Dataset 2: Executive Report

The goal of this project is to fit a model to a simulated data set, and to use the model to make forecasts for the next 13 values. The dataset appears to behave nicely in the sense that it is relatively flat and there are no obvious seasonal fluctuations. As a result, no transformations on the original data are required. The fitted model equation is:

$$X_t = 385.2898 - 0.0936X_{t-1} - 0.0904X_{t-2} - 0.7821X_{t-3} + Z_t - 0.1037Z_{t-1} + 0.0807Z_{t-2} - 0.8361Z_{t-3}$$

The forecasts and associated prediction intervals can be found in Table 2.1 and are graphically represented in Figure 2.1.

Time	Prediction	95% Prediction Interval: Lower Bound	95% Prediction Interval: Upper Bound
1501	222.5433	202.2581	242.8285
1502	171.3408	150.6646	192.0171
1503	223.2632	202.5862	243.9403
1504	174.8412	136.3375	213.3449
1505	214.7289	175.7362	253.7216
1506	174.7608	135.705	213.8167
1507	212.7701	166.5914	258.9488
1508	181.6264	134.8708	228.382
1509	212.3671	165.4967	259.2375
1510	182.5756	132.2585	232.8928
1511	206.9449	156.0364	257.8535
1512	183.3127	132.2845	234.341
1513	206.6236	153.8935	259.3537

Table 2.1: Forecasts 13 steps ahead, and accompanying bounds for 95% prediction intervals.

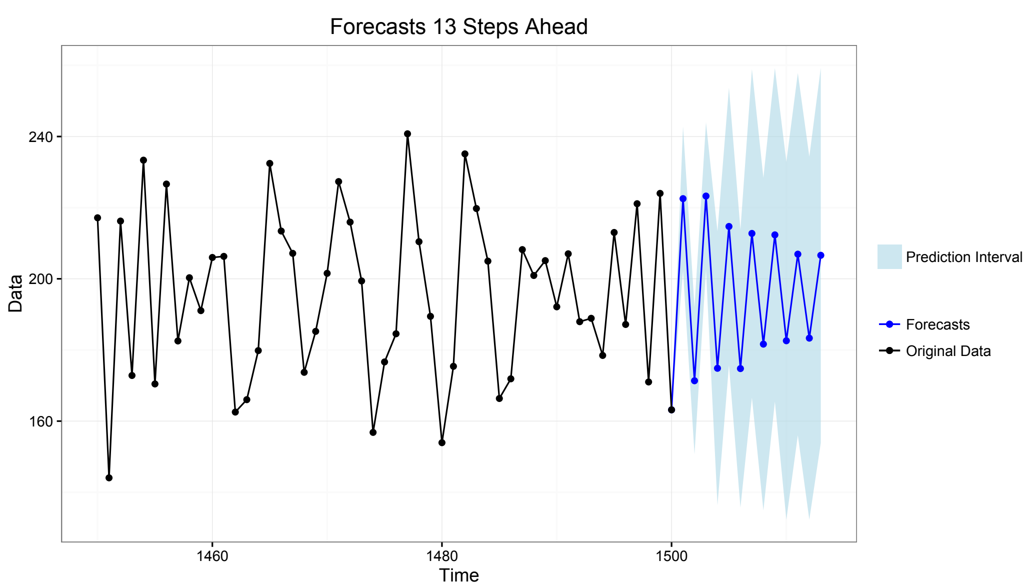

Figure 2.1: Plot of forecasts 13 steps ahead and 95% prediction interval.

Dataset 2: Technical Appendix

Examining the time plot of the data, there appear to be at least two periodicities in the data. Based on this information alone, at the very least an order 3 AR component is required.

The sample ACF and sample PACF, which can be seen in the top row of Figure 2.2, both exhibit interesting patterns. Both show large significant spikes at lag 3, which are immediately followed by decaying patterns repeated every 3 lags. It is difficult to determine where the values are actually significant and where the decay starts, but these patterns are characteristic of ARMA(p,q) processes.

