# An Exploration of Climatic Parameters in Delhi, India across time

# 2013 - 2017

## Time Series Final Report

## ST 566

Caitlon Vu (Methods)

Abraham Mendoza Jr. (Methods)

Maryanne Evans (Introduction)

Connor Crane (Discussion)

### Introduction

Temperatures in high-latitude regions have risen at twice the global average, 1.8 °C over the last 30 years (Schuur et al 2015). There is an abundance of research and data to explain climatic trends, particularly due to increasing temperatures, in circumpolar areas, specifically. For instance, moose are expanding their range across Alaska due to widespread vegetation growth, challenging historic community composition. Permafrost thaw is widespread across polar regions, exacerbating the presence of greenhouse gases atmospherically and contributing to further warming. Often, these research questions are addressed using Long Term Ecological Research (LTER) networks and data collection over carefully recorded periods of time. These data are necessary to discuss on-going, emergent trends and predict potential climatic outcomes across the Earth.

Although scientists understand there is rapid change impacting circumpolar regions, nuances in climatic parameters are rarely explored in subtropical and semi-arid regions, given their predictably nuanced subtlety. The health of ecosystems neighboring the equator should be considered for scientific investigation and data analysis, exploring temperature among other abiotic parameters. However, understandably, these regions receive less funding for research when compared to potent, rapidly changing regions of the globe, such as the ecosystems found in the Arctic or Antarctica. Temperature is emphasized as a direct driver, but other parameters such as atmospheric pressure, humidity and wind speed play a role in ecosystem health outcomes, also.

Given the problem of climate change is global, our group of data scientists chose to explore a climatic dataset from the open-source, community website “Kaggle” collected from the years 2013 through 2017, recording mean temperature, humidity, wind speed and mean pressure in Delhi, India. **This report** aims to explore the nuanced relationship of temperature in the sub-tropical, semi-arid climate of Delhi, India over four years, and determine if a predictive model generates an on-going seasonality and trend in the time series.

**Hypothesis:** There is seasonality, and a subtle increasingly trend, in temperatures over four years for climatic data from Delhi, India.

### Methods

#### Exploration

A series of statistical methodologies were applied to these data to explore four parameters of mean temperature, humidity, wind speed and mean pressure in Delhi, India across time from January 1st, 2013, and April 24th, 2017. The first step for analyzing these climatic parameters is plotting the raw time series and viewing temperature as the response variable against monthly data collection across time. It was found that these time series data are not stationary because the mean changes over time. Furthermore, it was noted that there is seasonality present. Since there is no change in variance over time, a transformation was not necessary nor was a transformation required to stabilize these data. Therefore, a log transformation was not applied or explored. The plot of the raw time series data shows a strong seasonality, with interval patterns or periodic fluctuations every 365 days.

Across these four years of data, there are four seasonal periods. This intuitively makes sense because the temperature changes are likely caused by seasons throughout the year. Therefore, we will estimate the seasonal component as 365 days in the process of making the time series stationary. There is a gradual upward trend over time (**Figure 1**). Meaning, we need to remove seasonality and trend to achieve stationarity by fitting a linear regression of the time series and analyzing residuals.

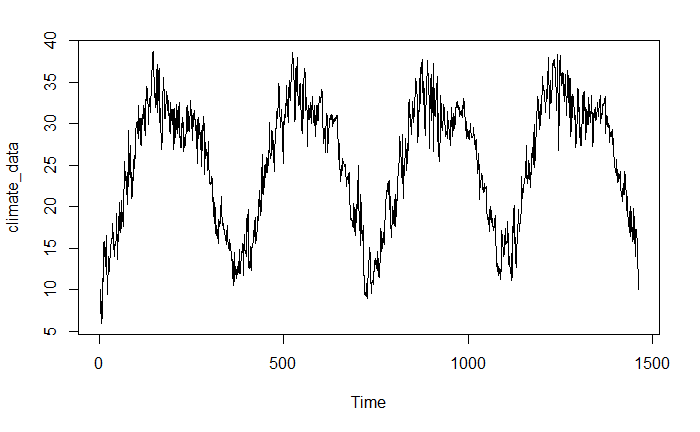


Figure 1. Climate time series data plotted before removing trend and seasonality.

