1: Forecasts 13 steps ahead, and accompanying bounds for 95% prediction intervals.

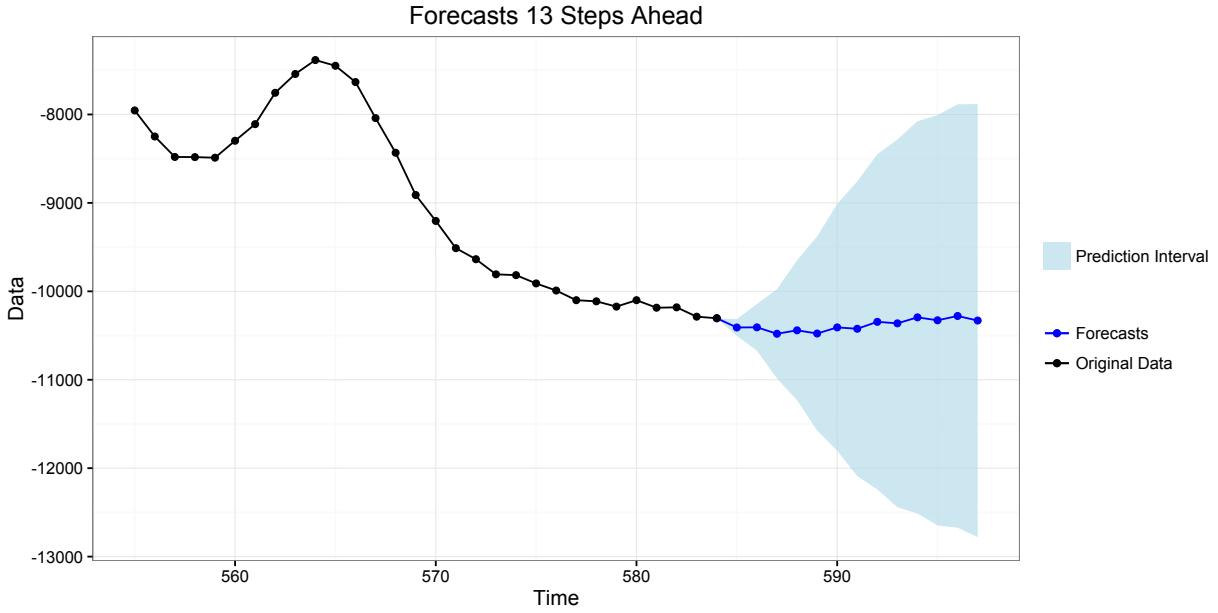

Figure 3.1: Plot of forecasts 13 steps ahead and 95% prediction interval.

