Part Two

To begin, we created a test data set since in part one we only had training and validation data. We also chose no change naive method as the benchmark, since it has the lowest MAPE out of four measurements we used in part one. After obtaining and evaluating the naive method in part one, our team will be trying more complex methods to see if we can get significantly more accurate forecasts. The objective of this second part is to use the training dataset to train smoothing and regression forecasting models and then generate one day forecasts to evaluate the performance of each model.

Our zone is the ComEd region, located in northern Illinois and includes the city of Chicago. It covers approximately 70 percent of the population in Illinois and the majority of the state's industrial companies. Its population, it approximately 3.4 million households.[1] As mentioned in part one, Illinois is a very industrialized area with high population density.

We found that many manufacturing plants and chemical processing plants run 24/7 when operating at capacity. It is commonly referred to as the 'third shift' in manufacturing and 'continuous production' in chemical processing (Kettle, 2001). While, we cannot tell with precision if that is the case for the majority of ComEd's industrial sector, we can expect to find a smaller difference between the electricity consumption on weekends, holidays, and weekdays. In addition, we expect temperature to play a less significant role in predicting electricity demand because manufacturing does not typically increase or decrease electricity consumption significantly due to changes in outside temperature.

To evaluate how much of our total available data to use in training, validation and test dataset, we considered how different the various years were with respect to daily electricity

[1] Regional demographics. ComEd

demand. To get the best out-of-sample performance, our training set should capture the various fluctuations in electricity demand due to randomness while ensuring there is no large systematic difference between each of the data sets.

Since our goal is a good forecasting model, we evaluated model performance on the validation data with multiple error measurements such as MAPE and Percentage-Bias. However, it is important to note that these measurements do not take into account the efforts needed to achieve such levels of performance. If our new methods are truly better at forecasting, they should significantly outperform our naïve model.

Evaluation of smoothing methods

Exponential smoothing methods are a simple and natural generalization of naive forecasts. For the purpose of developing a well performing smoothing models, we used State Space model and TBATS. We decided not to use other smoothing methods placed inside State Space models such as Holt-Winters, 2-Parameter Holt, and Simple Exponential Smoothing (SES) because as we mentioned in part one, there are two seasonal cycles for electricity demand in our region: annual seasonality and weekly seasonality. The Double Seasonal Holt-Winters method was not used because it cannot account for the structure on the errors, as will be shown in the regression section. We used TBATS model over BATS because not only can it can handle several cycles of seasonality, but it also allows for non-integer cycles of seasonality and performs faster than BATS.

For exponential smoothing methods we partitioned the data as follows: (1) Training dataset (year) = 2008-2013, (2) Validation dataset (year) = 2014-2016, and (3) Test dataset (year) = 2017-2018. This period was divided like that because even after we deactivate the Box-Cox transformation, only when we choose 2008-2013 as the training dataset can we generate a TBATS model that does not have a {0,0} ARMA errors.

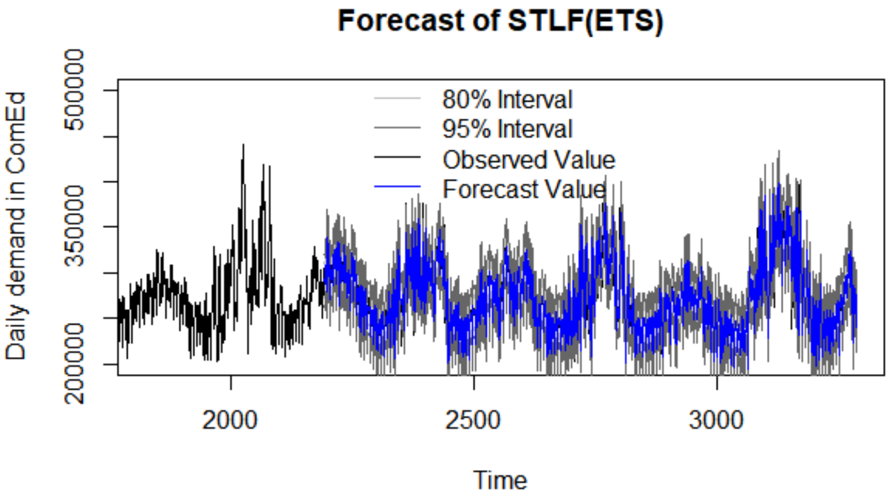


Figure 1 Forecast of STLF(ETS)

We used daily data for six years to forecast the next day's electricity demand in the rolling window in the forecast of STLF(ETS) and we got an MAPE of 4.76, which is better than the result of the naïve method.