#### Removing Trend and Seasonality

We remove the trend from the time series by extracting the residuals from the linear model. The time series no longer shows an upward trend (**Figure 2**). The mean is constant over time, but seasonality remains.

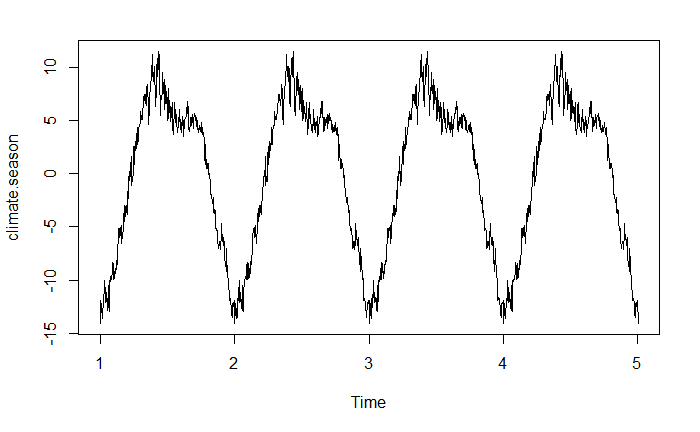


Figure 2. Climate time series data plotted after removing the positive trend.

We estimate the seasonality from the residuals of the linear model, and then remove the seasonality. We observe that the ‘Random’ plot after removing seasonality behaves like white noise. Therefore, we have successfully removed trend and seasonality from the time series to obtain a stationary series (**Figure 3**).

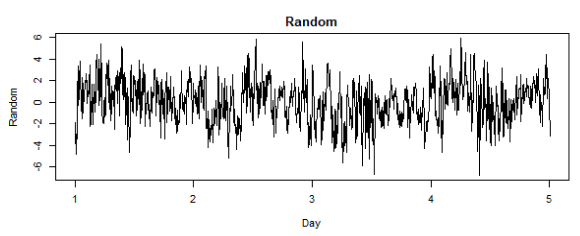


Figure 3. Stationary times series after removing trend and seasonality resembles a white noise process.

##### Fitting ARMA Model

We will fit an **ARMA model** to the stationary time series by making conclusions from the ACF and PACF plots. These plots show where autocorrelation may occur in these time series data. We observe that the ACF plot slowly decreases to zero and the PACF plot has a non-zero first lag. There are many models that may work for these data, therefore several models will be compared. There are also non-zero PACF values at the period lags, suggesting fitting a SARIMA model may be appropriate. Based on the ACF and PACF plots, we chose to compare a few model candidates: AR(1), AR(2), AR(3), ARMA(1,1), ARMA(2,2), ARMA(2,1), and ARIMA(1,1,0)x(0,1,0)12. The AIC value is used to compare these models, where the ARMA(2,1) model returns the smallest value of all the other models at 4990 (**Figure 4,** all model results in appendix).

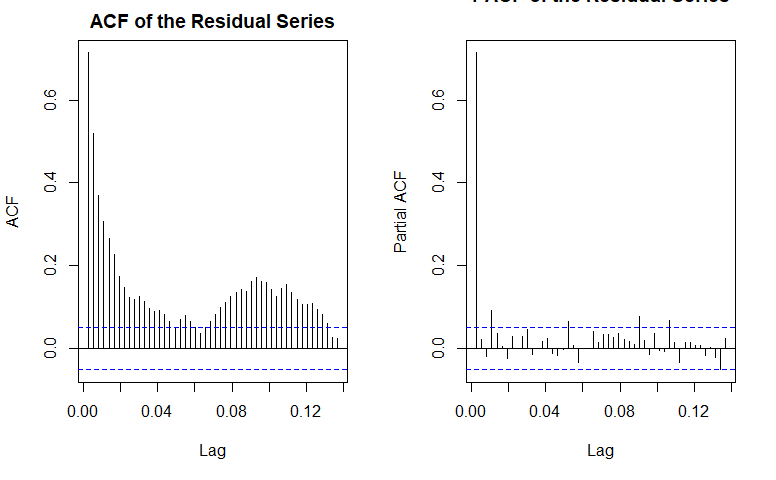


Figure 4. ACF and PACF plots for the stationary climate time series.

