Predicting Energy Usage in Colorado Using ARMA Processes

Elizabeth Rodriguez
University of Colorado Boulder
April 30, 2022

Introduction

This project was started by loading the Public Service Company of Colorado dataset from EIA into R. This dataset included the date, time, and energy consumption (in hertz). This data set spans from 2015 to 2022 and updates every hour for each day. This data set is an excellent tool for practicing how to predict numerical values while using time series processes to identify patterns or noise, such as energy consumption. Predicting energy consumption is also essential in many other studies as well, such as energy storage and energy generation. Energy is used every hour of the day to light our houses up, charge our phones, and heat up our food. It is part of our daily routine, and it is interesting to see how one state uses its time and energy. The purpose of this paper is to use different autoregressive moving average (ARMA) models to try to predict energy consumption in Colorado between 5 pm to 6 pm on May 1st, 2022, and therefore conclude whether or not ARMA processes are a reasonable and viable modeling effort for energy consumption.

Methods and Results

The following methodology was used to try to predict the energy consumption in Colorado between 5 pm to 6 pm on May 1st, 2022.
ARMA models are usually used when dealing with time-series data. An ARMA model, or autoregressive moving average model, comes from combining two models; an autoregressive model (AR model) and a moving average model (MA model). The autoregressive model takes into account past information in the data, also known as lagged values. In our case, it could be past temperatures, or energy consumption to help build a model to predict the future energy demand better. The moving average part of this model is modeling the error terms of the current model, along with the past errors. An ARMA(p,q) model has two input variables, p is the order of the AR model, and q is the order of the MA model.

The first method tried was to look specifically at past years' data on energy usage in Colorado on May 1st, between 5 and 6 pm, including the past data from 2016 to 2021 (6 years). From this data, I found a simple linear regression model including a negative intercept term with a coefficient of 72.74. The resulting equation is given by

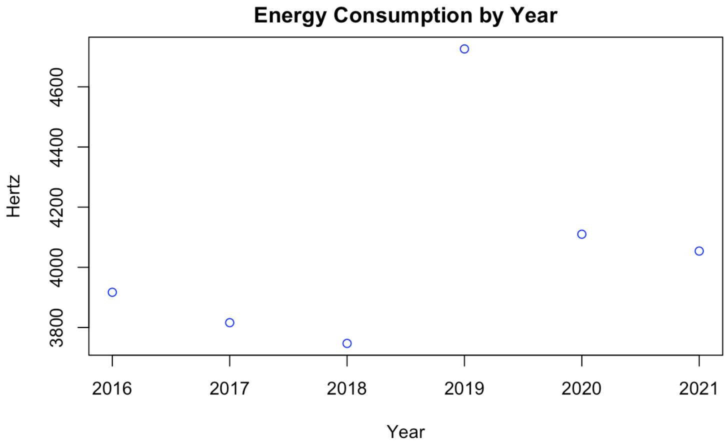
$$\beta_{energy} = -142769.79 + 72.74(year)$$

This gave a predicted energy consumption in Colorado on May 1st, 2022 between 5 pm and 6 pm to be 4310.49 hertz or 7.93 kWh. This does seem like a low estimate, but due to the regression equation only 6 points of data, it is not a bad estimate, but adding more variables might be helpful.

The next variable added to the regression equation was the average temperature in the last 6 years (2016 - 2021) on May 1st in Colorado Springs. After running this multiple linear regression model, I found a considerably higher adjusted R-squared value (0.749) in comparison to the first model (-0.0645). This tells me that adding the covariate ‘temp’ helps the model's predictive power. This gave the new model as

$$\beta_{energy} = -464200 + 232.7(year) - 22.24(temp)$$

Although I will not know for sure exactly what temperature it will be tomorrow, the National Weather Service gives an estimate of 62 degrees Fahrenheit. This, in turn, gives an estimated energy consumption of 4940.52 hertz, or 9.09 kWh. This is a significant jump from the prediction that the first model gave. The second model gives both the variable ‘year’ and the variable ‘temp’ a p-value of 0.03, meaning that these two predictors are significant in predicting energy usage. This is now the best prediction.



