Estimating Yokohama's Average Temperature

Bella Genolio (isabellagenolio@umail.ucsb.edu)

Abstract:

Day-to-day weather prediction is important to the majority of people around the world. It allows us to know what to wear, how to plan our day, and what transportation to use. Luckily, weather and temperature follow a seasonal time series and a simple model can be used to predict future values. This project takes Yokohama, Japan's daily average temperature time series data and uses a SARIMA model to fit the data and predict future temperatures. It then uses a second model, AIMAX, to fit the data again to make a model based on extraneous variables - average rainfall and sunshine duration. From these models, I was able to predict future values and see how other variables in the dataset impact the average temperature.

Introduction

My project takes in the average daily temperature in Yokohama, Japan from a study on cicada emergences from September 1st, 2002 to December 31, 2005. I initially chose this dataset in order to study the time series of cicada emergence in Japan. However, the data recorded only included two months out of the year - July and August. That means that around eighty percent of the data was missing and according to the research this is because there were virtually no cicada shells collected. I could not do individual years either, since there would be only sixty data points and it would be much more difficult to fit a model to such a small dataset. Instead, I switched to the observed average daily temperature as collected by this dataset so that I could fit a model and predict with this abundant dataset.

This dataset was originally used to examine the effects of meteorological factors on cicada emergence. What the researchers found was that the primary drive for cicada emergence was the emergence of the opposite sex, especially for male cicadas. They also found that temperature had a small, consistent effect on emergence and "precipitation and humidity were causally related to emergence" (Sugiura, 2021).

My goal for this dataset is simpler than its initial use - observe change in temperature and predict upcoming temperatures based on seasonality. I am applying a seasonal ARIMA model and an ARMAX model in order to achieve this. Predicting future temperatures will help people in the area prepare for the seasons, day to day weather, weather during special events, and any other parts of life where temperature is a factor.

Data

The original time range of this data set was from September 1st, 2002 to December 31st 2012 but for my project, I truncated the data so that the range only went until December 31st, 2005. This allowed me to maneuver a smaller dataset and make my models more focused. Besides date and average temperature, there are a number of other variables. These include hourly and daily rainfall, sunshine duration, average daily humidity, average vapor pressure, and maximum, minimum, and average daily temperature. The original reason for each of these variables was to examine the effects of weather on cicada emergence. Each observation in this data set is a single day, so the original dataset had 3,775 observations. The truncated data that I am using contains 1,218 observations - around a third of the original size.

Since the location of this data set is Yokohama, Japan, the average temperature is collected in celsius, as well as the highest and lowest temperatures. As for the other variables, cicada counts are integers, rainfall is measured in millimeters, sunshine duration is hours, vapor pressure is hectopascals, and humidity is a percentage. All of the weather variables come from the Japan Meteorological Agency website. The cicada counts come from shells collected by elementary and junior high students in the area (Sugiura, 2021).

Meteorological effects on cicada emergence is a phenomenon that has been studied by few. However, this insect has a great influence on its ecological surroundings and understanding how its environment impacts its emergence from the soil is crucial to find out the mechanisms behind the cicada. This is why Wataru Mukaimine, Kazutaka Kawatsu, and Yukihiko Toquenaga created and implemented this study.

Methodology

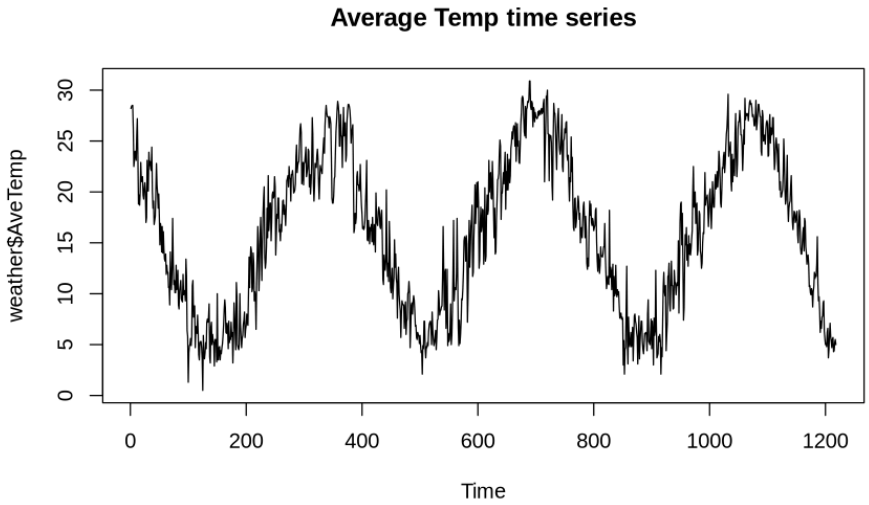
In order to predict future average temperatures of this project, I used two different methods: the SARIMA (p, d, q) x (P, D, Q) model and the ARMAX (p,q,b) model. Each of these models allow me to predict future values and see how other variables in the dataset impact the average temperature.

