An Exploration of Mean Temperature in Delhi, India across time 2013 - 2017

Time Series Final Report
ST 566

Caitlon Vu (Methods)
Abraham Mendoza Jr. (Methods)
Maryanne Evans (Introduction, Discussion)
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 resulting from increasing temperatures, specifically in circumpolar areas. For instance, permafrost thaw is widespread across the Arctic and contributes to greenhouse gas emissions, driving rising temperatures. This alarming implication is known because of scientific research focused on polar areas exhibiting permafrost thaw. Often, climate research questions are addressed using Long Term Ecological Research Networks, and carefully recorded data collection for extended periods of time. These data are necessary to discuss on-going, emergent trends and predict potential climatic outcomes across the planet.

Scientists know there is rapid change impacting circumpolar regions, but climatic parameters are rarely explored in subtropical and semi-arid regions, possibly given their predictably nuanced subtleties. The health of ecosystems neighboring the equator should be considered for scientific investigation. It is important to explore not only temperature, but other abiotic parameters such as atmospheric pressure, humidity, and wind speed, across many geographic regions to create a comprehensive understanding of climatic changes over time globally. Arguably, multiple parameters play an equally important role in ecosystem health. Climate change is a global concern, and 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 relationship of mean temperature across time in the sub-tropical, semi-arid region of Delhi, India over a timeframe of four years, and determine if a predictive model generates an on-going seasonality and trend explained in the time series.

> **Hypothesis:** There is seasonality and increasing trends in mean 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.

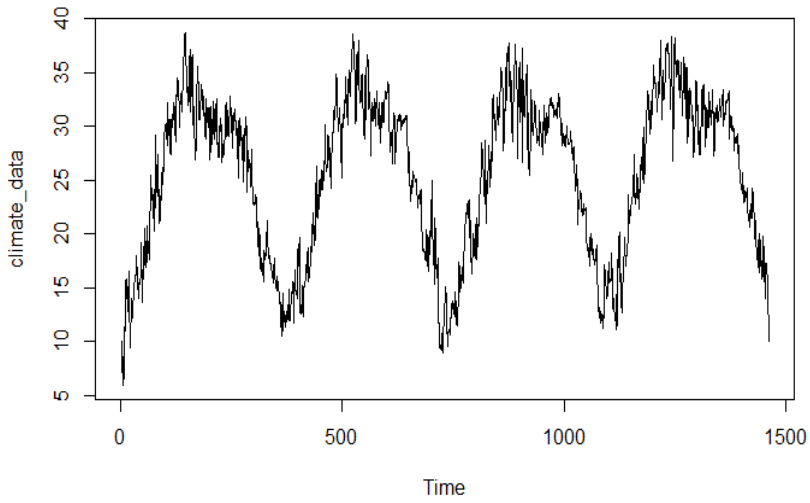

Figure 1. Climate time series data plotted before removing trend and seasonality.

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.

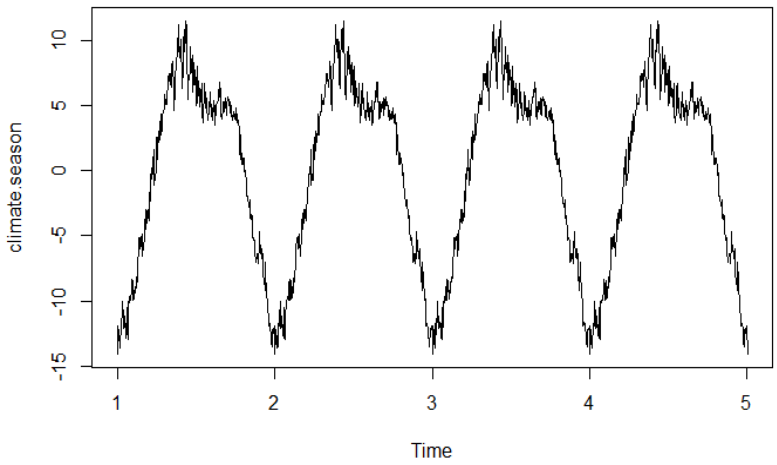

Figure 2. Climate time series data plotted after removing the positive trend.

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**).

