Deposits Dataset: Executive Report

The goal of this project is to fit a model to sedimentary layer thicknesses. The model will be used to make forecasts about future sedimentary thickness values as well as future weather patterns. The fitted model equation below is in terms of the original data:

$$X_t = X_{t-1}{}^{1.2291} X_{t-2}{}^{-0.2291} \exp\{Z_t - 0.8828 Z_{t-1}\}$$

The forecasts and associated prediction intervals can be found in Table 4.1 and are graphically represented in Figure 4.1.

Time	Prediction	95% Prediction Interval: Lower Bound	95% Prediction Interval: Upper Bound
625	10.03353	3.857917	26.09482
626	10.77181	3.917347	29.62004
627	10.94849	3.913274	30.63148
628	10.98937	3.882475	31.1055
629	10.99876	3.845518	31.45813
630	11.00092	3.807702	31.78299
631	11.00141	3.770249	32.1016
632	11.00152	3.733421	32.41893
633	11.00155	3.697261	32.73614
634	11.00155	3.661763	33.05353
635	11.00156	3.626909	33.37118
636	11.00156	3.592679	33.68913
637	11.00156	3.559057	34.00739

Table 4.1: Forecasts 13 steps ahead, and accompanying bounds for 95% prediction intervals.

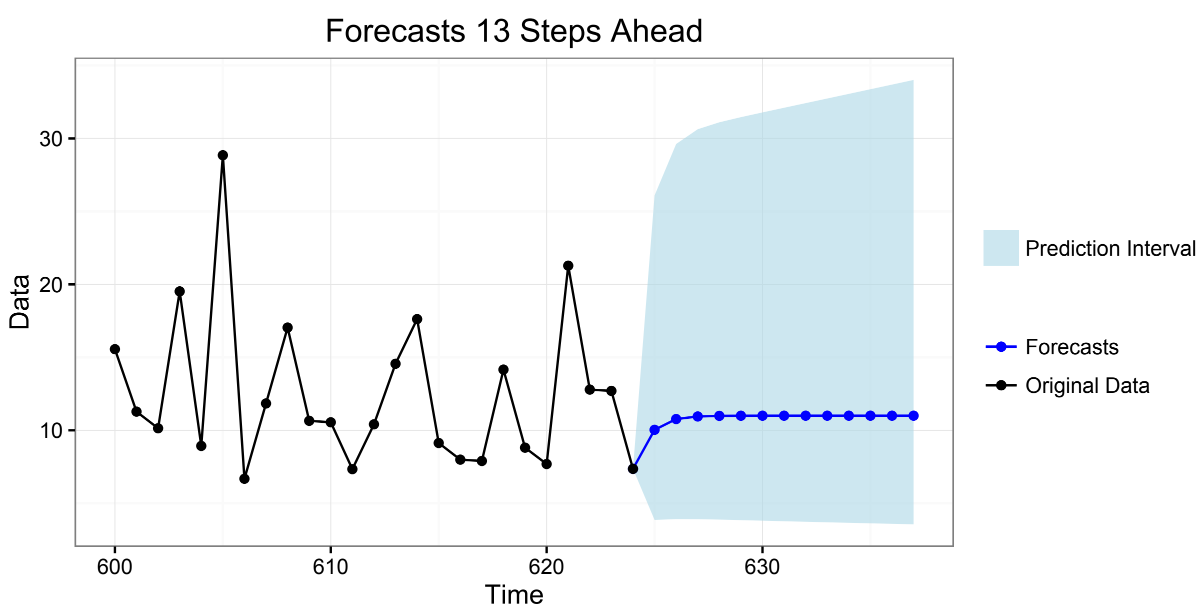

Figure 4.1: Plot of forecasts 13 steps ahead and 95% prediction interval.

Deposits Dataset: Technical Appendix

A. Model Selection and Model Fitting

The indication of non-stationarity revealed by the time plot of the data is also reflected in the sample ACF plot in Figure 4.2, which shows slow decay.