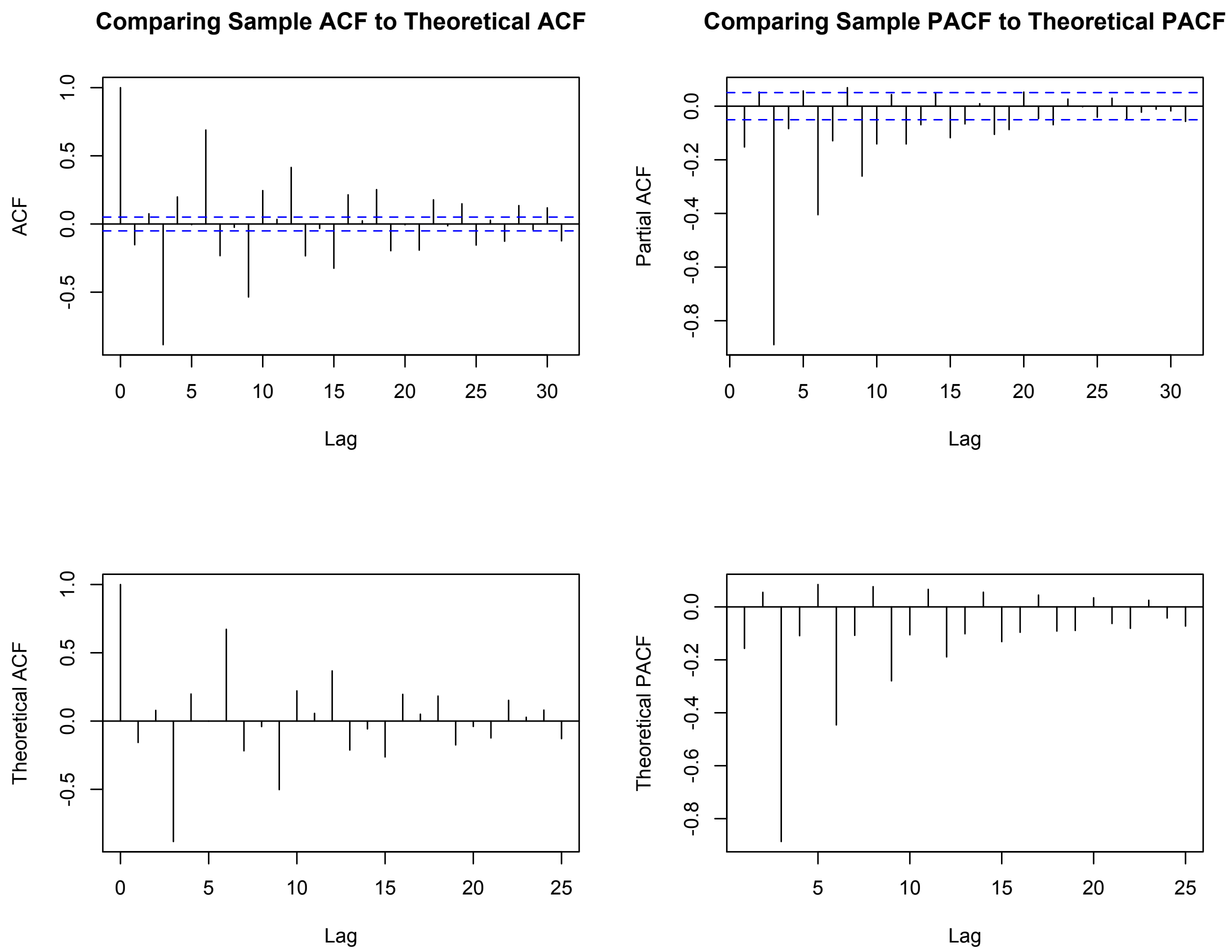

Figure 2.2: The top row plots the sample ACF and PACF of the data; the bottom row plots the theoretical ACF and PACF using the coefficients from the ARMA(3,3) fit. The close similarity between the sample and theoretical plots suggests the model is a good fit to the data.