For the TBATs we used daily data for six years to forecast the next day's electricity using a rolling window to keep the length of the training data set fixed. The MAPE is 13.41, which is worse than the naïve method. It might be the bizarre pattern of the recent data in 2012 due to the high temperature as we explained in linear regression part.

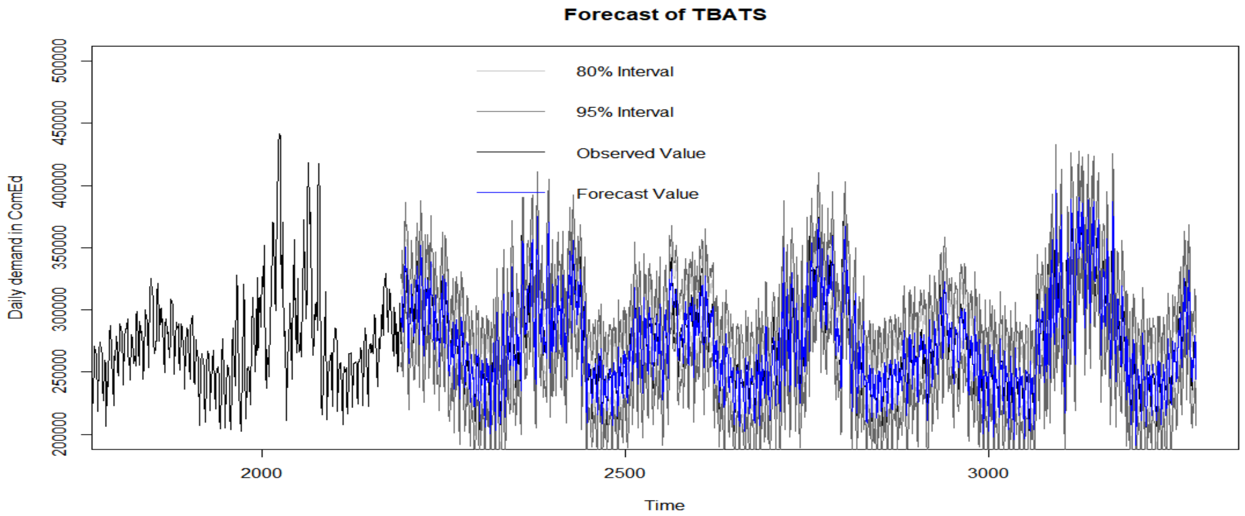


Figure 2 Forecast of TBATS using rolling window

Regression Methods

Explanatory variables

There are multiple explanatory variables that can explain electricity demand in our region. CDD and HDD are commonly used explanatory variables because they are directly related to increased cooling needs in the summer and increased heating needs in the winter.

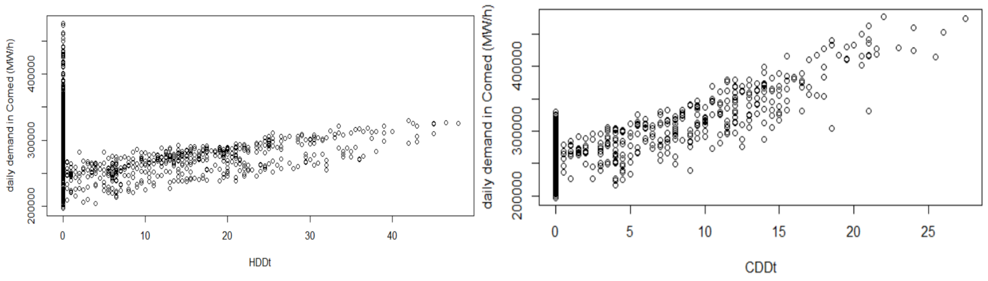


Figure 3 HDD (Tref 50F) and CDD (Tref 65F) effects

Their linear relationship with electricity demand can be seen in Figure 3. Given that the relationship between demand and HDD is not completely linear if we use a standard Tref of 65°F, we adjusted it to 50°F to omit the non linear effect.