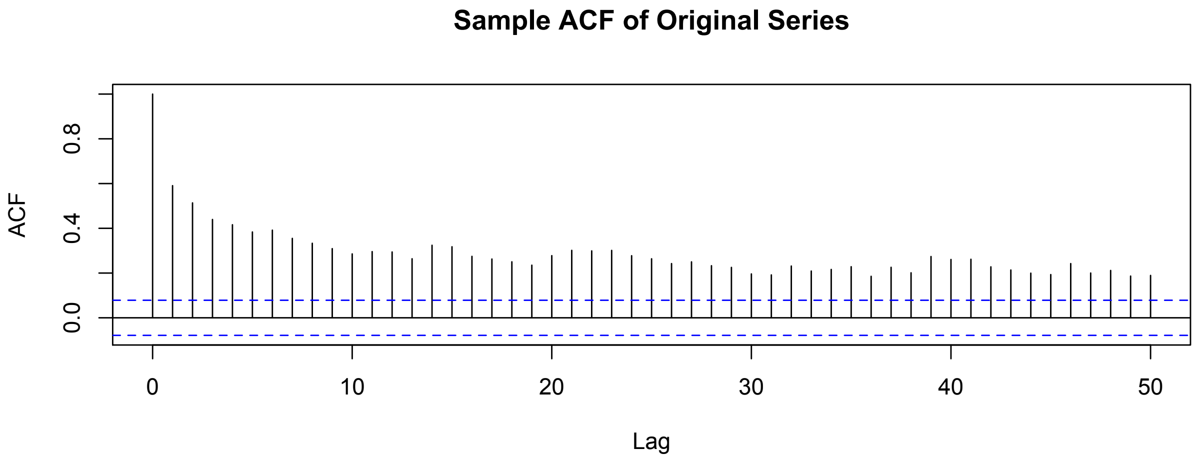

Figure 4.2: Sample ACF of the original series showing slow decay indicative of non-stationarity.

Because there are large spikes in the data and because the data is positive everywhere, a logarithmic transformation seems appropriate. However, applying the log transformation alone did not achieve stationarity: the sample ACF of the log-transformed data was nearly identical to the sample ACF of the original data in Figure 4.2. Therefore, the log-transformed data was transformed again by differencing once at lag 1. The sample ACF and PACF of the resulting stationary series are plotted in Figure 4.3.

The sample ACF cuts off at lag 1, while the sample PACF decays and continues on indefinitely. The periodogram in Figure 4.4 starts at a low point on the left and gradually increases over the entire frequency range. Therefore, together these three plots provide strong evidence for an MA(1). However, for the sake of completeness, ARIMA(p,1,q) candidates were considered for $0 \leq p \leq 2$ and $0 \leq q \leq 2$. The candidates were trained on data up to time 600.

AR	MA	AIC	BIC	Sigma^2	Log Likelihood	Adjusted SS Residual	Significant Coefficients
1	2	826.5815	844.1626	0.2291374	-409.2908	0.2314558	No
1	1	826.2122	839.3979	0.2298098	-410.1061	0.2313548	Yes
2	1	826.8616	844.4426	0.2292611	-409.4308	0.2315808	No
2	2	828.6552	850.6315	0.2291984	-409.3276	0.2323009	No
0	2	829.3611	842.5469	0.2310624	-411.6805	0.2326157	Yes
0	1	843.714	852.5045	0.2375301	-419.857	0.2383259	Yes
2	0	908.045	921.2308	0.2638341	-451.0225	0.2656078	Yes
1	0	941.6503	950.4408	0.2800478	-468.8252	0.280986	Yes
0	0	1039.317	1043.7123	0.3308371	-518.6585	0.3308371	N/A

Table 4.2: Candidate model evaluation. Greyed-out rows indicate candidate models that were not adequate due to unnecessary model terms.

