Mid Term Electricity Demand Forecasting

A Saikia
M.S. Data Science
Indiana University - Bloomington
arsaikia@iu.edu

I. INTRODUCTION

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

The data is for operational electricity demand for Victoria, Australia. 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 2014.
- Temperature : maximum daily temperatures for Melbourne

The data has 365 timepoints for each day in the year 2014 'Demand' is the criterion of interest and 'Temperature' is the predictor variable.

III. DATA ANALYSIS

A. Time Domain and Frequency Domain

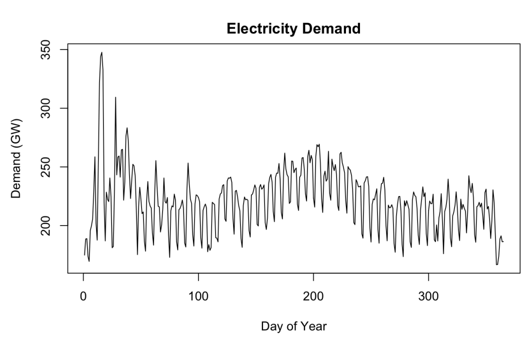

Fig. 1: Daily electricity demand for the year 2014 in Victoria, Australia

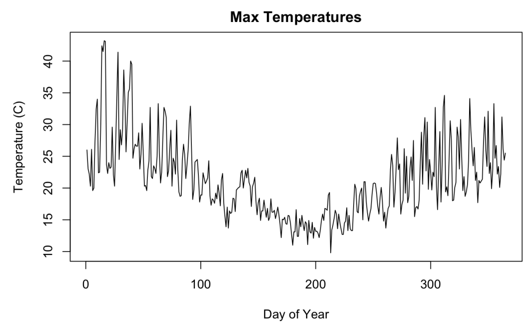

Fig. 2: Maximum daily temperatures for the year 2014 recorded by BOM site at Melbourne, Victoria

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.

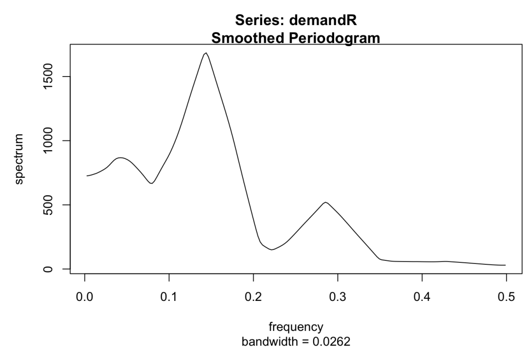

Fig. 3: Periodogram for residuals of Demand after fitting a polynomial of order 6 to de-trend the series

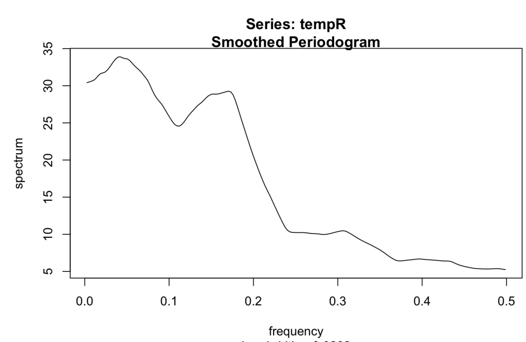

Fig. 4: Periodogram for residuals of Temperature after fitting a polynomial of order 4 to de-trend the series

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 at $f_1' = .1706667$ denoting some periodicity every 6 days. So we will try to remove these using regression by fitting sinusoids at these frequencies.

B. Autocorrelation and Spectral Density

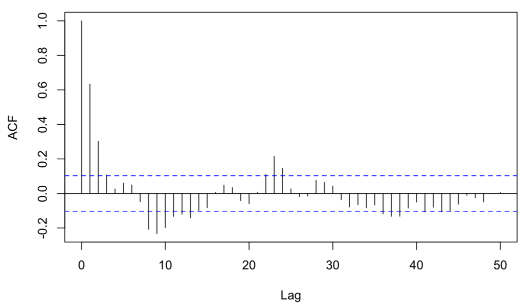

Fig. 5: Autocorrelation of the residuals in Demand after making the series stationary

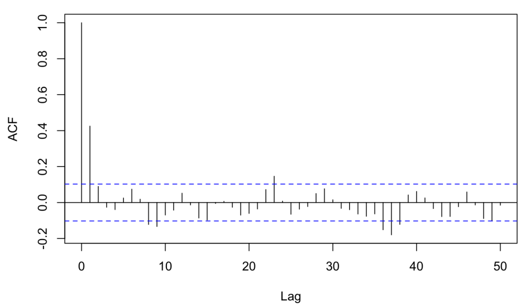

Fig. 6: Autocorrelation of the residuals in Temperature after making the series stationary

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. From figure 5 it looks like Demand is an AR(2) process while Temperature (see figure 6) is MA(1) process.

