Electricity Demand Forecasting

Dhairya Jayesh Chheda
MS Data Science - Residential
Indiana University - Bloomington
Bloomington, IN
dhchhed@iu.edu

I. Introduction

Time Series data are collected based on certain periods which have constant values (e.g., daily, weekly, or monthly). Prediction is one of the objectives of the time series analysis by identifying the model from previous data and assuming the current information will also occur in the future. [1]

Forecasting loads on power systems is an integral activity embedded in the long-term system operation planning processes and in the processes of the ongoing control of its operation. The system cannot function without accurate forecasts. This is because electricity cannot be stored in large quantities. The demand must be covered on an ongoing basis with production, with the limitations resulting from the flexibility of the production units and the requirements of the reliability and safety of the system operation. The accuracy of forecasts translates into the costs of production, transmission, and the degree of reliability of electricity supply to consumers. Inflated forecasts lead to the maintenance of too many generating units in order to meet the safety requirements to ensure an adequate margin of reserve capacity. Underestimated forecasts have the opposite effect—too few generating units are planned, which are not able to cover the actual demand. In such a situation, additional units with quick start-up are intervened in the traffic, generating additional operating costs. [2]

Hence, electricity demand forecasting is an essential tool for energy management, maintenance scheduling and investment decisions in the energy markets. Electricity demand for a region depends on economic variables – oil prices, stock prices, exchange rates; demographic circumstances – holidays, population, and most importantly climatic conditions – temperatures, humidity, etc. In this project we want to measure how daily temperatures affect electricity demand for a region. We will also investigate how accurately temperature can be used to forecast the demand in mid-term (8 weeks/2 months).

II. Data

As of 2024, approximately 6.7 million people reside in Victoria, Australia's second most populated state. Most of them, 5 million, live or work in Melbourne, state's capital. During 2020, Australia was among the first to close international borders, followed by a closure of interstate borders. Victoria introduced some of the strictest pandemic-related restrictions on business activity that resulted in a significant portion of its population working from home.

The dataset covers 365 days between 1 January 2020 and 31 December 2020. The data is for operational electricity demand for Victoria, Australia [3]. Operational demand is the demand met by local scheduled generating units, semi-scheduled generating units, non-scheduled intermittent generating units of aggregate capacity larger than 30 MW, and by generation imports to the region. There are 2 measures –

- Demand : Total electricity demand in GW for Victoria, Australia, every day during 2020.
- Temperature : Maximum daily temperature during the day for Victoria, Australia.

The data has 365 timepoints for each day in the year 2020. **'Demand' is the criterion of interest and 'Temperature' is the predictor variable.**

III. Data Analysis

A. Time Domain and Frequency Domain

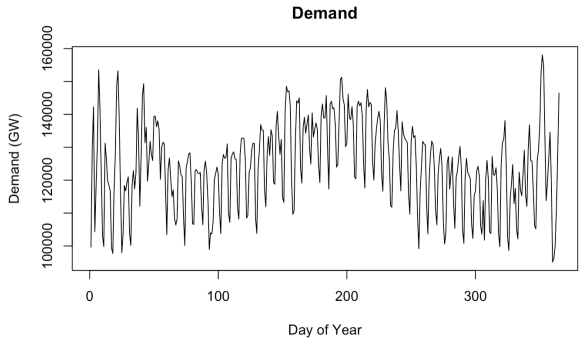

Fig. 1. Daily electricity demand for the year 2020 in Victoria, Australia.

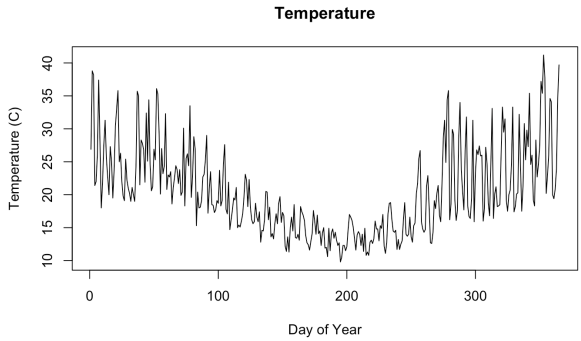

Fig. 2. Maximum daily temperatures for the year 2020 in Victoria, Australia.