The first model I am using is the SARIMA (p, d, q) x (P, D, Q), the seasonal autoregressive integrated moving average model. Like the ARIMA model, the AR part of this model explains the past values, the I indicates differencing which makes the data stationary, and the MA explains the previous error. In addition, the SARIMA model also takes into account seasonality. With these variables, we are then able to forecast values in the data set by following

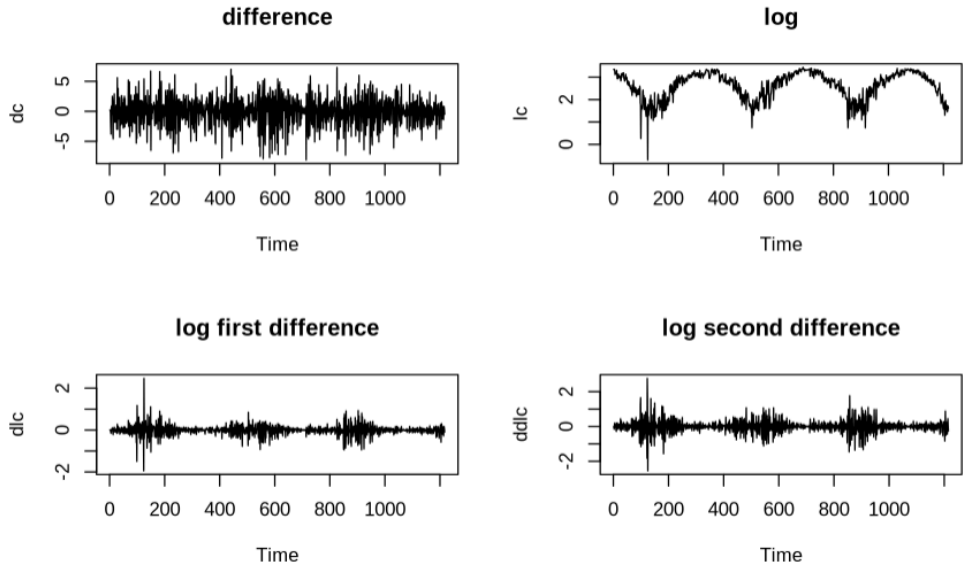
For the second model, I chose to use the ARMAX (p,q,b) model with three variables - average temperature, sunshine duration, and total daily rainfall. All of these variables are in the same dataset and are daily values, so it is simple to create a time series model using these. Like the SARIMA model, the ARMAX or autoregressive moving average model with exogenous inputs model is an expansion of the classic ARMA model. The input values are slightly different, it takes the same auto regressive p variable and moving average q variable, though the b variable represents the exogenous inputs terms. These variables allow for the modeling of the average temperature using past values of itself, past errors, and external predictors.