The spectral density for Temperature (see figure 8) has a cosine shape which means it is MA processes. Although spectral density for Demand also has a cosine shape (see figure 7), we saw above that it is an AR process. Therefore, Demand is probably an ARMA process.

C. Cross-Correlation and Coherence

From the cross-correlation plot in figure 9 we can say that both series are leading each other at lag of 1. Also from the squared coherence (see figure 10) there seems to be some correlation between the two series at $f = .17$ and $f = .32$ although the correlation is not that high.

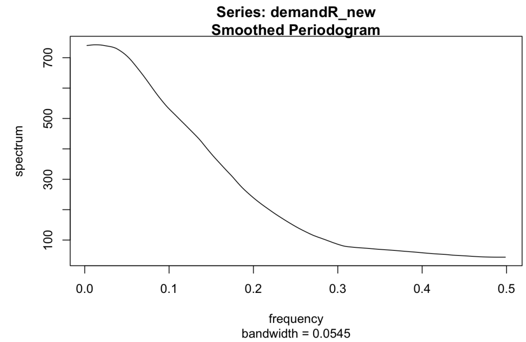

Fig. 7: Spectral Density of the residuals for Demand

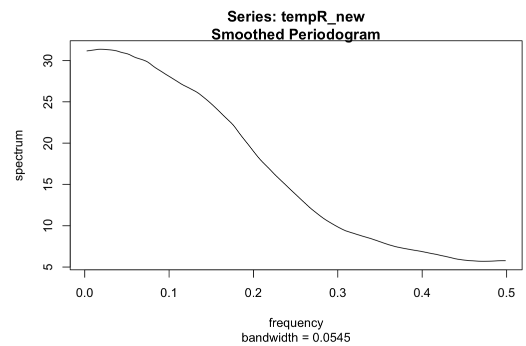

Fig. 8: Spectral Density of the residuals for Temperature

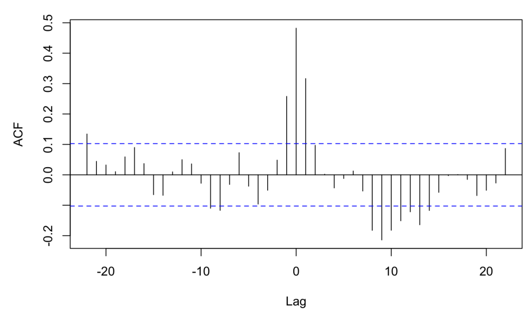

Fig. 9: Cross-correlogram between residuals of demand and temperature

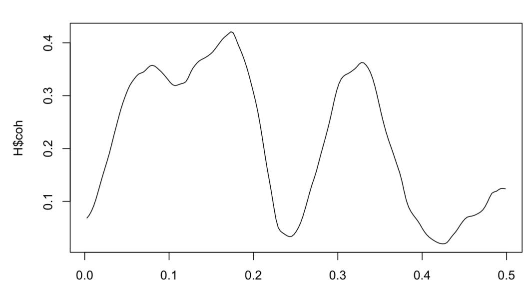

Fig. 10: Cross-spectral analysis between residuals of demand and temperature

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 and $x(t)$ is the predictor. $w(t)$ is white noise. f_1, f_2 are frequencies where $y(t)$ shows systematic cycles in the periodogram (figure 3).

Simple Model -
$y(t) = b_0 + b_1 \cdot y(t-1) + b_2 \cdot x(t) + b_3 \cdot cos(2\pi f_1 t) + b_4 \cdot sin(2\pi f_1 t) + b_5 \cdot cos(2\pi f_2 t) + b_6 \cdot sin(2\pi f_2 t) + w(t)$

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

Third Model -
$y(t) = b_0 + b_1 \cdot y(t-1) + b_2 \cdot x(t) + b_3 \cdot x(t-1) + b_4 \cdot x(t-2) + b_5 \cdot cos(2\pi f_1 t) + b_6 \cdot cos(2\pi f_2 t) + b_7 \cdot sin(2\pi f_2 t) + w(t)$

Since the simple model is nested within the complex model we can do a 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 61062
## 2    350 49596  5     11466 16.183 2.321e-14 ***
## ---
## 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 get a p-value < 0.01 significance level, which means we can reject the Null hypothesis and both models (simple vs. complex) are not same. By looking at the AIC and BIC metric (Table I) we can see that the 'Third Model' with an adjusted $R^2 = 0.798$ (see figure 12) is the best as it has lowest AIC and BIC scores.