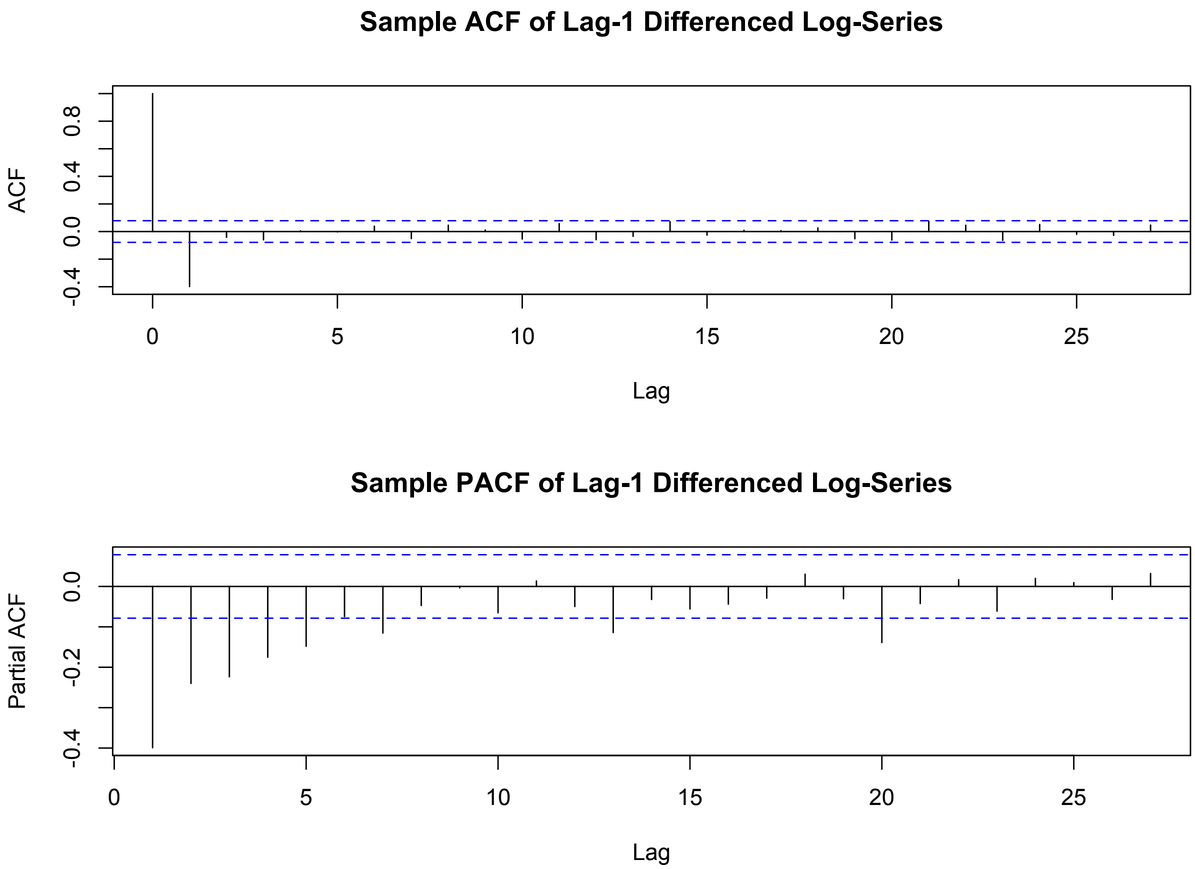

Figure 4.3: Sample ACF and PACF of the lag-1 difference of the log series showing stationarity

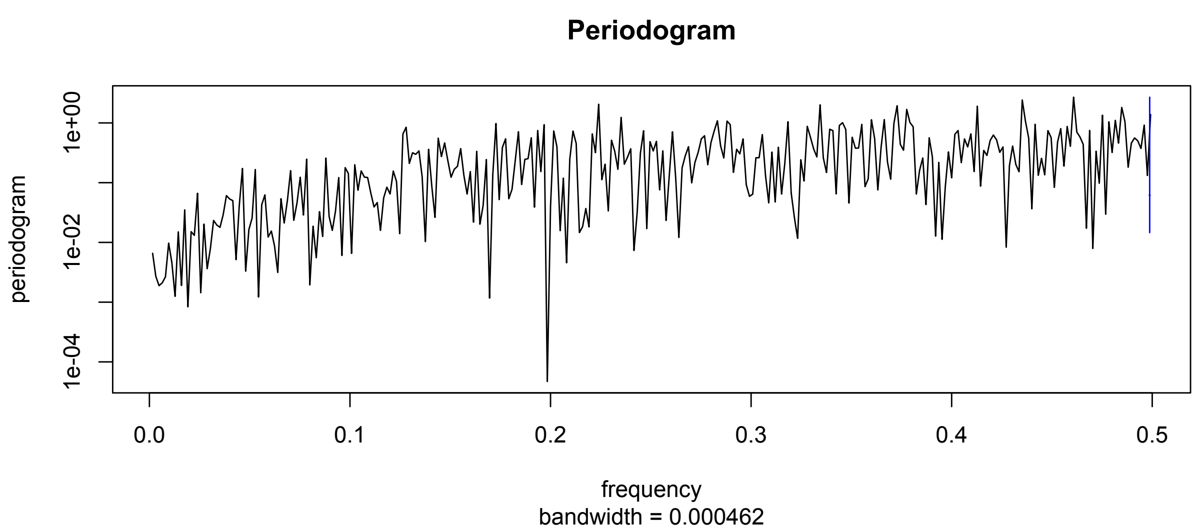

Figure 4.4: Periodogram of the lag-1 difference of the log series.