Results

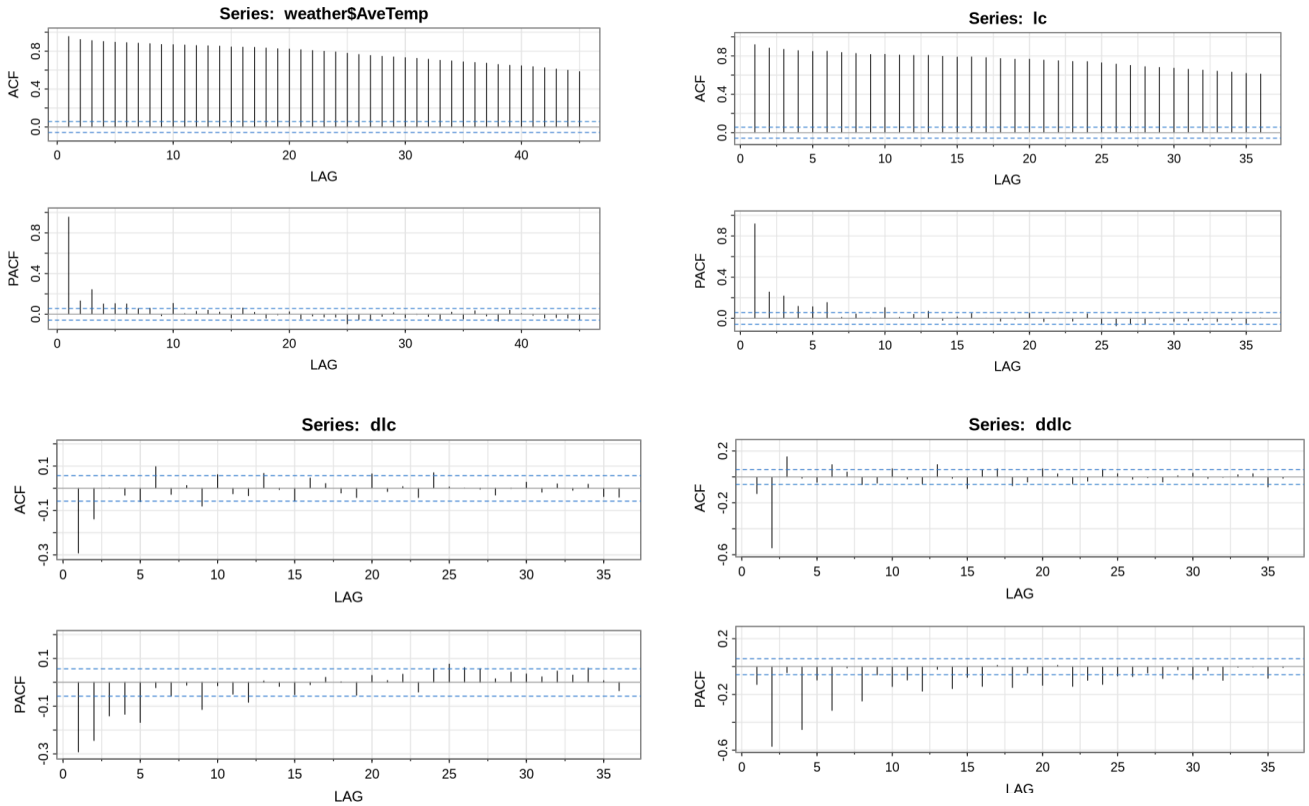
In order to visualize the data, I plotted the average temperature on a regular time series graph:



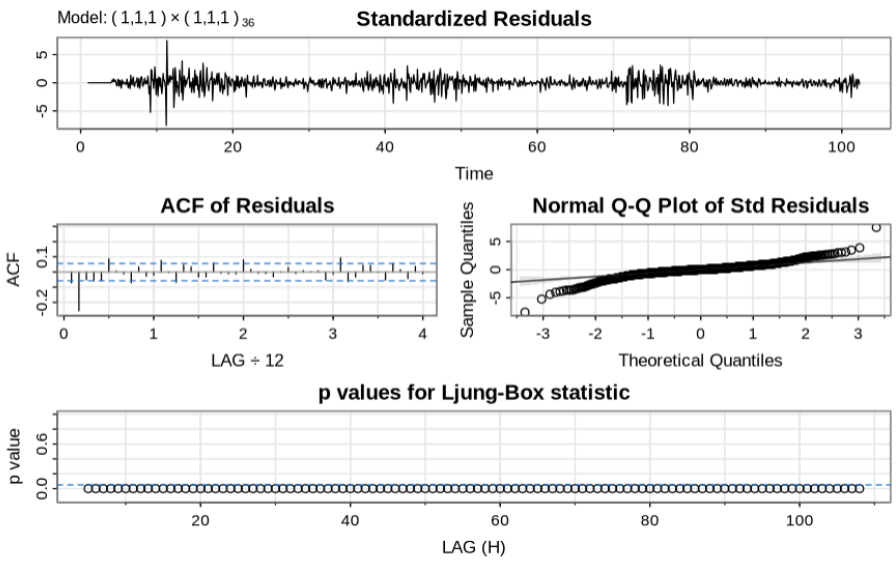
As you can see, it follows a fairly regular seasonal curve, this data starts in September and ends in December three years later, so we can see the start of the decline in temperature at the beginning of the plot, the four winter lows, and the three summer peaks. I then transformed the data, taking the difference, log, log first difference, and log second difference. Here are the plots:



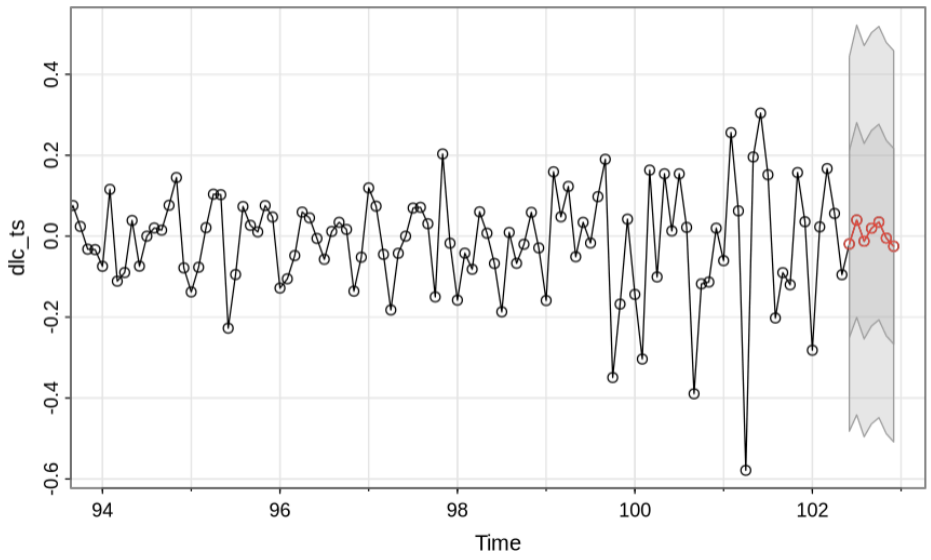
As you can see, these plots completely transform the data, but with the three log-transformed plots we can still see the same peaks and valleys as the original dataset. Now that we have our transformed data, we can take a look at all of the ACF and PACF plots:



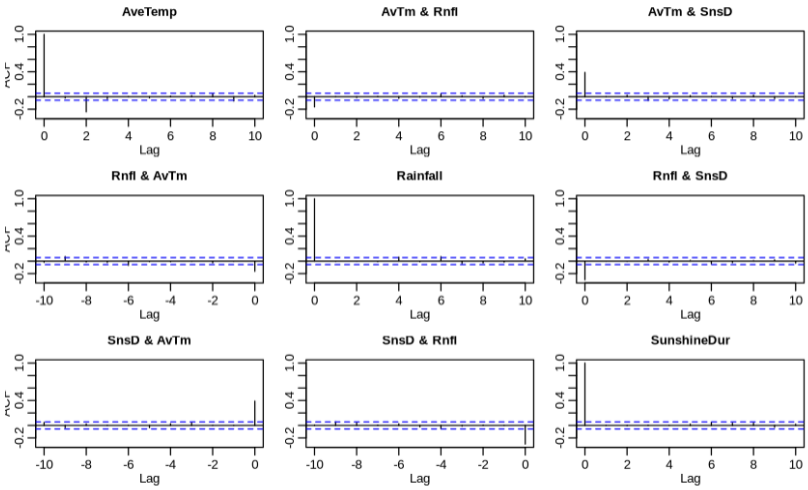
From these plots, I determined that the best data to use was the first difference log transformation. First, I transformed it into a time series dataset so that the analysis would work properly. I then compared two plots where the maximum lag was 12 and 36, and decided to stick with 36 for my seasonal ARIMA model. Now it was time to select p, d, q and P, D, Q candidates. I went through four versions of the SARIMA model with different values and ended up with the SARIMA (p = 1, d = 1, q = 1, P = 1, D = 1, Q = 1, S = 36) model because the AICc was the closest to zero at -0.0003.