#### Model Diagnosis and Analysis

We check the residuals of our selected ARMA(2,1) model to evaluate the fit. The model is evaluated with diagnostics using R function “*tsdiag()*”*.* The residuals appear reasonable, and the process is centered at *zero* without a trend (**Figure 5, top**).

The residual ACF and PACF plots resemble white noise, and there are no erroneous autocorrelations in the residuals (**Figure 5, middle**, PACF plot can be seen in appendix). The p-values are larger than 0.05, which suggests residuals come from white noise (**Figure 5, bottom**). The ‘*qqplot*’ supports the assumption of normal distribution of the residuals because the data points fall along the straight line (**Figure 6**). The model fitting performs reasonably well, which is supported by these data explorations.

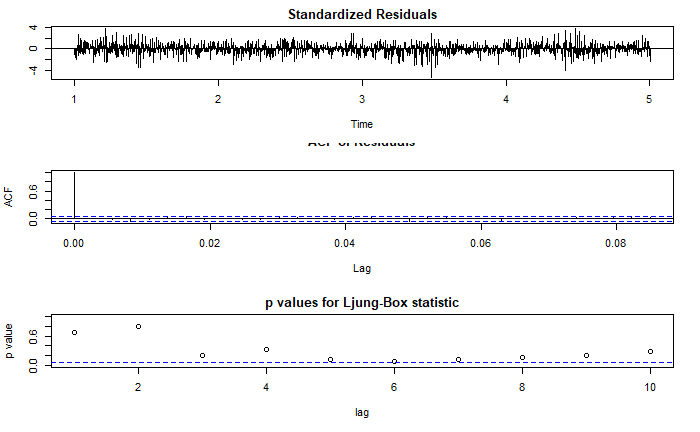


Figure 5. Results from tsdiags() show a well-fitted model.

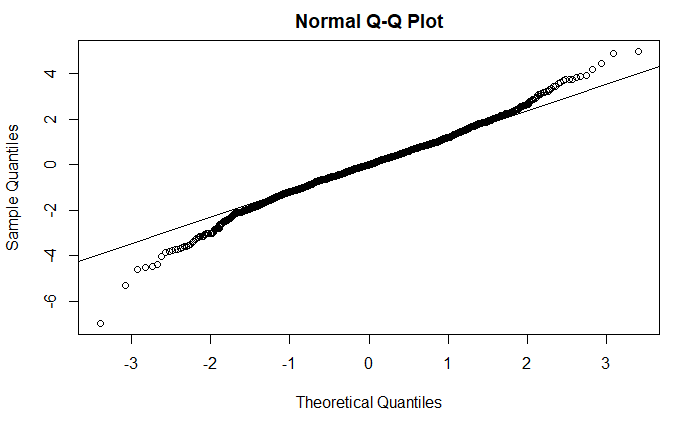
****

Figure 6. ‘qqplot’ of the residuals support assumption of normal distribution.

#### Model Prediction and Forecasting

To forecast the next year of temperature data, we can use the R function “*predict()*” to predict the next 365 (annual) data points. We add the trend and seasonality back into the time series model for the prediction to obtain the plot of the newly predicted data below.

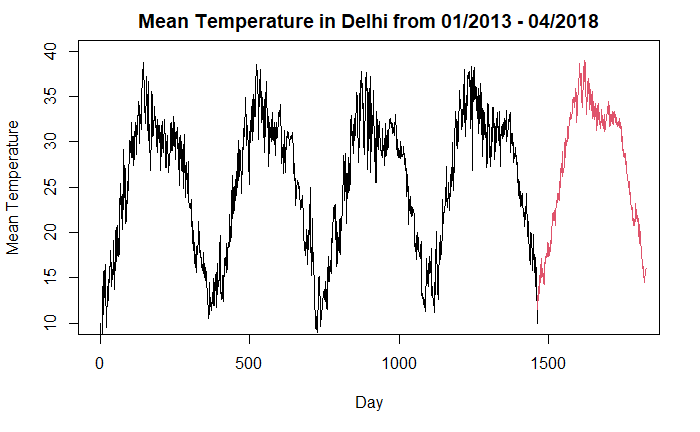


Figure 7. Predicted values from the fitted ARMA(2,1) model for the next year of temperature.