The Periodogram, which can be seen in the top row of Figure 2.3, clearly suggests an ARMA(3,3). In the middle of the plot there is a sharp peak followed by sharp dip, which implies order 2 for both the AR and MA components. The movement of the periodogram going toward the left boundary (frequency 0.0) is not simply movement away from the peak: it appears to be heading toward another dip at the boundary. Therefore, the MA order is increased by one.

Similarly, the movement going toward the right boundary suggests the AR order should also be increased by one. Thus, an ARMA(3,3) seems appropriate.

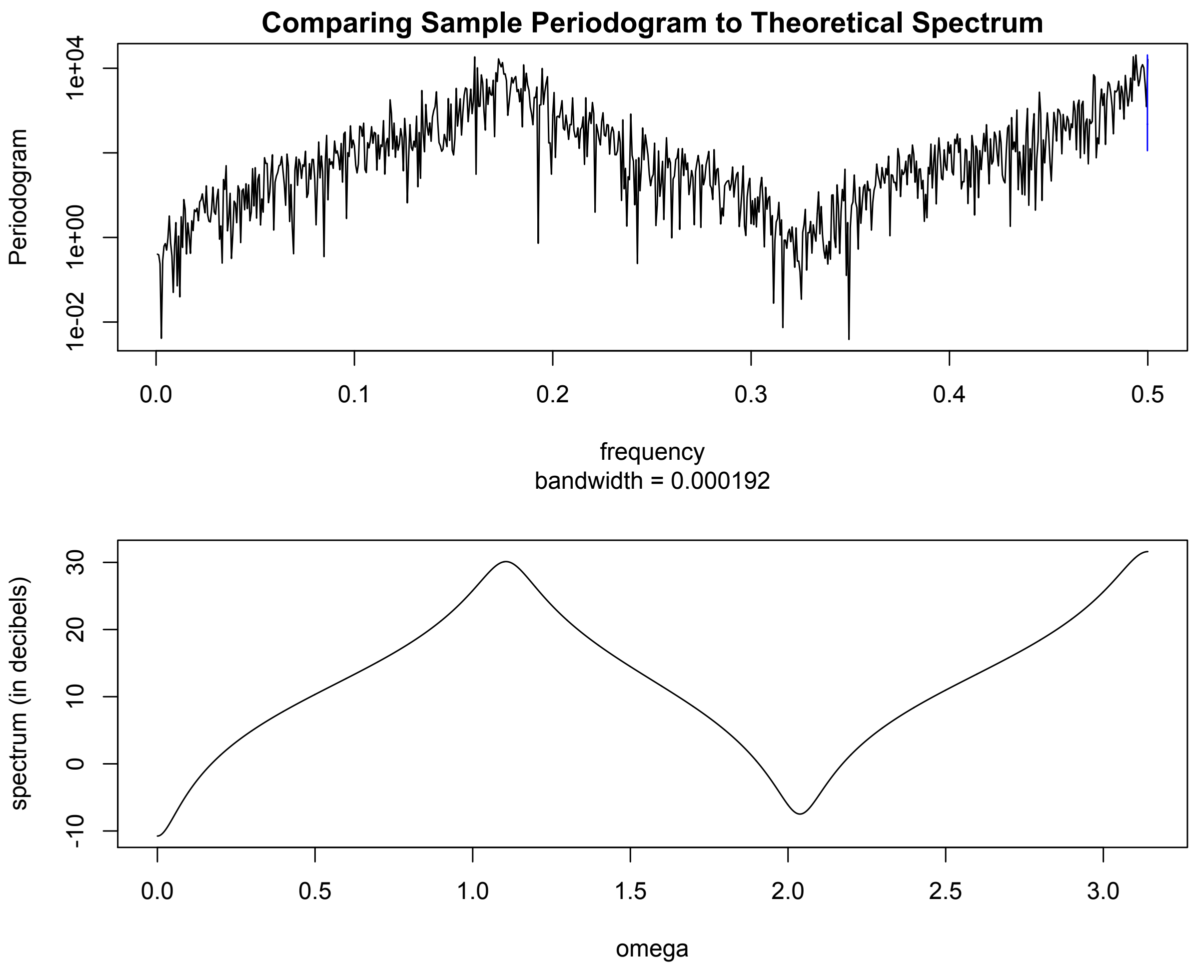

Figure 2.3: The top row plots the periodogram data; the bottom row plots the theoretical spectrum using the coefficients from the ARMA(3,3) fit. The close similarity between the sample and theoretical plots suggests the model is appropriate.

Based on these plots, ARMA(p,q) candidates were considered for $3 \leq p \leq 6$ and $3 \leq q \leq 6$.The candidates were trained on data up to time 1430. 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 2.2 lists the candidates as well as several evaluation criteria. The table is sorted based on overall performance across columns 3 through 7. The overall performance is assessed by sorting each column individually, assigning an integer rank to each model, and then finally summing the rank over all columns. Please note this is simply a heuristic used to get a general idea of which models are performing better than others. The greyed-out rows of Table 2.2 represent candidate models that were excluded from consideration due to the following issues:

1. Statistically insignificant coefficients

2. NA's produced when calculating standard errors of coefficients
3. Convergence warnings when fitting the model

AR	MA	AIC	BIC	Sigma^2	Log Likelihood	Adjusted SS Residual	Significant Coefficients