Dataset 3: Technical Appendix

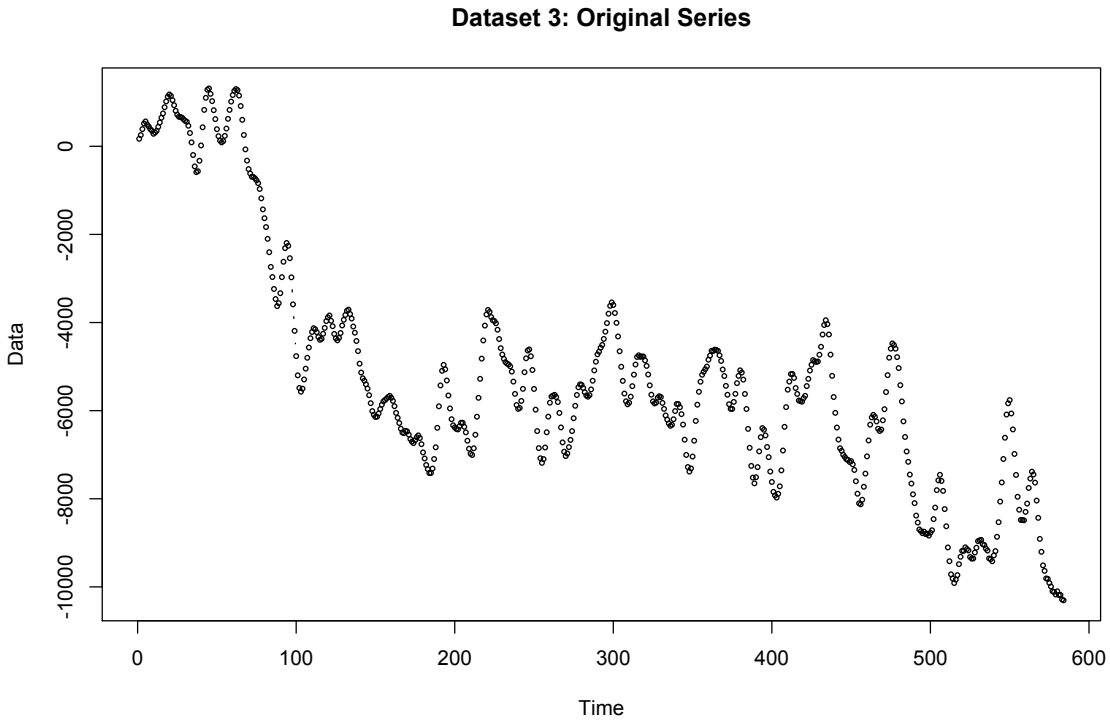

Figure 3.2: Plot of original series showing downward trend. Suggests differencing required to achieve stationarity

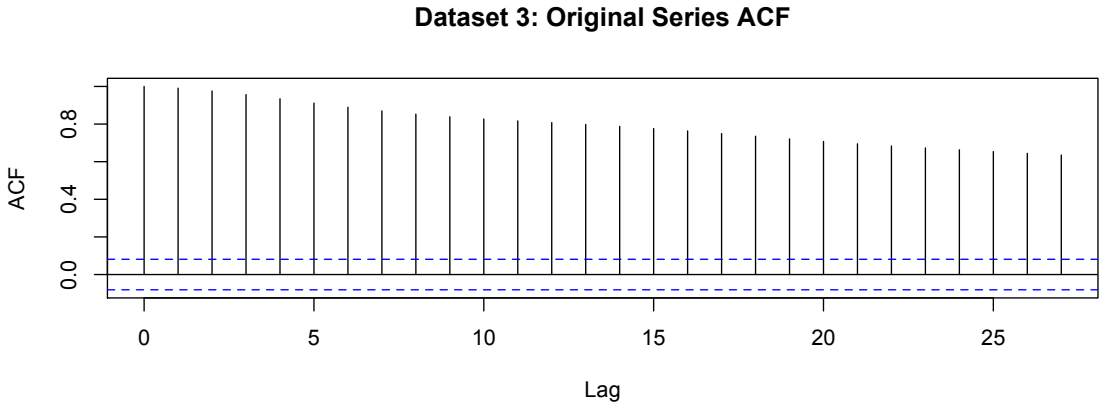

Figure 3.3: Plot of autocorrelation function showing slow decay indicative of non-stationarity

In order to achieve stationarity, the original series was differenced once at lag 1. The differencing removed the downward trend, and the resulting series appears to be stationary. Evidence of stationarity can be seen in the top left cell of Figure 3.4, which shows relatively fast sinusoidal decay in the sample ACF plot of the differenced series. Thus, ARIMA(p,1,q) models were deemed appropriate.

The sample ACF, sample PACF, and Periodogram of the lag 1 differenced data can be seen in the top rows of Figure 3.4 and Figure 3.5, respectively. Based on these plots, ARIMA(p,1,q) candidates were considered for $3 \leq p \leq 5$ and $0 \leq q \leq 3$. The candidates were trained on data up to time 490. This time was chosen because there appeared to be similarities in the time plot of the original series between this time and the end of the series for which forecasts need to be made. Table 3.2 lists the candidates as well as several evaluation criteria.