Both series in figure 1 and 2 have a quadratic trend with cyclic components. We can therefore de-trend both series using regression and then investigate the noise (residuals) using auto-correlation and spectral analysis.

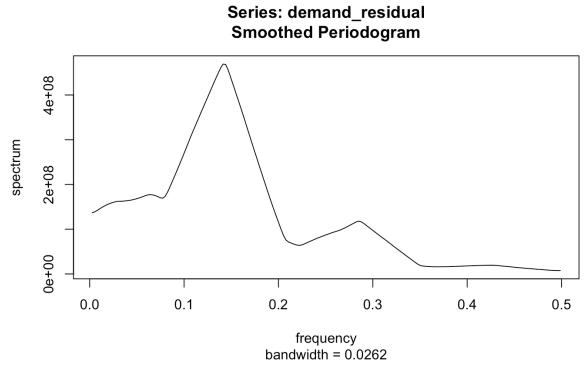

Fig. 3. Periodogram for residuals of Demand after fitting a polynomial of order 6 to de-trend the series.

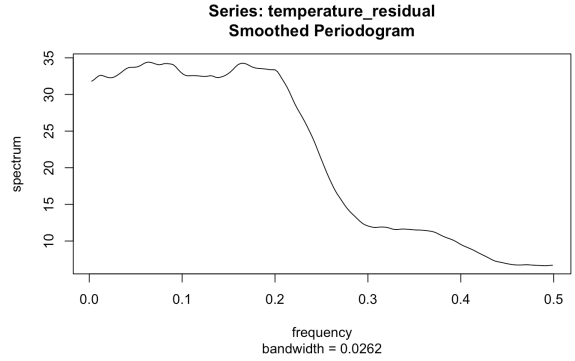

Fig. 4. Periodogram for residuals of Temperature after fitting a polynomial of order 4 to de-trend the series.

From figure 3 and 4, we can see that both 'Demand' and 'Temperature' have systematic cycles. For Demand, we have spikes at f_1 = .1428571 and f_2 = .2857143 denoting some periodicity every 7 days and 3.5 days respectively; and for Temperature, we have a small spike at f_3 = .1706667 denoting some periodicity every 6 days. So, we will try to remove these using regression by fitting sinusoids at these frequencies.

B. Auto-correlation and Spectral Density

To make both series stationary, we used order 6 and 4 polynomials to de-trend demand and temperature respectively. Further, we used sinusoids to remove the systematic cycles. The ACF for Demand (see figure 5) shows a significant spike at lag 0, which is always 1 because a dataset is perfectly correlated with itself at lag 0. The ACF quickly drops off to within the confidence bounds (the blue dashed lines) from lag 3 onwards. This drop-off suggests that there is little to no auto-correlation in the time series at lags greater than 0, indicating that the noise is likely white noise. Similarly, the ACF for Temperature shows a spike at lag 0. The remaining lags are mostly within the confidence bounds, which means

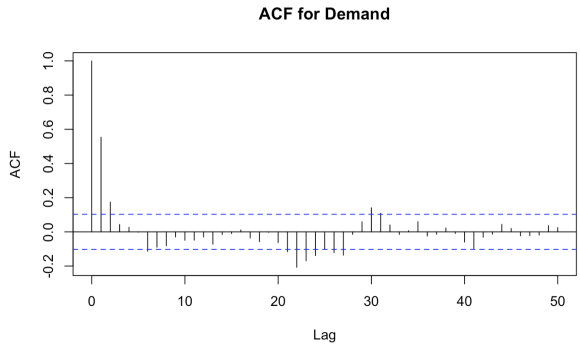

Fig. 5. Auto-correlation of the residuals of Demand after making the series stationary.

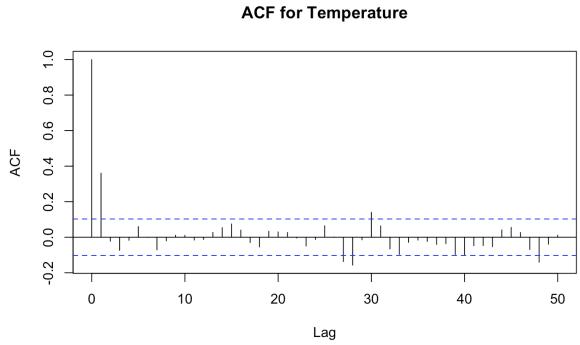

Fig. 6. Auto-correlation of the residuals of Temperature after making the series stationary.

they are not statistically significant. This pattern also indicates that the time series may be white noise, as there is no evidence of auto-correlation at any lag.