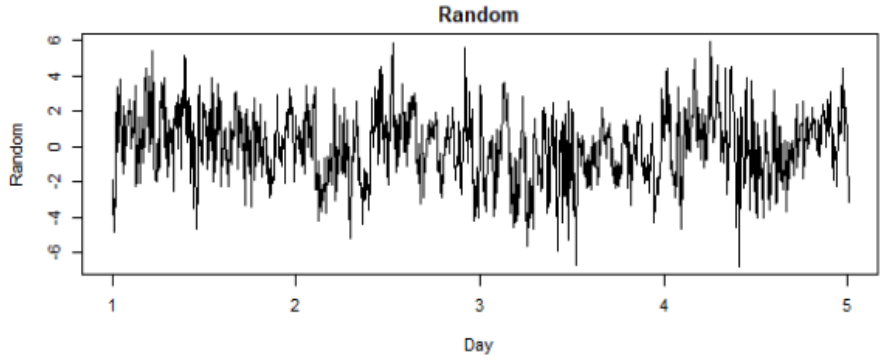

Figure 3. Stationary times series after removing trend and seasonality resembles a white noise process.

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).

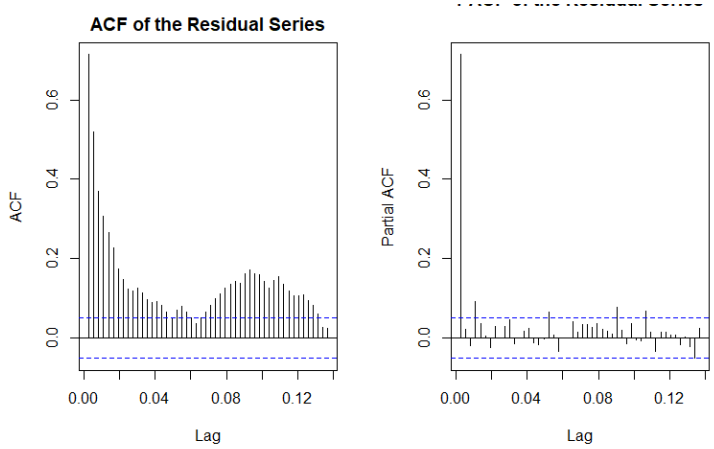

Figure 4. ACF and PACF plots for the stationary climate time series.

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.

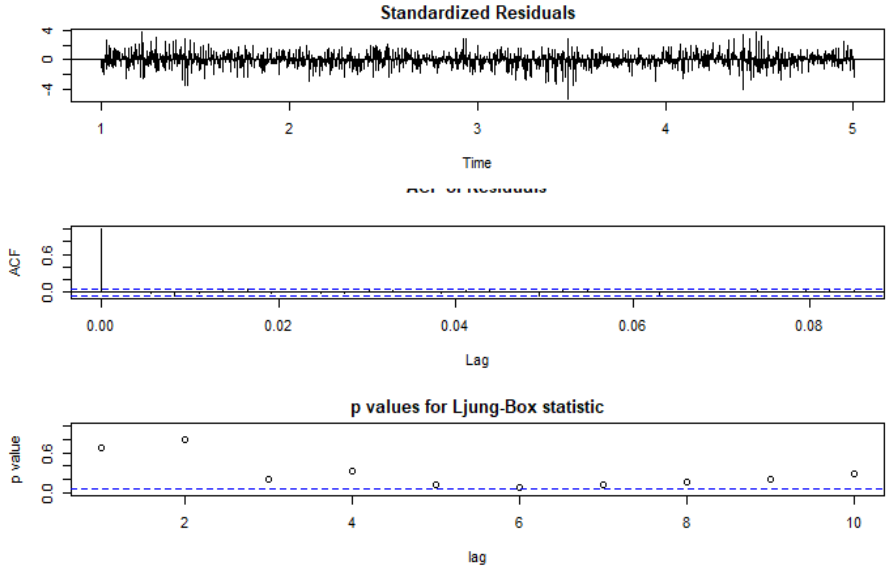

Figure 5. Results from tsdiags() show a well-fitted model.