5	4	10701.45	10754.11	102.01	-5340.727	103.3836	No
6	3	10701.47	10754.12	102.0098	-5340.735	103.3835	No
5	6	10701.63	10764.81	101.7317	-5338.815	103.3947	No
6	4	10702.16	10760.08	101.9177	-5340.082	103.4367	No
3	6	10701.49	10754.14	102.0118	-5340.745	103.3854	No
3	3	10701.07	10737.93	102.4157	-5343.534	103.3553	Yes
4	6	10702.48	10760.4	101.9407	-5340.24	103.46	No
4	5	10701.65	10754.3	102.0246	-5340.824	103.3985	No
6	5	10703.28	10766.46	101.8524	-5339.639	103.5174	N/A
3	4	10703.03	10745.15	102.413	-5343.515	103.4986	No
3	5	10703.23	10750.62	102.2818	-5342.616	103.5123	No
4	3	10703.03	10745.16	102.4132	-5343.516	103.4989	No
5	3	10703.56	10750.95	102.3054	-5342.78	103.5363	No
6	6	10706.36	10774.81	101.932	-5340.18	103.7457	No
5	5	10704.41	10762.33	102.0797	-5341.206	103.6011	N/A
4	4	10705.05	10752.44	102.4147	-5343.526	103.6469	N/A

Table 2.2: Candidate model evaluation. Greyed-out rows indicate candidate models that were not adequate.

All candidate models with AR order or MA order greater than 3 had at least one unnecessary parameter. The estimated coefficients for the ARMA(3,3) candidate were all significant, and the residual diagnostics showed no indication of model inadequacy. Thus, it is the chosen model. The R output from fitting the model on the entire dataset is included below.

```
Call:
arima(x = p2.data.demean, order = c(3, 0, 3), include.mean = FALSE, method
= "ML")

Coefficients:
          ar1      ar2      ar3      ma1     ma2      ma3
      -0.0936  -0.0904  -0.7821  -0.1037  0.0807  -0.8361
s.e.   0.0164   0.0165   0.0165   0.0137  0.0151   0.0161

sigma^2 estimated as 102.9:  log likelihood = -5608.25,  aic = 11230.49
```

The intercept in the model equation is calculated using the following formula:

$$\bar{X}(1 - \hat{\alpha}_1 - \hat{\alpha}_2 - \hat{\alpha}_3)$$

The residual diagnostics for the ARMA(3,3) model fit to the entire dataset do not indicate any issues.