AR	MA	AIC	BIC	Sigma^2	Log Likelihood	Adjusted SS Residual	Significant Coefficients
4	1	5166.29	5191.444	2185.722	-2577.145	2250.143	No
5	1	5167.653	5196.999	2182.774	-2576.826	2256.61	No
3	1	5164.739	5185.701	2187.709	-2577.369	2242.745	Yes
3	2	5166.33	5191.484	2185.984	-2577.165	2250.414	No
5	2	5169.845	5203.384	2183.602	-2576.923	2267.051	No
5	3	5170.161	5207.892	2176.082	-2576.08	2268.878	N/A
4	2	5168.631	5197.977	2187.349	-2577.315	2261.34	N/A
3	3	5169.344	5198.691	2190.1	-2577.672	2264.184	No
4	3	5170.347	5203.886	2186.705	-2577.174	2270.273	No
5	0	5233.647	5258.801	2516.772	-2610.824	2590.951	Yes
4	0	5257.061	5278.023	2652.446	-2623.531	2719.174	No
3	0	5328.226	5344.996	3084.461	-2660.113	3148.855	No

Table 3.2: Candidate model evaluation. Greyed-out rows indicate candidate models that were not adequate.

Of the two remaining candidate models, ARIMA(3,1,1) performed much better at predicting 13 steps ahead (sum of squared prediction error ~27% smaller) than ARIMA(5,1,0). Furthermore, residual diagnostics of ARIMA(5,1,0) showed several significant values in the ACF of the residuals. Residual diagnostics for ARIMA(3,1,1) showed no issues. Thus, ARIMA(3,1,1) is the chosen model. The R output from fitting the model on the entire dataset is included below.

```
Call:
arima(x = p3.data, order = c(3, 1, 1), include.mean = FALSE, method = "ML")

Coefficients:
          ar1     ar2      ar3     ma1
      0.6938  0.8485  -0.8418  0.9007
s.e.  0.0222  0.0092   0.0221  0.0189

sigma^2 estimated as 2246:  log likelihood = -3080.48,  aic = 6170.96
```

Due to the differencing of the data, the coefficients listed in the output above will not match those listed in the model specification at the beginning of this report. An alternative and equivalent model specification using the coefficients from the R output follows.

$$(X_t - X_{t-1}) = 0.6938(X_{t-1} - X_{t-2}) + 0.8485(X_{t-2} - X_{t-3}) - 0.8418(X_{t-3} - X_{t-4}) + Z_t + 0.9007Z_{t-1}$$

All residual diagnostics of the ARIMA(3,1,1) model fit to the entire dataset are satisfactory. The remaining plots illustrate the adequacy of the chosen model.

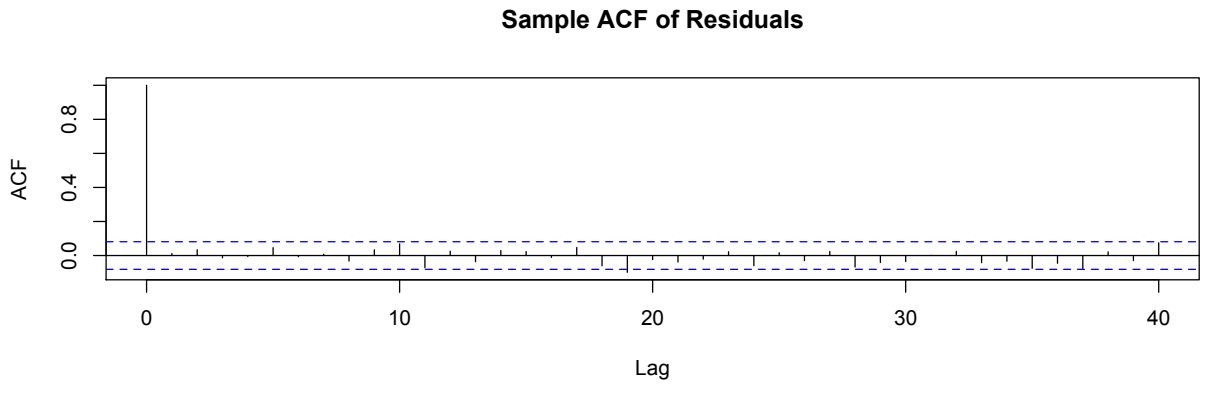

Figure 3.4: Plot of the sample ACF of the residuals. There is one ACF value (lag 19) that is slightly above the boundary for significance; however, the bounds are 95% confidence levels, so 1 in 20 can be expected.