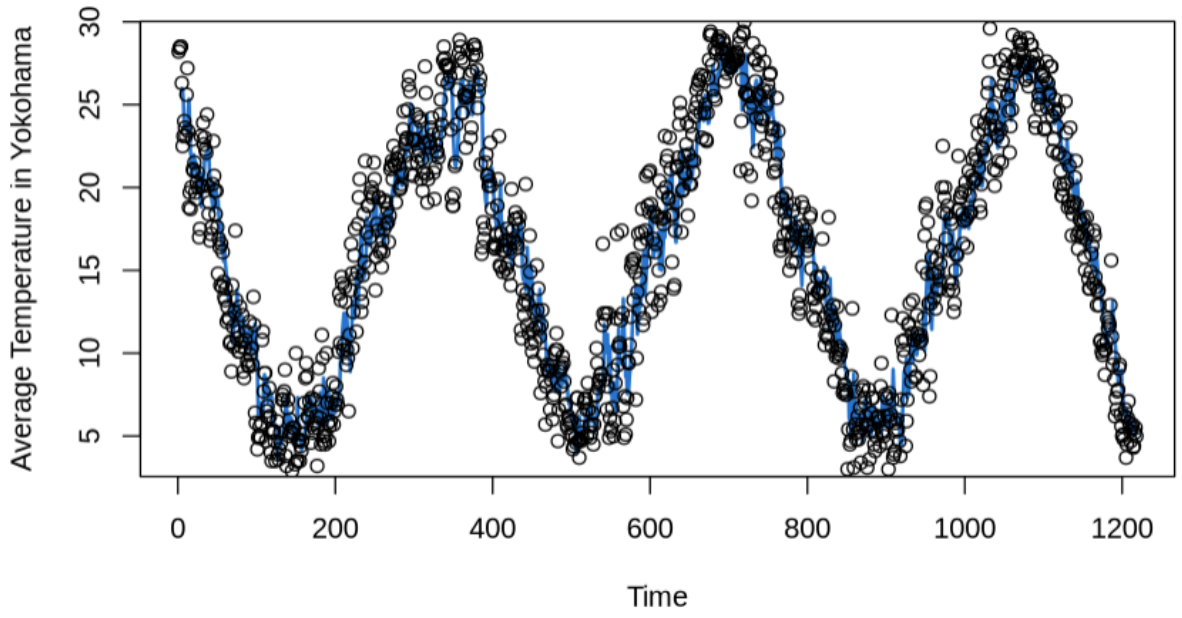
As we can see from the residuals plot, this SARIMA model was a good fit for our data. Finally, I plotted the forecasting model with seven predicted points to represent predicting a week of average temperatures.



The next model that I fit was the ARMAX model. I began by combining the average temperature, average rainfall, and sunshine duration so I could then fit a vector AR model via least squares.



Then I determined that the best lags for the ar portion were 2 and 3, and for the ma portion the lag was 2. Then it was time to fit the model.



As you can see, the fit was very close to the actual points with only a few that were slightly off the path. We can see that rainfall and sunshine duration are helpful predictors for the average temperature.

Conclusion and Future Study

In this project, I analyzed the average daily temperature in Yokohama, Japan, using two different time series models: the SARIMA and ARMAX models. The SARIMA model effectively captured the seasonal patterns in temperature, allowing for accurate short-term forecasting. The ARMAX model incorporated external meteorological factors—sunshine duration and total daily rainfall—to aid in fitting the average temperature time series.

Throughout the analysis of this project, I found that both models provide valuable insights into temperature prediction. The SARIMA model allowed me to take seasonality into account, while the ARMAX model incorporated other weather prediction variables. With these two models, local residents and meteorologists can anticipate coming temperatures, allowing for easy planning of daily activities, anticipation of necessary transportation, and preparation for seasonal events.

While the study I completed focuses solely on average temperature forecasting, a future expansion of this research could include more complex models that predict other variables as well, such as rainfall or sunshine duration. This expansion could also include hourly temperature predictions rather than daily. Another area that this research could extend to is longer climate patterns. With more data than just three years, this research could assess warming temperatures due to climate change in the area, multi-year patterns, and repeated high and low temperatures over time.