Residual diagnostics reveal ARIMA(0,1,1) to be inadequate due to several significant residual ACF values. The residual diagnostics for ARIMA(1,1,1) and ARIMA(0,1,2) do not show any issues. Although ARIMA(0,1,2) predicts holdout data slightly better (sum of squared prediction error ~10% less), ARIMA(1,1,1) performs better across all other evaluation criteria. Referring back to the periodogram in Figure 4.4, an argument could be made that the gradual increase calls for an AR component of order 1. The sample ACF, sample PACF, and periodogram do not provide any clear evidence to support an MA component of order 2. Therefore, the chosen model is ARIMA(1,1,1). The R output from fitting the model on the entire dataset is included below.

```
Call:
arima(x = deps.log, order = c(1, 1, 1), include.mean = FALSE, method =
"ML")

Coefficients:
         ar1      ma1
      0.2291  -0.8828
s.e.  0.0523   0.0295

sigma^2 estimated as 0.2284:  log likelihood = -424.56,  aic = 855.13
```

The residual diagnostics for the ARIMA(1,1,1) model fit to the entire dataset do not indicate any issues. The following plots illustrate the appropriateness of the ARIMA(1,1,1) fit to the data.

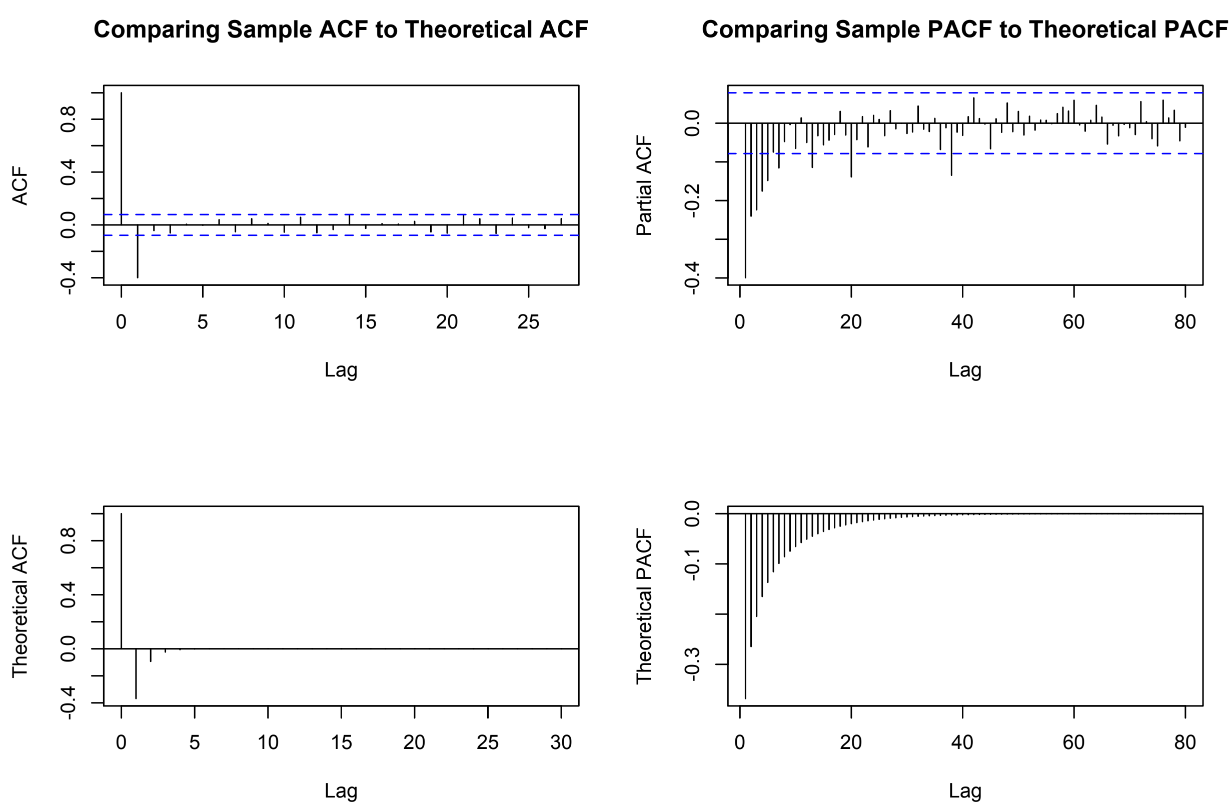

Figure 4.5: The top row plots the sample ACF and sample PACF of the lag-1 difference of the log series; the bottom row plots the theoretical ACF and theoretical PACF using the coefficients from the ARIMA(1,1,1) fit. The close similarity between the sample and theoretical plots suggests the model is a good fit to the data.