TABLE I: AIC and BIC values for linear difference equation models

Models	df	AIC	BIC
Simple	8	2899	2930
Complex	13	2834	2884
Third	9	**2831**	**2866**

```
summary(res3) # Third model
```

```
##
## Call:
## lm(formula = Y ~ dY_1 + x + dx_1 + dx_2 + cos1 + cos2 + sin2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.976  -6.454  -0.329   6.158  55.017
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.03336    0.62668   0.053 0.957577
## dY_1          0.84260    0.02541  33.160  < 2e-16 ***
## x             1.37980    0.15441   8.936  < 2e-16 ***
## dx_1         -0.78314    0.18750  -4.177 3.73e-05 ***
## dx_2         -0.62132    0.15313  -4.058 6.10e-05 ***
## cos1         14.97108    0.89842  16.664  < 2e-16 ***
## cos2         -3.31586    0.91484  -3.625 0.000332 ***
## sin2        -15.02833    0.89932 -16.711  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.92 on 354 degrees of freedom
## Multiple R-squared:  0.8019, Adjusted R-squared:  0.798
## F-statistic: 204.7 on 7 and 354 DF,  p-value: < 2.2e-16
```

Fig. 12: Summary of the linear difference equation model 3

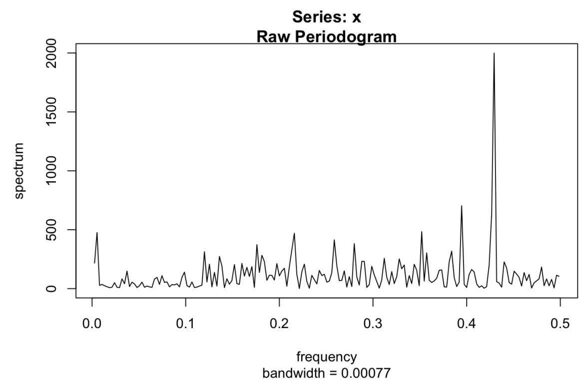

Fig. 13: Spectral analysis of the residuals from 'Third Model'

B. ARMAX

Here we have 5 competing ARMAX models of varying complexity in terms of autoregressive and moving average components. $y(t)$ is the criterion and $x(t)$ is the predictor. $w(t)$ is white noise. f_1, f_2 are frequencies where $y(t)$ shows systematic cycles in the periodogram (figure 3). When we looked at the periodogram of the residuals from the previous best model i.e. Third 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 -
$y(t) = b_0 + b_1 \cdot y(t-1) + b_2 \cdot y(t-2) + b_3 \cdot y(t-3) + b_4 \cdot x(t) + b_5 \cdot x(t-1) + b_6 \cdot x(t-2) + b_7 \cdot cos(2\pi f_1 t) + b_8 \cdot sin(2\pi f_1 t) + b_9 \cdot cos(2\pi f_2 t) + b_{10} \cdot sin(2\pi f_2 t) + b_{11} \cdot w(t-1) + b_{12} \cdot w(t-2) + b_{13} \cdot w(t-3) + w(t)$

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

Model 3 -
$y(t) = b_0 + b_1 \cdot y(t-1) + b_2 \cdot x(t) + b_3 \cdot x(t-1) + b_4 \cdot x(t-$

$2) + b_5 \cdot cos(2\pi f_1 t) + b_6 \cdot sin(2\pi f_1 t) + b_7 \cdot cos(2\pi f_2 t) + b_8 \cdot sin(2\pi f_2 t) + b_9 \cdot w(t-1) + b_{10} \cdot w(t-2) + b_{11} \cdot w(t-3) + w(t)$

Model 4 -

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

Model 5 -

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

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 **Model 5** has lowest AIC and BIC scores, and therefore is our **champion model**.

TABLE II: AIC and BIC values for ARMAX models

Models	df	AIC	BIC
Model 1	15	2344	2400
Model 2	13	2340	2388
Model 3	13	2343	2392
Model 4	12	2339	2383
Model 5	13	**2331**	**2380**

V. Results

A. Residual Analysis

After we have chosen our champion model we want to make sure if 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.