Upon looking at the auto-correlation plot for demand, it looks like it is an AR process. But from the spectral density of demand (see figure 7), we infer that it is cosine, following an MA process. Since it is a combination / mixture of both, we can say that Demand is probably an ARMA process. Upon looking at the auto-correlation plot for temperature, it is evident that it is an MA process and the same can be proved

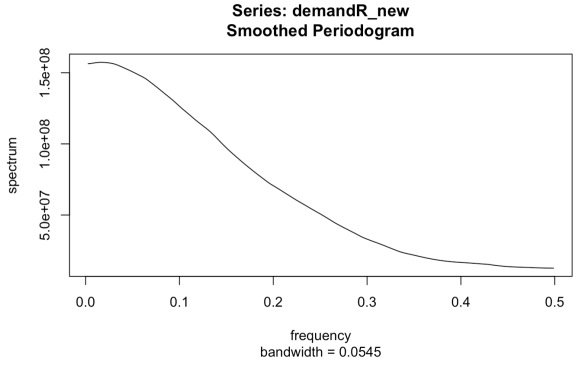

Fig. 7. Spectral Density of the residuals for Demand.

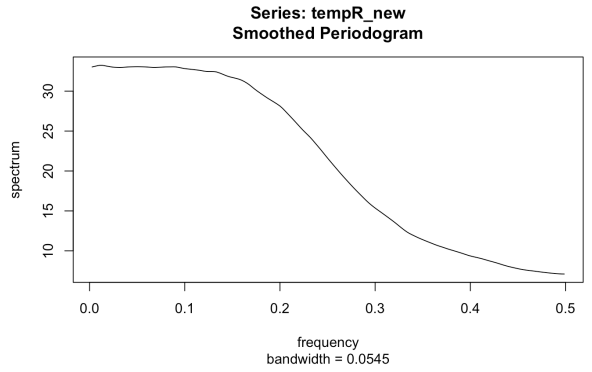

Fig. 8. Spectral Density of the residuals for Temperature.

by looking at the spectral density of temperature (see figure 8), which has a cosine shape. The poitive and negative correlation over different lags indicate that both the variables follow a cyclic systematic trend.

C. Cross-Correlation and Coherence

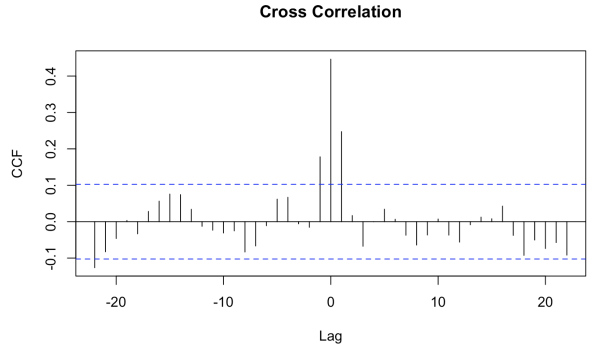

Fig. 9. Cross-correlation between residuals of demand and temperature.

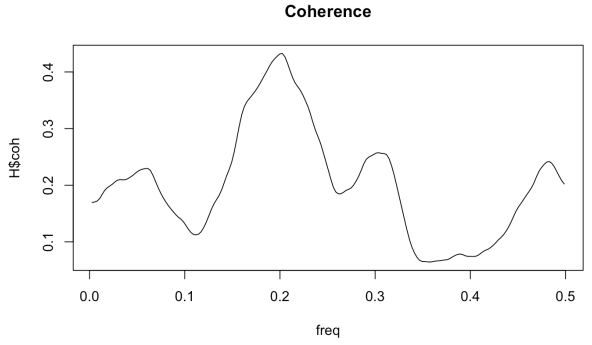

Fig. 10. Cross-spectral analysis between residuals of demand and temperature.