### Discussion

Notes for Connor’s results/discussion:

* Better results may have been obtained by comparing more ARMA or SARIMA models
* Alternative approach would have been to using function diff() to difference the time series data
* If it were provided, we could have also used a ‘test’ data set to evaluate the performance of our data by comparing our ARMA(2,1) predictions with real results
* Overall the ARMA(2,1) model performs fairly well and the forecast reflects the behavior of the time series very well.
* The forecast shows increasing temperature in the following year, which is what we expected

### Appendix

|  |
| --- |
| --- |
|  | ## Introduction |
|  |  |
|  | ### Exploration |
|  |  |
|  | Data |
|  |  |
|  | ```{r} |
|  | climate\_data <- read.csv("DailyDelhiClimateTrain.csv") |
|  | climate\_data <- ts(climate\_data$meantemp) |
|  | plot(climate\_data) |
|  | ``` |
|  |  |
|  | ### Fit Linear Trend to the Climate Data |
|  |  |
|  | ```{r} |
|  | climate.time <- time(climate\_data) |
|  | climate.fit.trend <- lm(climate\_data ~ climate.time) |
|  | climate.pass.trend <- ts(climate.fit.trend$fitted.values, deltat = 1/12) |
|  | head(climate.pass.trend) |
|  | ``` |
|  |  |
|  | ### Removing Trend & Seasonality |
|  |  |
|  | ```{r} |
|  | #Removing Linear Trend |
|  | climate.res <- climate.fit.trend$residuals |
|  |  |
|  | #Formatting into time series |
|  | climate.res <- ts(climate.res, deltat = 1/365) |
|  |  |
|  | #Remove Seasonality |
|  | climate.day <- factor(cycle(climate.res)) |
|  | fit.season <- lm(climate.res ~ climate.day) |
|  | climate.season <- ts(fit.season$fitted, deltat = 1/365) |
|  | plot(climate.season) |
|  |  |
|  | climate.rand <- ts(climate.res - climate.season, deltat = 1/365) |
|  | ``` |
|  |  |
|  | ### Plotting Trend, Seasonality, and Residuals |
|  |  |
|  | ```{r} |
|  | par(mfrow = c(3,2)) |
|  | #Original Time Series Plot |
|  | plot(climate\_data, xlab = "Day", ylab = "Mean Temp", main = "Original Time Series") |
|  |  |
|  | #Plotting Trend |
|  | plot(climate.pass.trend, xlab = "Day", ylab = "Trend", main = "Trend") |
|  |  |
|  | #Plotting Seasonality |
|  | plot(climate.season, xlab = "Day", ylab = "Seasonality", main = "Seasonality") |
|  |  |
|  | #Plotting White Noise |
|  | plot(climate.rand, xlab = "Day", ylab = "Random", main = "Random") |
|  |  |
|  | #Plotting Resduals |
|  | plot(climate.res, xlab = "Day", ylab = "Residuals", main = "Residuals") |
|  | ``` |
|  |  |
|  | ### Fitting ARMA Models |
|  |  |
|  | ```{r fig.height=5} |
|  | #Checking ACF and PACF Plots |
|  | par(mfrow = c(1,2)) |
|  |  |
|  | acf(climate.rand, main = "ACF of the Residual Series", lag.max = 50) |
|  | pacf(climate.rand, main = "PACF of the Residual Series", lag.max = 50) |
|  | ``` |
|  |  |
|  | Trying Different ARIMA models and SARIMA models and looking at lowest AIC Values |
|  |  |
|  | AR(1) |
|  |  |
|  | ARMA(1,1) |
|  |  |
|  | ARMA(2,1) |
|  |  |
|  | ARIMA (1,1,0) X (0,1,0) |
|  |  |
|  | ```{r} |
|  | #AR 1 Model |