References

Japan Meteorological Agency. "日ごとの値: 横浜（神奈川県) 2002年9月（日ごとの値） 主な要素." *過去の気象データ検索*, 2002, https://www.data.jma.go.jp/stats/etrn/view/daily_s1.php?prec_no=46&block_no=47670&year=2002&month=9&day=1&view=.

"SARIMA (Seasonal Autoregressive Integrated Moving Average)." *GeeksforGeeks*, https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/.

Shumway, Robert H., and David S. Stoffer. *Time Series Analysis and Its Applications.* Springer. Fourth Edition, 2017.

Sugiura, K., & Tainaka, K. (2021). Data from: Long-term monitoring of cicada emergence reveals effects of climate change on phenology. Dryad, Dataset. https://doi.org/10.5061/dryad.6wwpzgmxf

Wikipedia contributors. "Autoregressive Integrated Moving Average." *Wikipedia, The Free Encyclopedia*, Wikimedia Foundation, March 21, 2025. https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average.

Appendix

SARIMA code:

```
## model setup
dc <- diff(weather$AveTemp)
lc <- log(weather$AveTemp)
dlc <- diff(lc)
ddlc <- diff(dlc, 2)

par(mfrow = c(2,2))
ts.plot(dc, main = "difference")
ts.plot(lc, main = "log")
ts.plot(dlc, main = "log first difference")
ts.plot(ddlc, main = "log second difference")
```

```
# look at P/ACF of plot without transformations
acf2(weather$AveTemp)

## initial P/ACF plots
acf2(lc, max = 36)
acf2(dlc, max = 36) # lets use this
acf2(ddlc, max = 36)

# in order to get the forecast to work, must transform to ts
dlc_ts <- ts(dlc, frequency = 12)
# take a look at just dlc
acf2(dlc, max = 12)
acf2(dlc, max = 36) # this looks best

# determine which (p,d,q) candidates
sarima(dlc_ts, p = 1, d = 1,q = 1, P = 1,D = 1,Q = 1, S = 36)  # AICc = -0.0002923867, this is very good
sarima(dlc_ts, 1,0,1, 1,1,1, 36)  # AICc = -0.1409866, this is worse
sarima(dlc_ts, 1,1,1, 0,1,1, 36)  # AICc = -0.001353737, removing D did not help
sarima(dlc_ts, 1,1,0, 1,1,1, 36)  # AICc = 0.6503312, much worse

# we can see the first choice is our best
sarima.for(dlc_ts, n.ahead = 7, p = 1, d = 1, q = 1, P = 1, D = 1, Q = 1, S = 36)
```

ARMAX code:

```
AveTemp <- as.numeric(as.character(weather$AveTemp))
Rainfall <- as.numeric(as.character(weather$Rainfall))
SunshineDur <- as.numeric(as.character(weather$SunshineDur))

x = cbind(AveTemp, Rainfall, SunshineDur)
summary(VAR(x, p=1, type='both'))

VARselect(x, lag.max=10, type="both")
summary(fit <- VAR(x, p=2, type="both"))

acf(resid(fit), 10)
serial.test(fit, lags.pt=12, type="PT.adjusted")
```

```
library(marima)

# define the model shape, how many variables, and p,q values
model <- define.model(kvar=3, ar=c(2,3), ma=c(2))
arp <- model$ar.pattern # extract the AR and MA patterns
map <- model$ma.pattern

# define the residuals for the average temperature data
avetemp.d <- resid(detr <- lm(AveTemp~ time(AveTemp), na.action=NULL))

# define the data as a matrix using the residual value for average temperature and the other two variables
xdata <- matrix(cbind(avetemp.d, Rainfall, SunshineDur), ncol=3)

# fit the data to the ARIMAX model
fit <- marima(xdata, ar.pattern=arp, ma.pattern=map, means=c(0,1,1),penalty=1)

# residual analysis
innov <- t(resid(fit)); plot.ts(innov); acf(innov, na.action=na.pass)

pred <- ts(t(fitted(fit))[,1], start=start(AveTemp), freq=frequency(AveTemp)) +
  detr$coef[1] + detr$coef[2]*time(AveTemp)

plot(pred, ylab="Average Temperature in Yokohama", lwd=2, col=4); points(AveTemp)
```