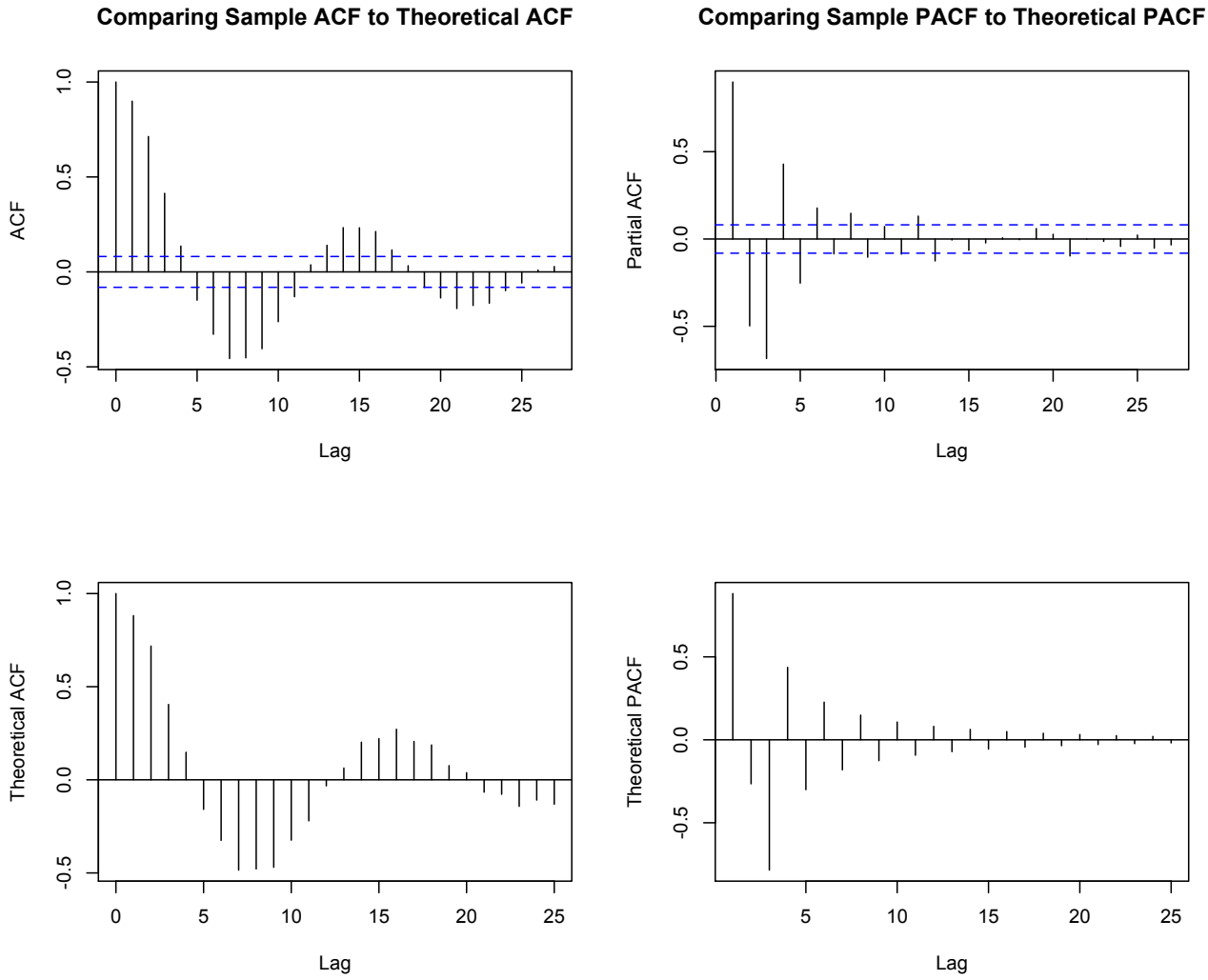

Figure 3.5: The top row plots the sample ACF and sample PACF of the lag-1 differenced data; the bottom row plots the theoretical ACF and theoretical PACF using the coefficients from the ARIMA(3,1,1) fit. The close similarity between the sample and theoretical plots suggests the model is a good fit to the data.