From the cross-correlation plot in figure 9, we can say that both series are leading each other at lag of 1. The plot is symmetric around lag 0 because the correlations at negative lags are the same as those at positive lags, but in reverse order. The plot shows that for most lags, the correlations are within the confidence bounds (indicated by the blue dashed lines), suggesting that there are no significant correlations at those lags. The peak of the cross-correlation plot is at a positive lag, suggesting that the demand is anticipating the temperature. Moreover, from the spectral density coherence plot (see figure 10), there are peaks, suggesting the presence of multiple periodic components in the data. The values in the coherence plot fluctuate more than in the squared coherence plot and reach higher peaks. This is the typical appearance of a coherence function. There appear to be peaks around 0.05, 0.2, 0.33 and just before 0.5 frequencies, which suggests that there is some degree of correlation between the two signals at these frequencies, although the correlation is not that high.

IV. Model Estimation

A. Linear Difference Equation with Noise

Here we have 3 competing difference equation models. The simple model is nested within the complex model and the third model is not nested. $y(t)$ is the criterion, i.e., 'Demand' and $x(t)$ is the predictor, i.e., 'Temperature'. $w(t)$ is white noise. f_1 and f_2 are frequencies where $y(t)$ shows systematic cycles in the periodogram (figure 3).

Simple Model -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot x(t) \\ &+ b_3 \cdot cos(2\pi f_1 t) + b_4 \cdot sin(2\pi f_1 t) \\ &+ b_5 \cdot cos(2\pi f_2 t) + b_6 \cdot sin(2\pi f_2 t) + w(t) \end{aligned}$$

Complex Model -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot y(t-2) + b_3 \cdot y(t-3) \\ &+ b_4 \cdot x(t) + b_5 \cdot x(t-1) + b_6 \cdot x(t-2) + b_7 \cdot x(t-3) \\ &+ b_8 \cdot \cos(2\pi f_1 t) + b_9 \cdot \sin(2\pi f_1 t) \\ &+ b_{10} \cdot \cos(2\pi f_2 t) + b_{11} \cdot \sin(2\pi f_2 t) + w(t) \end{aligned}$$

Third Model -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) \\ &+ b_2 \cdot x(t) + b_3 \cdot x(t-1) + b_4 \cdot x(t-2) \\ &+ b_5 \cdot cos(2\pi f_1 t) + b_6 \cdot cos(2\pi f_2 t) \\ &+ b_7 \cdot sin(2\pi f_2 t) + w(t) \end{aligned}$$

Since the simple model is nested within the complex model, we can do an F-test to check if the variance captured by both models is same.

```
anova(res1, res2)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ dY_1 + x + cos1 + sin1 + cos2 + sin2
## Model 2: Y ~ dY_1 + dY_2 + dY_3 + x + dx_1 + dx_2 + dx_3 + cos1 + sin1 +
##     cos2 + sin2
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    355 1.6496e+10
## 2    350 1.2927e+10  5 3569063398 19.327 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 11. Result of ANOVA between the simple and complex model.

By looking at the results of ANOVA (figure 11), we see that we get a p-value < 0.01 significance level, which means we

TABLE I
AIC & BIC METRIC COMPARISON FOR LINEAR DIFFERENCE EQUATION MODELS

Model Type	df	AIC	BIC
Simple	8	7427	7458
Complex	13	**7349**	**7399**
Third	9	7429	7464