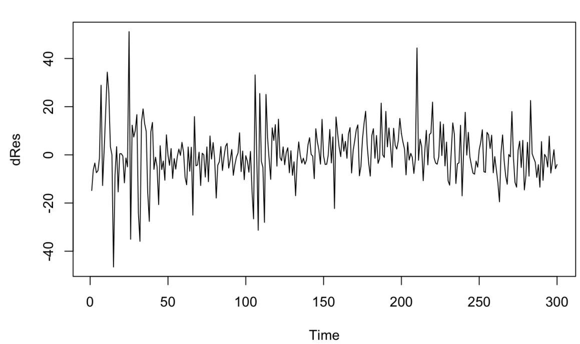

Fig. 14: Residuals after fitting champion model

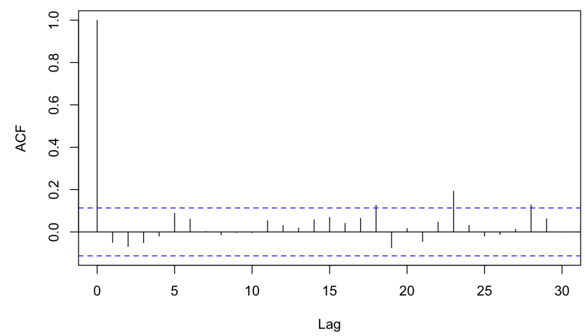

Fig. 15: Autocorrelation of residuals after fitting champion model

```
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

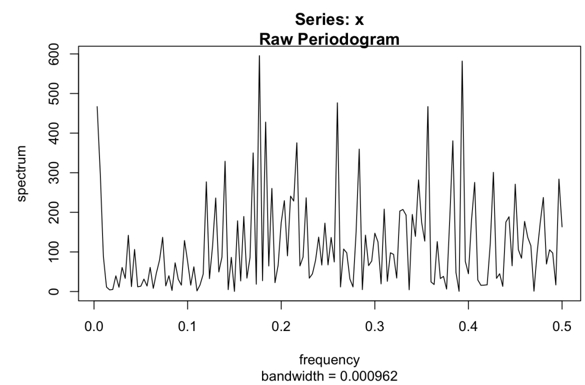

Fig. 17: Spectral analysis of the residuals from champion model

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 autocovariance at different frequencies. Thus, the residuals are white noise.

B. Parameter Interpretation

The model has order 2 feedback terms and order 1 noise component. The AR(2) term and MA(1) term are significant and positive. AR(1) is insignificant as 95% confidence interval of the coefficient estimates will include 0. The model also has exogenous inputs which drive the mean of the output. There are sinusoids which capture cyclic/seasonal trends. Almost all sinusoids are statistically significant except cos3. $x(t)$ and $x(t-1)$ capture the effect of systematic input i.e temperature on the output. Since the coefficients