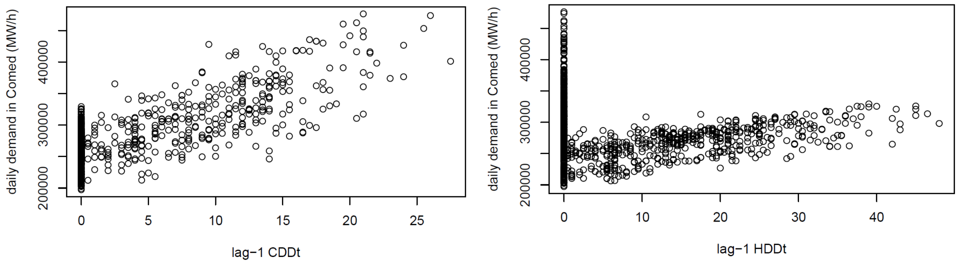


Figure 4 CDD and HDD effects

Lag-CDD’s and Lag-HDD’s shown in figure 4 can also be important explanatory variable because consecutive days of increased temperatures or decreased temperatures can cause heat or coldness to accumulate without wearing off, leading to higher than normal use of cooling or heating utilities.

As seen in Figure 5, days of the week and holidays, including both moving and fixed holidays, are another important variable that can help explain the lower electricity demand on certain days even when temperature may be equal. It is lower because less people go to work. Furthermore, dividing the holidays into moving and fixed will allow us to capture the effects of various types of holidays in more detail. For example, a fixed holiday like Christmas and New Year's is celebrated more widely versus a moving holiday like Easter or Columbus day, which may have less participants. Consequently, we added a fixed-holiday lag-1 since most fixed holidays are considered very important and can trigger a different dynamic even one day before. For example, December 24th and December 31th may not be holidays but many people may not work their regular hours on these days.

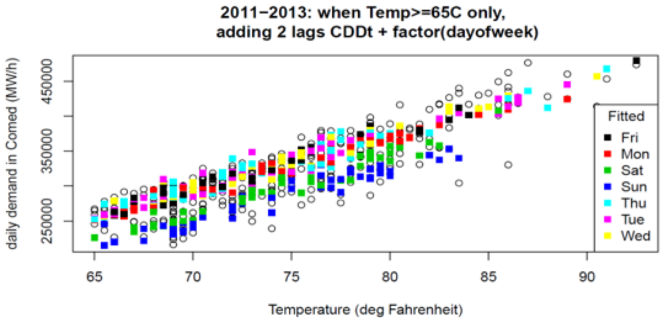


Figure 5 Linear model including day of the day effect

Humidity is another relevant climate variable that can explain the demand of energy. This is especially true during the summer when high humidity can amplify the needs for cooling. Although the relationship doesn't appear completely linear, allowing this variable into the regression model gives a lower adjusted R-squared value when compared with the adjusted R-squared value without it. The wind speed (including wind chill) and precipitation variables were discarded as they do not show any clear direct relation that could potentially be useful to explain our dependent variable.

Partition of the set

For our regression methods we partitioned the dataset as follows: (1) Training data = 2005-2013, (2) Validation dataset = 2014-2016, and (3) Test dataset = 2017-2018. As can be seen in Figures 6 and 7, there is similar behavior between temperature and electricity demand for 2005-2013 and 2014-2016. Thus, by training our regression models on data from 2005-2013 and validating on 2014-2016, we can be assured that our out-of-sample performance is not compromised due to significantly different behavior for in-sample and out-of-sample data.

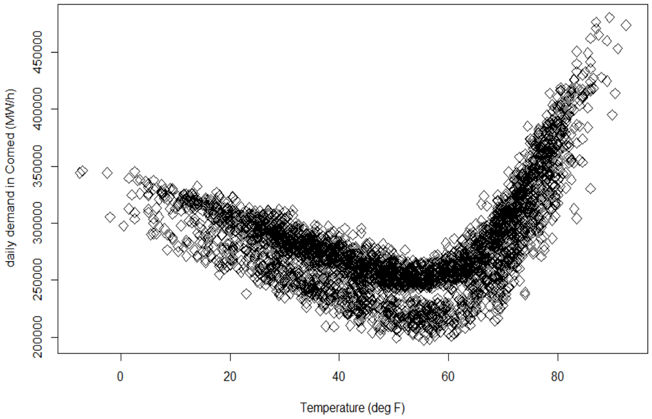


Figure 6 Temperature vs. demand training set (2005 to 2013)