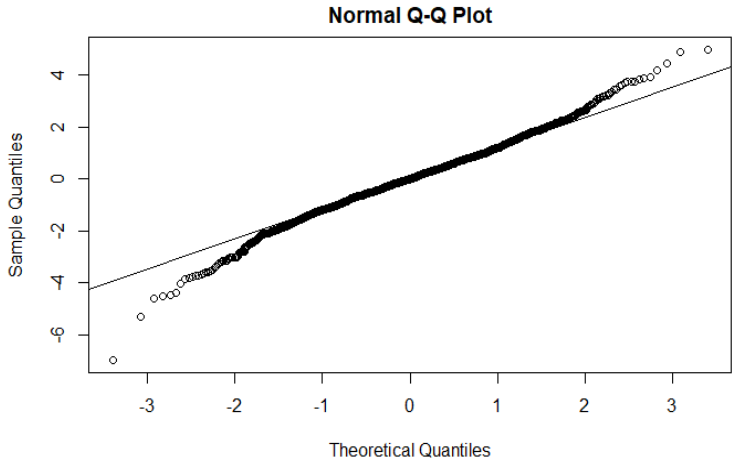

Figure 6. 'qqplot' of the residuals support assumption of normal distribution.

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.

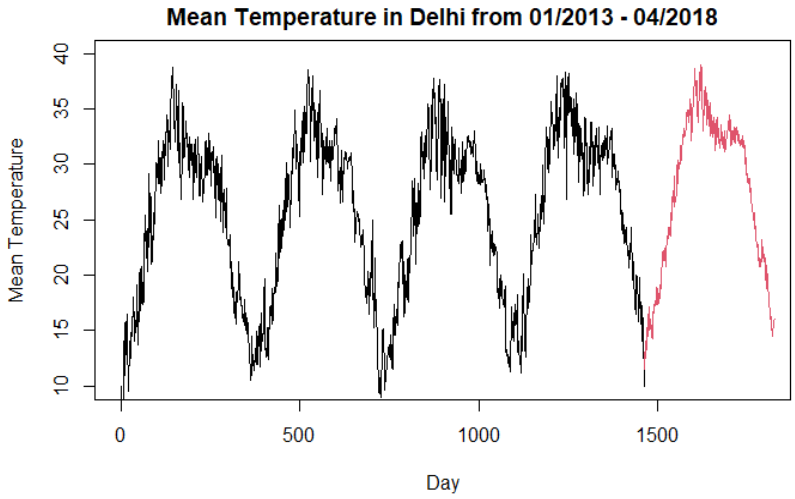

Figure 7. Predicted values from the fitted ARMA(2,1) model for the next year of temperature.

Discussion

The time series analyses (explained in the **Methods Section**) in this report were used to explore our choice dataset of climatic parameters from Delhi, India and postulate about climate change in semi-arid and subtropical regions around the globe.

> **Hypothesis:** There is seasonality and increasing trends in mean temperatures over four years for climatic data from Delhi, India.
>
> **Null Hypothesis:** There is no seasonality or trends in the mean temperature over four years for the climatic data from Delhi, India.

We have decided to **accept our hypothesis**, rejecting the null hypothesis as an outcome of our statistical exploration of the mean temperature dataset, which was collected over four years' time (2013- 2017). **There are strong, statistically supported results from our ARMA(2, 1) model**, suggesting it may be beneficial to explore the proposed models (included in the **R Appendix**), comparing the resulting AIC values, and examining the ACF plots, PACF plots, and possible forecasts. A "*test*" dataset may also benefit this study, offering a more thorough analysis, since these data only account for years up until 2017, leaving a current data gap for the last six years. The forecasts predicted in this report have occurred in real time, limiting these findings to the past (2018), opposed to predicting reasonable future outcomes for Delhi, India's mean temperature. If our group was able to access the mean temperature values recorded with identical protocol used for data collection for the relevant, study Kaggle dataset, we could construct a robust method for evaluating the supported model.

Given additional time, an interesting elaboration for this study includes all the climatic parameters (mean temperature, humidity, mean pressure, and wind speed) to develop a comprehensive exploration of abiotic outcomes for Delhi, India, detecting any other changes that may be occurring in correlation to mean temperature increases. Overall, these results mirror the initial expectations for these data. The chosen model exemplifies expected outcomes, supported by the model diagnostics (**Methods Section**). The ACF and PACF plots show white noise, the QQ plot suggests the normality assumption is satisfied, and the forecast reasonably explains these data. With confidence, we accept the study hypothesis, and predict similar trends for the future in Delhi, India.

R Code 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.