```
##
## Call:
## arima(x = Y[1:300], order = c(2, 0, 1), xreg = X5[1:300, ])
##
## Coefficients:
##          ar1     ar2     ma1  intercept       x    dx_1     cos1     sin1
##       0.0514  0.7120  0.8592     3.4703  1.6943  0.5229  10.0945  17.0108
## s.e.  0.1501  0.1364  0.1317     5.0087  0.1762  0.1758   1.1534   1.1504
##         cos2      sin2     cos3     sin3
##       4.7097  -10.1311  -0.9484  -2.1452
## s.e.  0.6415    0.6417   0.5008   0.4978
##
## sigma^2 estimated as 126.9:  log likelihood = -1152.94,  aic = 2331.87
```

Fig. 18: Coeff. and std errors of estimates for the champion ARMAX(2, 0, 1) model

are positive it means there is a positive correlation between temperature and electricity demand.

VI. Conclusion

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 $\sim$ 8 weeks/2 months for forecasting. 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. 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.

In conclusion we can see that daily temperatures works well is estimating the electricity demand over time. Since our data was for 2014 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 here `https://github.com/arunavsk/Time-Series-S650`.

VIII. Acknowledgements

The author wishes to thank Prof Jerome Busemeyer for his instructions and support. This work was part of Time Series Analysis (STAT-S650) course for Spring 2020 at Indiana University, Bloomington.

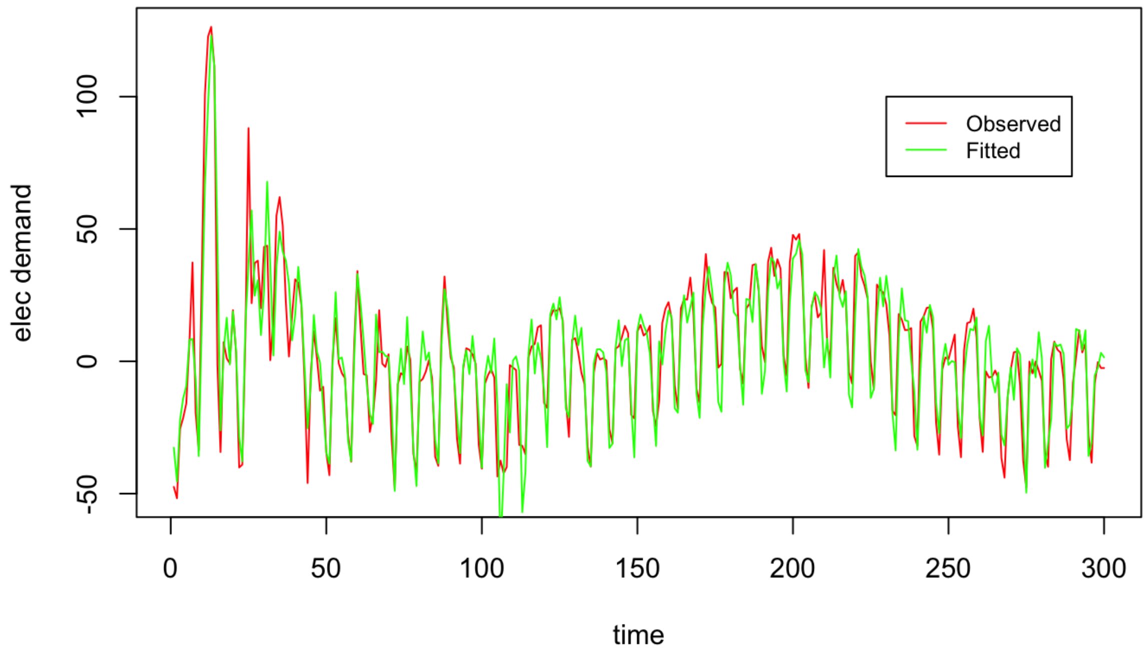

Fig. 19: Observed data and fitted values from the champion model

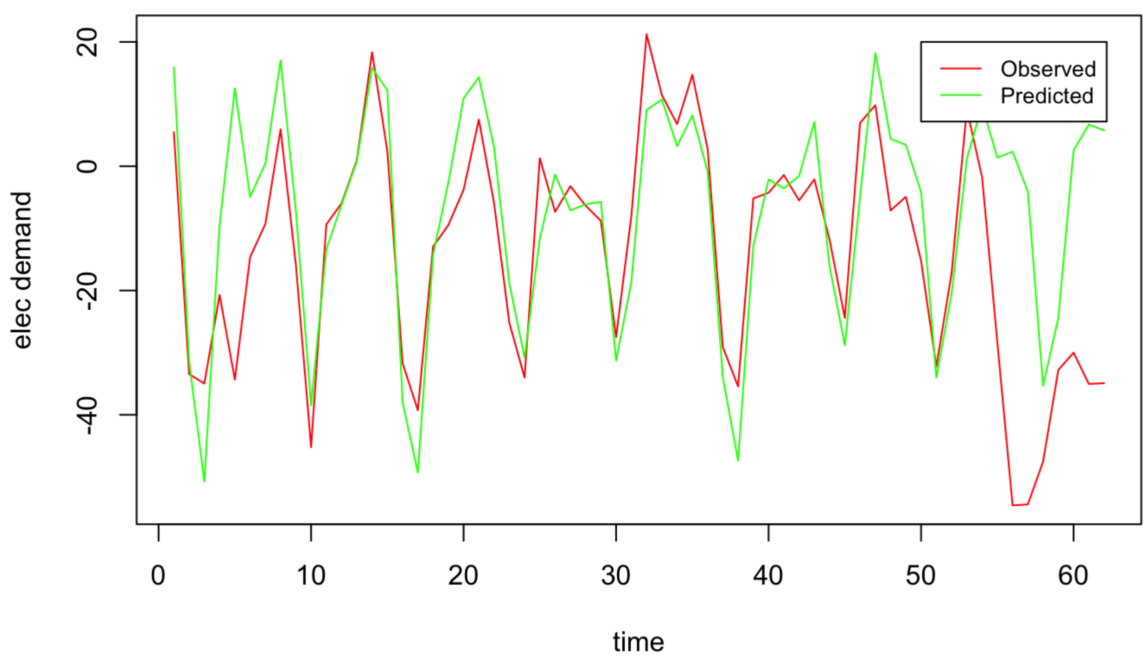

Fig. 20: Observed data and predicted values from the champion model on the hold out data