```
##
## Call:
## lm(formula = Y ~ dY_1 + dY_2 + dY_3 + x + dx_1 + dx_2 + dx_3 +
##     cos1 + sin1 + cos2 + sin2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20207.8  -3496.1    -64.4   3320.4  18845.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.918e+01  3.195e+02  -0.154  0.87776
## dY_1         7.431e-01  5.162e-02  14.396  < 2e-16 ***
## dY_2        -8.292e-02  6.493e-02  -1.277  0.20243
## dY_3         1.051e-01  5.061e-02   2.076  0.03867 *
## x            5.281e+02  7.099e+01   7.439 7.89e-13 ***
## dx_1        -3.796e+02  8.571e+01  -4.429 1.27e-05 ***
## dx_2        -1.241e+02  8.761e+01  -1.416  0.15754
## dx_3        -2.096e+02  7.540e+01  -2.780  0.00573 **
## cos1         7.410e+03  7.056e+02  10.503  < 2e-16 ***
## sin1        -4.461e+03  5.982e+02  -7.458 6.97e-13 ***
## cos2        -5.688e+03  5.849e+02  -9.725  < 2e-16 ***
## sin2         2.280e+03  5.782e+02   3.944 9.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6077 on 350 degrees of freedom
## Multiple R-squared:  0.7901, Adjusted R-squared:  0.7835
## F-statistic: 119.7 on 11 and 350 DF,  p-value: < 2.2e-16
```

Fig. 12. Summary of the 'Complex' linear difference equation model.

can reject the Null Hypothesis (H_0) and accept the Alternate Hypothesis (H_1), i.e. both the models (simple vs complex) are not same. By looking at the AIC and BIC metric (Table I), we can see that the 'Complex Model' with an adjusted R^2 = 0.7835 (see figure 12) is the best as it has the lowest AIC and BIC scores.

B. ARMAX

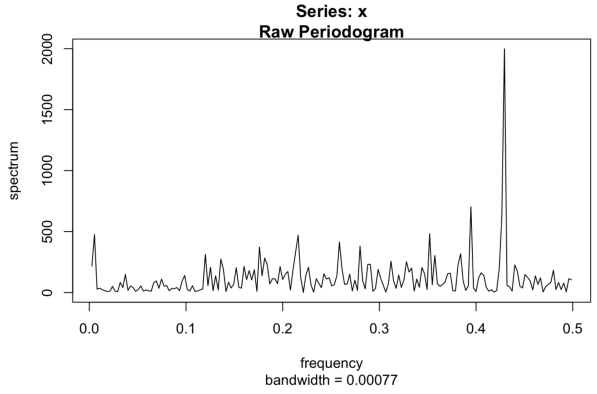

Fig. 13. Spectral analysis of the residuals of 'Complex Model'.

Here we have 5 competing ARMAX models of varying complexity in terms of autoregressive and moving average components. $y(t)$ is the criterion, i.e., Demand and $x(t)$ is the predictor, i.e., Temperature. $w(t)$ is white noise. f_1 and f_2 are frequencies where $y(t)$ shows systematic cycles in the periodogram (figure 3). When we look at the periodogram of the residuals from the previous best model i.e., Complex Model (figure 13), we saw there was a spike around f_3 = .429333333 , so we added sinusoids at f_3 to take care of this in our ARMAX models.

Model 1 -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot y(t-2) + b_3 \cdot y(t-3) \\ &+ b_4 \cdot x(t) + b_5 \cdot x(t-1) + b_6 \cdot x(t-2) \\ &+ b_7 \cdot cos(2\pi f_1 t) + b_8 \cdot sin(2\pi f_1 t) \\ &+ b_9 \cdot cos(2\pi f_2 t) + b_{10} \cdot sin(2\pi f_2 t) \\ &+ b_{11} \cdot w(t-1) + b_{12} \cdot w(t-2) + b_{13} \cdot w(t-3) + w(t) \end{aligned}$$

Model 2 -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot y(t-2) + b_3 \cdot y(t-3) \\ &+ b_4 \cdot x(t) + b_5 \cdot x(t-1) + b_6 \cdot x(t-2) \\ &+ b_7 \cdot cos(2\pi f_1 t) + b_8 \cdot sin(2\pi f_1 t) \\ &+ b_9 \cdot cos(2\pi f_2 t) + b_{10} \cdot sin(2\pi f_2 t) \\ &+ b_{11} \cdot w(t-1) + w(t) \end{aligned}$$

Model 3 -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) \\ &+ b_2 \cdot x(t) + b_3 \cdot x(t-1) + b_4 \cdot x(t-2) \\ &+ b_5 \cdot cos(2\pi f_1 t) + b_6 \cdot sin(2\pi f_1 t) \\ &+ b_7 \cdot cos(2\pi f_2 t) + b_8 \cdot sin(2\pi f_2 t) \\ &+ b_9 \cdot w(t-1) + b_{10} \cdot w(t-2) + b_{11} \cdot w(t-3) + w(t) \end{aligned}$$

Model 4 -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot y(t-2) + b_3 \cdot y(t-3) \\ &+ b_4 \cdot x(t) + b_5 \cdot x(t-1) \\ &+ b_6 \cdot cos(2\pi f_1 t) + b_7 \cdot sin(2\pi f_1 t) \\ &+ b_8 \cdot cos(2\pi f_2 t) + b_9 \cdot sin(2\pi f_2 t) \\ &+ b_{10} \cdot w(t-1) + w(t) \end{aligned}$$

Model 5 -

$$\begin{aligned} y(t) = b_0 &+ b_1 \cdot y(t-1) + b_2 \cdot y(t-2) \\ &+ b_3 \cdot x(t) + b_4 \cdot x(t-1) \\ &+ b_5 \cdot cos(2\pi f_1 t) + b_6 \cdot sin(2\pi f_1 t) \\ &+ b_7 \cdot cos(2\pi f_2 t) + b_8 \cdot sin(2\pi f_2 t) \\ &+ b_9 \cdot cos(2\pi f_3 t) + b_{10} \cdot sin(2\pi f_3 t) \\ &+ b_{11} \cdot w(t-1) + w(t) \end{aligned}$$

TABLE II
AIC & BIC METRIC COMPARISON FOR ARMAX MODELS

Model Type	df	AIC	BIC
Model 1	15	6068	6123
Model 2	13	6065	6113
Model 3	13	6067	6115
Model 4	12	**6064**	**6108**
Model 5	13	6089	6137

Since the models are not nested, we can use AIC and BIC to compare and choose the best model. From Table II, we can see that Model 4 has the lowest AIC and BIC scores, and therefore is our champion model.

V. Results

A. Residual Analysis

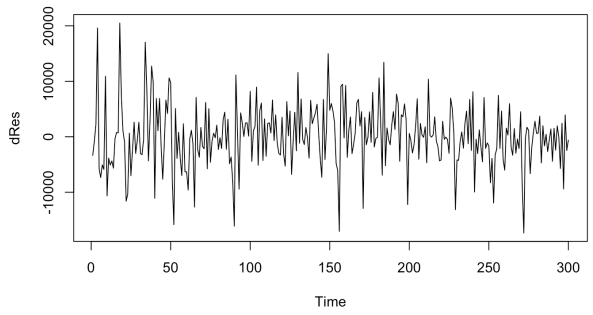

Fig. 14. Residuals after fitting champion model.

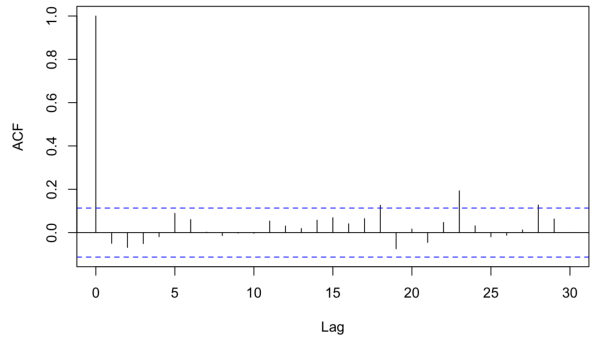

Fig. 15. Auto-correlation of residuals after fitting champion model.

After we have chosen our champion model, we want to make sure that the residuals are white noise. From figure 14, we observe the residuals are centered around zero and variance looks constant. The auto-correlation plot (figure 15) also does not show correlation at different lag components.