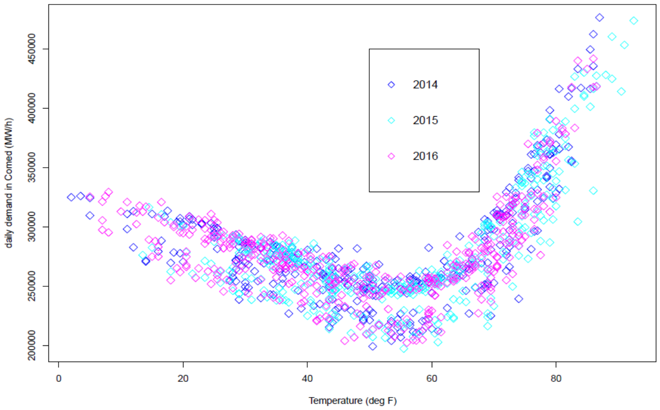


Figure 7 Temperature vs. demand validation set (2014 to 2016)

Development of Regression Model

We started by fitting a standard linear multiple regression model using the 10 explanatory variables stated above: $Y_t = \widehat{\beta_0} + factor(day\ of\ the\ week) + factor(moving\ holidays)$

$$+ factor(fixed\ holidays) + \hat{\beta}_{cdd} * CDD_t + \hat{\beta}_{hdd} * HDD_t + \hat{\beta}_{humidity} * Humidity_t$$

$$+\hat{\beta}_{lag1\ cdd} * CDD_{t-1} + \hat{\beta}_{lag2\ cdd} * CDD_{t-2} + \hat{\beta}_{lag1\ hdd} * HDD_{t-1} + \hat{\beta}_{lag2\ hdd} * HDD_{t-2} + \varepsilon_t$$

Where $\varepsilon_t \approx N(0, \sigma^2)$

Evaluation of Regression Methods

As stated in Chapter 6, the linear regression model is based on five assumptions: (1) E(ε)= 0, (2) Var(ε)is constant (3) $\varepsilon_i$are not correlated (4) $\varepsilon_i$follow a normal distribution and (5) Explanatory variables are not collinear. However, as one can see in Figure 8, the variance of the residual is constant, but does not follow a normal distribution according to the Q-Q plot. The three outlier points tagged on this graph as outliers are all the effect of abnormally high temperatures during different times of the years.

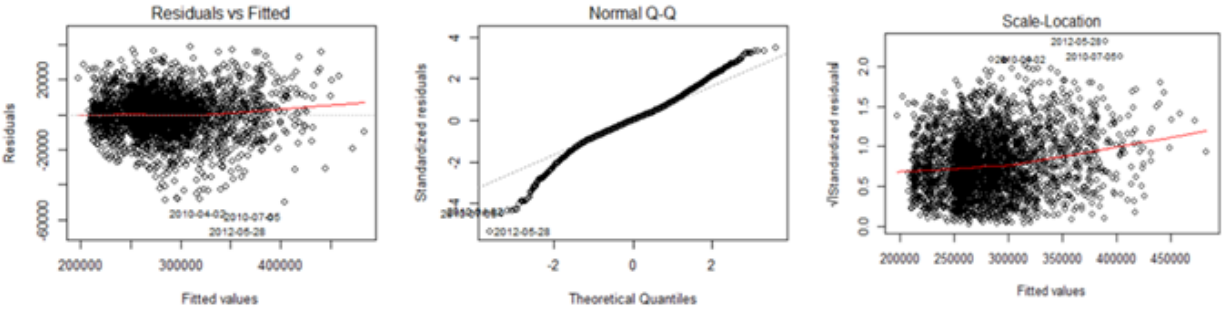
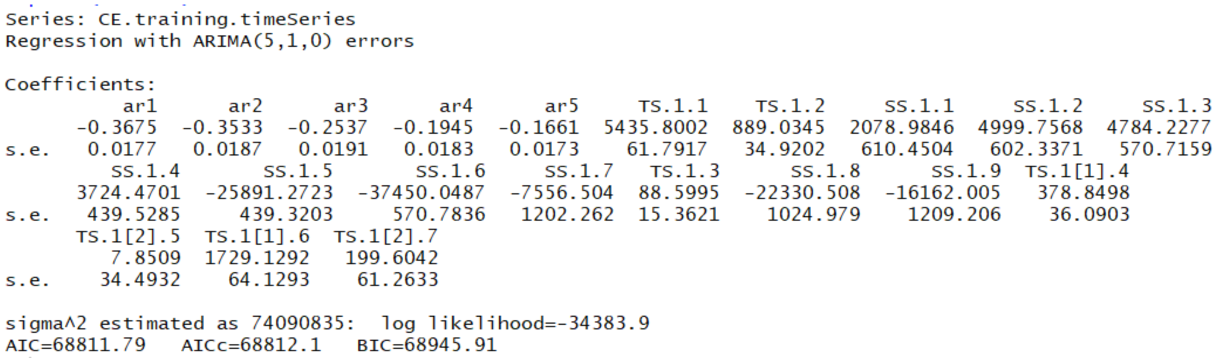