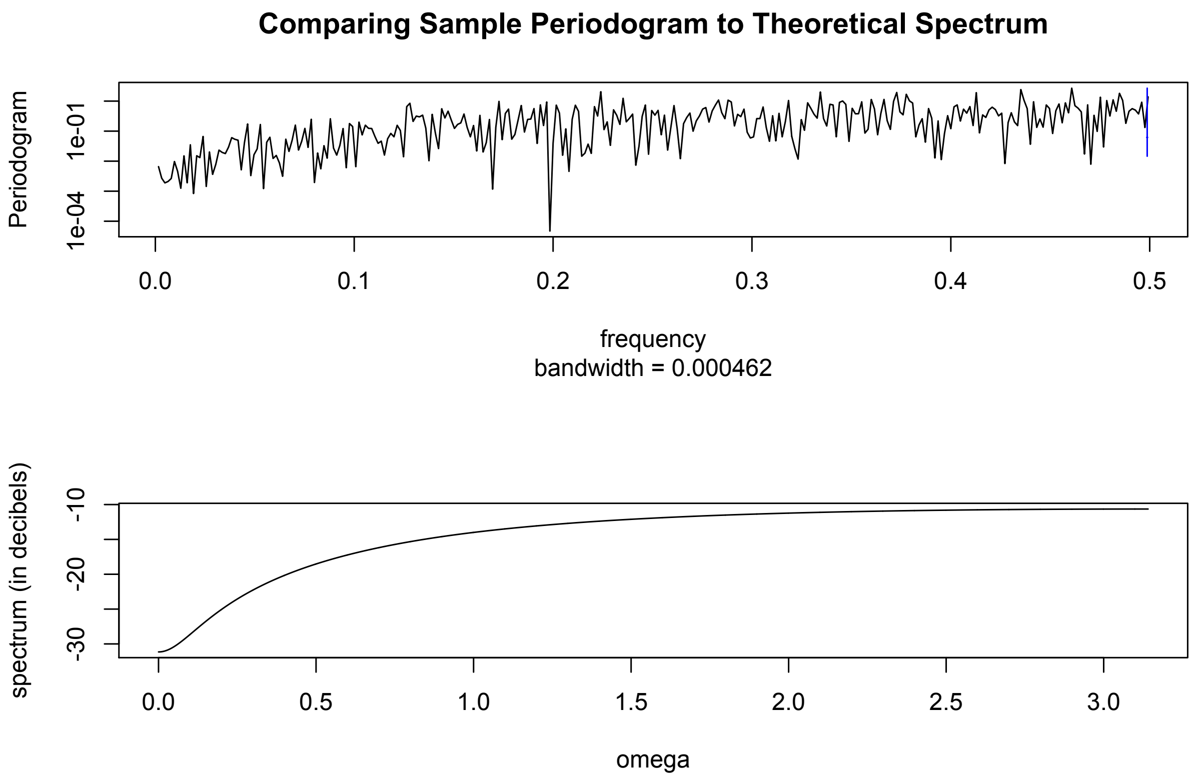

Figure 4.6: The top row plots the periodogram of the lag-1 difference of the log series; the bottom row plots the theoretical spectrum using the coefficients from the ARIMA(1,1,1) fit. The close similarity between the sample and theoretical plots suggests the model is appropriate.

B. Comments on Predictions

Due to the logarithmic transformation and the subsequent differencing at lag 1, the forecasts flatten out quickly after 2 steps ahead. As a result, the forecasts beyond 2 steps ahead will not capture the year-to-year fluctuations present in the original data, but instead will convey information about the likely average temperature over the period. For more accurate predictions of the year-to-year fluctuations, it will be more useful to limit forecasts to 2 steps ahead and to update the model regularly with newly available data.