```
Box.test(dRes,lag = 10, type = 'Box-Pierce')
```

```
##
##  Box-Pierce test
##
## data:  dRes
## X-squared = 6.5552, df = 10, p-value = 0.7667
```

Fig. 16. Result of Box-Pierce test.

From Box-Pierce test (see figure 16), we can say that there is no significant evidence of auto-correlation up to lag 10 components. The periodogram (see figure 17) also looks flat with no significant peaks suggesting no auto-covariance at different frequencies. Thus, the residuals are white noise.

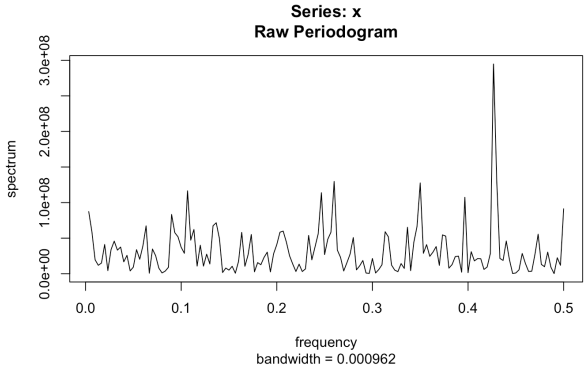

Fig. 17. Spectral analysis of the residuals of Champion Model.

B. Parameter Interpretation

The model includes 3 autoregressive terms and 1 moving average term. The AR(1) term seems to be significant and large relative to its standard error, suggesting a potentially strong influence on the model. The AR(2) term is significant but negative, while the MA(1) term is also significant and negative. The exogenous variable x included in the model is positively correlated with the dependent variable, implying that as x increases, so does the output. The sinusoids cos1 and sin1 are included to capture cyclic or seasonal trends. Since the coefficients are positive, it means there is a positive correlation between temperature and electricity demand.