Figure 8 Residual plots of the linear model

In addition, as typical with most time series data, the Durbin-Watson test (p−value = $2.2e - 16$) shows that the residuals seem to have a certain structure or dependency. Thus, the null hypothesis is being rejected, and there is an autocorrelation among residuals. The R-squared for our linear regression model was 0.9336.

Linear regression with ARIMA errors

In order to allow the model to incorporate the serial dependence structure of the errors and obtain residuals that are independent of each other. We used the auto-arima function in R that gives the best possible error structure by minimizing the AICC, which gave us ARIMA (5,1,0) as shown in Figure 9:

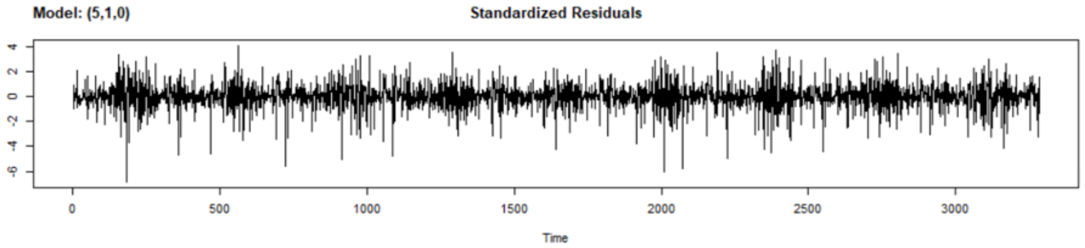


```
Series: CE.training.timeSeries
Regression with ARIMA(5,1,0) errors

Coefficients:
          ar1      ar2      ar3      ar4      ar5     TS.1.1    TS.1.2     SS.1.1     SS.1.2     SS.1.3
      -0.3675  -0.3533  -0.2537  -0.1945  -0.1661  5435.8002  889.0345  2078.9846  4999.7568  4784.2277
s.e.   0.0177   0.0187   0.0191   0.0183   0.0173    61.7917   34.9202   610.4504   602.3371   570.7159
         SS.1.4       SS.1.5       SS.1.6     SS.1.7   TS.1.3      SS.1.8      SS.1.9  TS.1[1].4
      3724.4701  -25891.2723  -37450.0487  -7556.504  88.5995  -22330.508  -16162.005   378.8498
s.e.   439.5285     439.3203     570.7836   1202.262  15.3621    1024.979    1209.206    36.0903
      TS.1[2].5  TS.1[1].6  TS.1[2].7
         7.8509  1729.1292   199.6042
s.e.    34.4932    64.1293    61.2633

sigma^2 estimated as 74090835:  log likelihood=-34383.9
AIC=68811.79   AICc=68812.1   BIC=68945.91
```

***Figure** 9 Output of the ARIMA model*

Checking assumptions

To validate again for the five assumptions stated above we used the Sarima function to make the fit of the model including ARIMA (5,1,0) errors. After incorporating this structure for the residuals, we ran the regression again, and although the ACF of the residuals was zero, the Q-Q plot still indicated that the residuals are not normally distributed. Even more, on the Ljung Box we can see that errors appear to be correlated (figure 10).