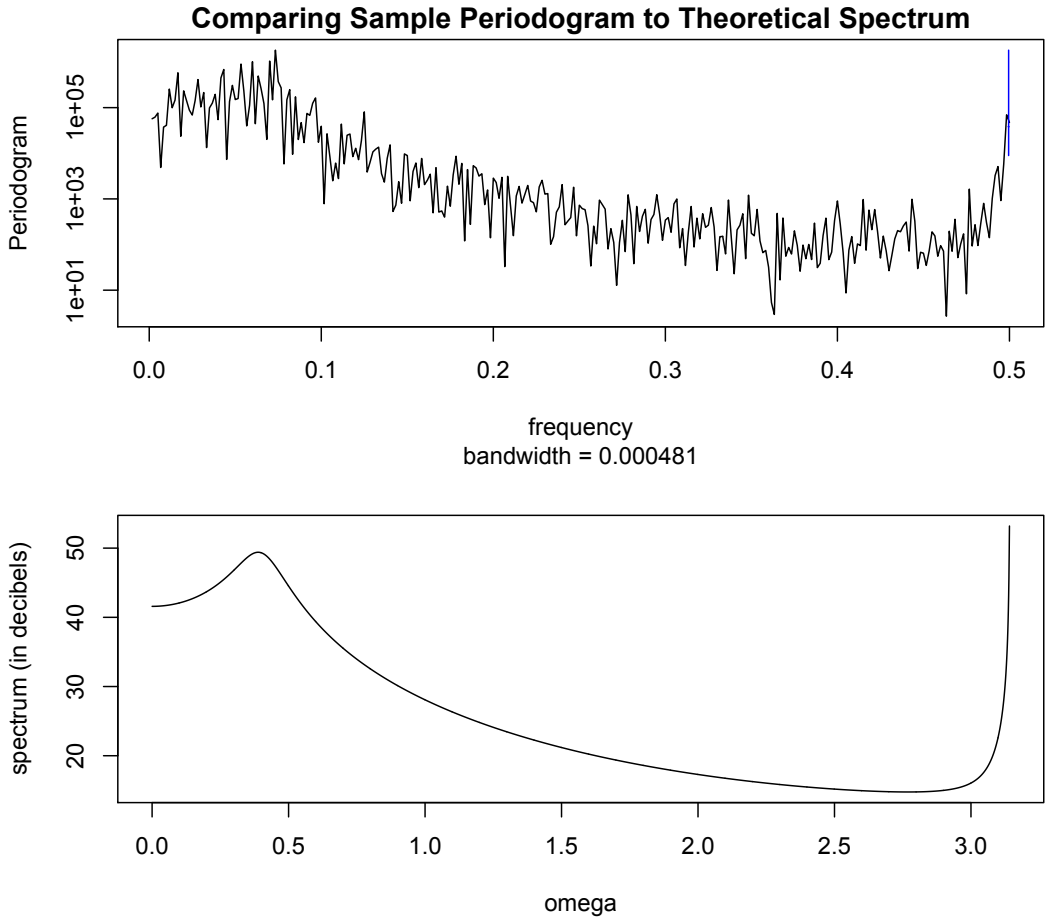

Figure 3.6: The top row plots the periodogram of the lag-1 differenced data; the bottom row plots the theoretical spectrum using the coefficients from the ARIMA(3,1,1) fit. The close similarity between the sample and theoretical plots suggests the model is appropriate.