```
##
## Call:
## arima(x = Y[1:300], order = c(3, 0, 1), xreg = X4[1:300, ])
##
## Coefficients:
##          ar1      ar2     ar3      ma1  intercept         x     dx_1       cos1
##       1.5970  -0.7609  0.1565  -0.8644  -1218.516  371.4880  53.4114  10587.821
## s.e.  0.0681   0.1006  0.0601   0.0409   5261.804   84.9184  82.8536    661.493
##            sin1        cos2        sin2
##       1425.4040  -4893.3461  -1209.8638
## s.e.   660.0537    366.6816    366.5829
##
## sigma^2 estimated as 32281251:  log likelihood = -3020.02,  aic = 6064.04
```

Fig. 18. Coeff. and Std Errors of estimates for the champion ARMAX model.

VI. Conclusion

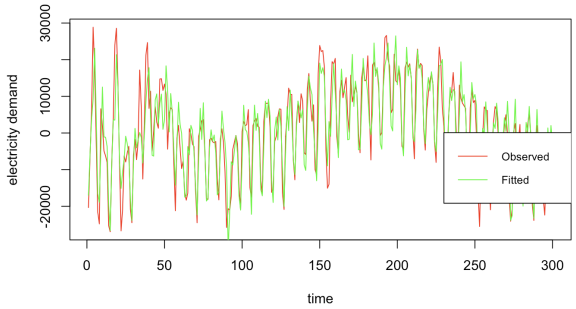

Fig. 19. Observed data vs Fitted values from the champion model.

From figure 19, we can see the model estimates the observed data very closely. The model captures the peaks very accurately but is over estimating the troughs.

Since our goal was to see how well the model can be used for forecasting, we used the first 300 datapoints for estimating the model parameters and latter 65 datapoints, i.e., 8 weeks / 2 months for forecasting.

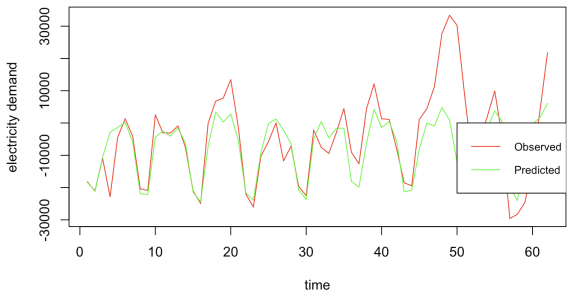

Fig. 20. Observed data vs Predicted values from the champion model on the hold out data.

From figure 20, we see the model does a pretty good job in predicting the future values of demand for the first few cycles. The performance degrades for future time points which is expected since ARIMA works for short / mid-term and not long-term predictions.
In conclusion, we can infer that daily temperatures works well in estimating the electricity demand over time. Since our data was for one year, we were restricted by number of predictors which could be taken into consideration. If data is available for more time points, it would be interesting to see how the other factors like humidity and holidays affect the demand.

VII. Source Code

Source code for all experiments and results can be found at https://github.com/DhairyaC/Electricity-Demand-Forecasting.

Acknowledgment

The author wishes to thank Prof. Jerome R. Busemeyer for his guidance and support. This work was part of the Time Series Analysis (STAT-S 650) course for Spring 2024 at Indiana University, Bloomington.

References

[1] V. Puspita and Ermatita, “Time series forecasting for electricity consumption“, Journal of Physics, vol. 1196, no. 1.
[2] P. Pelka, “Analysis and forecasting of monthly electricity demand time series“, vol. 1196.
[3] “Daily Electricity Price and Demand Data“, Kaggle dataset.