|  | fit.ar1 <- arima(climate.rand, order = c(p = 1, d= 0, q = 0), method = "ML", include.mean = F) |
|  |  |
|  | #ARMA(1,1) Model |
|  | fit.arma11 <- arima(climate.rand, order = c(p= 1, d = 0, q = 1), method = "ML", include.mean = F) |
|  |  |
|  | #ARMA(2,1) |
|  |  |
|  | fit.arma21 <- arima(climate.rand, order = c(p= 2, d = 0, q = 1), method = "ML", include.mean = F) |
|  |  |
|  | #ARIMA(1,1,0)x(0,1,0) |
|  | fit.s.ar1 <- arima(climate.rand, order = c(1, 1, 0),seasonal = list(order = c(0, 1, 0), period = 12)) |
|  |  |
|  | fit.ar1 #AIC Value 4990.84 |
|  | fit.arma11 #AIC Value 4992.3 |
|  | fit.arma21 #AIC Value 4990.47 |
|  | fit.s.ar1 #AIC Value 6113.25 |
|  | ``` |
|  |  |
|  | Fitting a more complicated AR(2), AR(3), & ARMA(2,2) |
|  |  |
|  | ```{r} |
|  | #AR 2 Model |
|  | fit.ar2 <- arima(climate.rand, order = c(p = 2, d= 0, q = 0), method = "ML", include.mean = F) |
|  |  |
|  | #AR 3 Model |
|  | fit.ar3 <- arima(climate.rand, order = c(p = 3, d= 0, q = 0), method = "ML", include.mean = F) |
|  |  |
|  | fit.ar2 #AIC Value of 4992.27 |
|  | fit.ar3 #AIC Value of 4993.74 |
|  |  |
|  | #ARMA(2,2) |
|  |  |
|  | fit.arma22 <- arima(climate.rand, order = c(p= 2, d = 0, q = 2), method = "ML", include.mean = F) |
|  | fit.arma22 #AIC Value of 4991.64 |
|  | ``` |
|  |  |
|  | Model of Preference: ARMA(2,1) Code Above |
|  |  |
|  | ### Model Diagnosis |
|  |  |
|  | ```{r} |
|  | tsdiag(fit.arma21) |
|  | ``` |
|  |  |
|  | Checking for Normality |
|  |  |
|  | ```{r} |
|  | res <- fit.arma21$residuals |
|  | qqnorm(res) |
|  | qqline(res) |
|  | ``` |
|  |  |
|  | Checking ACF and PACF Plots to see if Residuals appear to be white noise |
|  |  |
|  | ```{r} |
|  | par(mfrow = c(1,2)) |
|  | acf(res, main = "ACF for Residuals") |
|  | pacf(res, main = "PACF for Residuals") |
|  | ``` |
|  |  |
|  | ### Forecasting |
|  |  |
|  | ```{r} |
|  | #Predicting Random Section |
|  | ran.pred <- predict(fit.arma21, n.ahead = 365) |
|  |  |
|  | #Predicting Trend |
|  | pred.time <- seq(1463, by = 1, length = 365) |
|  |  |
|  | trend.pred <- predict(climate.fit.trend, newdata = data.frame(climate.time = pred.time)) |
|  |  |
|  | #Predicting Seasonality |
|  | season.pred <- fit.season$fitted[1:365] |
|  |  |
|  | #Sum of Predictions |
|  | pred <- ts(ran.pred$pred + trend.pred + season.pred, start = c(1463, 1), deltat = 1) |
|  | ``` |
|  |  |
|  | ### Prediction Plot |
|  |  |
|  | ```{r} |
|  | plot(climate\_data, xlab = "Day", ylab = "Mean Temperature", main = "Mean Temperature in Delhi from 01/2013 - 04/2018", xlim = c(0, 1800), ylim =c(10, 40)) |
|  | lines(pred, col = 2) |
|  |  |
|  | ``` |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### Citations

Schuur, E. A. G., A. D. McGuire, C. Schadel, G. Grosse, J. W. Harden, D. J. Hayes, G. Hugelius, C.D. Koven, P. Kuhry, D. M. Lawrence, S. M. Natali, D. Olefeldt, V. E. Romanovsky, K. Schaefer, M. R. Turetsky, C. C. Treat, and J. E. Vonk (2015) Climate Change and the Permafrost Carbon Feedback. Nature 520:171-179.