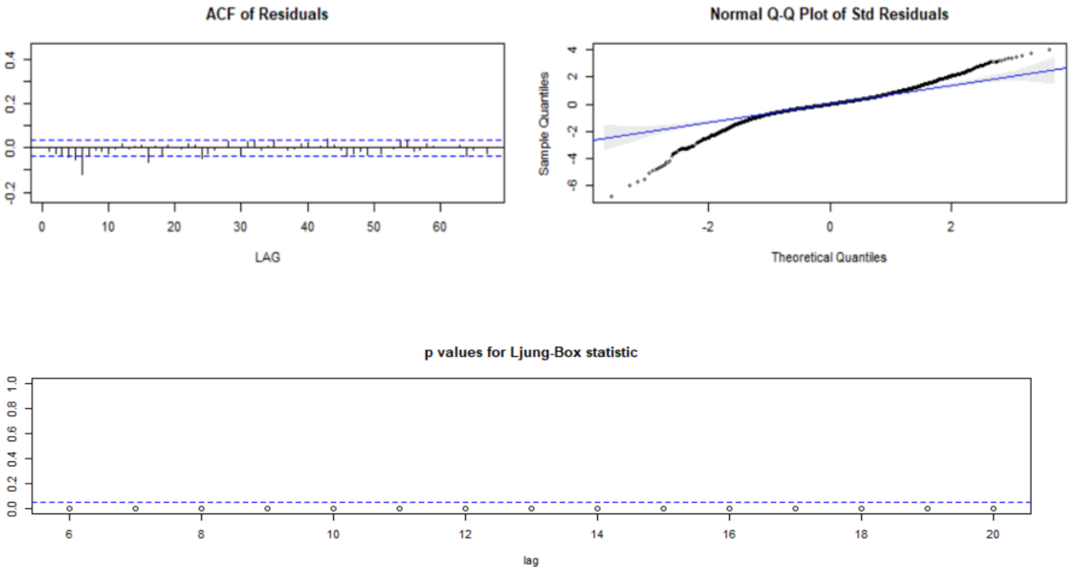


Figure 10 Diagnostic plot of the residuals for the ARIMA model

We tried to fix this problem by taking further differences and also by including MA(q) terms but none of this generated a better behaviour for the errors.

Forecasting results

Using the model ARIMA(5,1,0), we used the predict function to forecast for the entire validation set (2014-2016). When doing the forecasting we did not have to forecast any of the explanatory variables, since the forecast for Yt only involves variables at t-1 or before. We computed the MAPE for this period as 5.5170. Thus, the MAPE is lower than with the naive method (MAPE = 6.32)

Conclusion

After running different methods for the training data set, we were able to compare the results from two different types of forecasts. We got the result of the linear regression which relies on the capacity of the explanatory variables to explain the dependent variable and the result of TBATS from the exponential smoothing family which uses only historical data to model our series as a combination of different seasonal and components. After running the regression and incorporating an ARMA structure for the errors, we found out that the assumptions of the regressions were not satisfied completely. Nevertheless, we were able to make forecasts for three full years obtaining an MAPE which beated the naive method in terms of the MAPE (5.51 vs. 6.32) . On the other hand, the exponential smoothing methods has mixed results, the ETS showed a very accurate forecast (4.76 MAPE) while the TBATS, didn't performed as expected (13.41 MAPE); This could be explained by the choice of the length of the training data set, which might have be too large for a 1 day ahead forecast.

Bibliography

Clean Technica. (n.d.). *PJM Interconnection territory*. Retrieved from https://cleantechnica.com/2012/03/19/grid-operators-report-details-energy-market-shift-to-clean-energy/pjm-interconnection-territory/

Comed. (n.d.). *Regional Demographics*. Retrieved from Northerm Illinois location: https://www.comed.com/DoingBusinessWithUs/Pages/RegionalDemographics.aspx

Kettle, M. (2001, Aug 3) *So long, American work culture* : https://www.theguardian.com

National Weather Service Forecast Office. (2019). Retrieved from Chicago, IL: https://w2.weather.gov/climate/xmacis.php?wfo=lot

NewsMax. (2019). *Top 5 Industries in Illinois*. Retrieved from https://www.newsmax.com/fastfeatures/industries-illinois-chicago-economy/2015/04/05/id/636484/

Schellong, W. (2011). Energy Demand Analysis and Forecast. *Energy Management Systems. University of Cologne, Germany.* Retrieved from https://www.intechopen.com/books/energy-management-systems/energy-demand-analysis-and-forecast

U.S Energy Information Administration. (2019). *Illinois, state profile and energy estimates*. Retrieved from https://www.eia.gov/state/?sid=IL#tabs-2

Weather Underground. (2019). *Chicago O'hare International Station*. Retrieved from Historical: https://www.wunderground.com/history/monthly/us/il/chicago/KORD/date/2019-2

Wikipedia. (2019). *Economy of Illinois*. Retrieved from https://en.wikipedia.org/wiki/Economy_of_Illinois