The next method tried was to fit an ARMA and AR model to the vector of energy consumption between 2016 to 2021. I started by trying an ARMA(1,1) process. This gave the equation

$$X_t = 4075.485 + 0.4074X_{t-1} + e + e_{t-1}$$

The log-likelihood for this model came out to be 42.61 and had an AICC number of 93.21. Now I would like to compare these numbers with the outcome of using an AR(1) process. Fitting the data to an AR model outputs the equation

$$X_t = 4063.8614 - 0.0917X_{t-1}$$

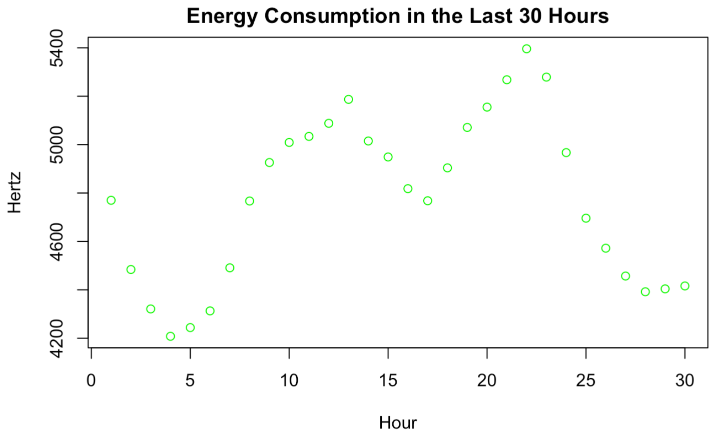
The estimate for phi in this model came out negative rather than in the ARMA model where it came out to be positive. The AR(1) model had a log-likelihood of 43.14 and an AICC number of 92.28. Due to both the log-likelihood and AICC being higher in the AR model than in the ARMA model, I decided to do the next prediction using the AR(1) model. This estimated 3692.1096 hertz of energy consumption in Colorado on May 1st, 2022 between 5 and 6 pm.

I tried to fit an AR(2) and an AR(3) process to see if adding any more terms into the model would give a better prediction, both models gave the exact same phi coefficient as -0.214 and an intercept of 4075.345. The log-likelihood came out to be 42.98, which is lower than the AR(1) model, but not by much. Both models also had an AICC of 93.96, which is only slightly higher than the AICC in the AR(1) model. The model given by using an AR(2) (p parameters over 2 apply as well) was

$$X_t = 4075.3449 - 0.1327X_{t-1} - 0.2144X_{t-2}$$

Which in turn gave the prediction of 2656.195. This prediction is worse than the prediction obtained by using the AR(1) process, and therefore that is the best prediction thus far.
I decided to next look at just a day's worth of data, so I took the most recent 30 hours of energy consumption in Colorado and ran a linear regression with the same variables as the one prior, the temperature at every hour and the hour itself. This model gave very significant predictors and an adjusted R-squared of 0.3275 which is sufficiently smaller than the first multiple linear regression model ran. The model was given by

$$\beta_{energy} = 2739.273 + 17.650(day) + 29.452(temp)$$



Using this model, and the exact temperature prediction of 62 degrees, gave an estimate of 5112.447.
I also tried fitting an AR(1) model to this data like the last data set.

$$X_t = 4711.5480 + 0.8950X_{t-1}$$

This model turned out to not be the most accurate due to the difference in how much data there was. The prediction was 8663.868, which is too high to accept.

Conclusion

Although many models were tried and many different applications were used to get said models, the best model was the model using the last 30 hours of energy consumption and multiple linear regression was the best prediction. I believe that because these last hours were on the weekend, and because they are very close to the time of the predicted value, this estimate is better than the others that I computed. Although I was hoping to be able to use ARMA models, it is difficult to obtain an ARMA model that fits data as easily and well as regression models. This might be because of the added parameters. For example, using an ARMA(1,1) is different than using an ARMA(1,2). To try to find the best parameters to get the model to fit the past data and give the best predictions is a long process, one that regression has perfected and is made easy. And although regression is usually an easier and quicker process, I do believe that ARMA processes will catch up and be used for more forecasting in the future.

References

https://forecast.weather.gov/MapClick.php?CityName=Colorado+Springs&state=CO&site=PUB&textField1=38.8632&textField2=-104.76&e=0#.Ym21lxPMLvU

https://weatherspark.com/h/d/3685/2021/5/1/Historical-Weather-on-Saturday-May-1-2021-in-Colorado-Springs-Colorado-United-States#Figures-Temperature