Delhi Climate

Professor Soyer

Forecasting Analytics

Violet Chen - ychen20@gwu.edu

Abdurrahman Tauqueer - abtauqeer@gwu.edu

Stephen Gaffney - gaffney24@gwu.edu

1. Introduction

The data that we chose to do our analysis on was the daily climate data of Delhi, India. Each observation is one day's recorded data. We have one target variable and three predictor variables. Our target variable is meantemp, which is the mean temperature in Celcius for that day. Our first predictor variable is the humidity percentage for that day, which will be represented by a number 0-100. Our second predictor variable is the wind speed in m/s for the day. Our third predictor variable is the mean pressure in hectoPascals (hPa) for the day.

We chose this data set because we are very interested in seeing what effects that time has on Delhi's weather trends. Through analyzing this data we will be able to see not only the effects that our predictor variables may have on our target variable but also the effect that time will have on our target variable. We will be able to analyze if our data exhibits any seasonal/cyclical trends.

2. Univariate Time Series Models

We used a hold-out sample of around 15%. Given that we had a total of 1462 observations (four years), the number of observations used for fit is [1:1250]. The data starts with observation 1 on January 1, 2013, and ends with observation 1250 on June 5th, 2016.

For basic understanding of the data set, we plot the time series plot of our target variable meantemp (shown below on Fig 1), which exhibits seasonality.

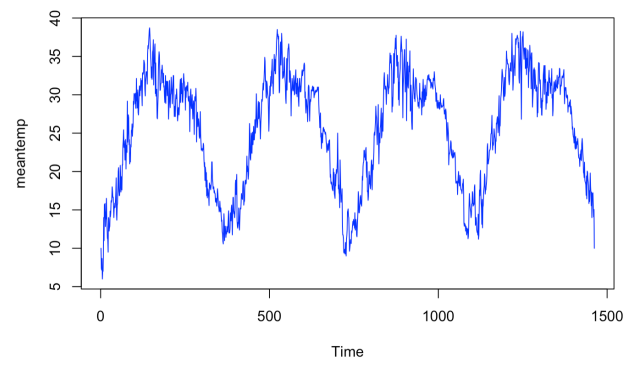


Fig 1

We did boxplots of meantemp against month to further investigate the seasonal behavior (Fig 2) . Looking at the boxplots, the seasonality occurs in a monthly manner.

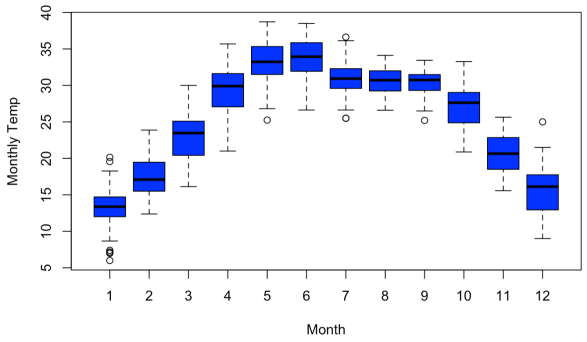


Fig 2

(Our SACF of meantemp is shown below on Fig 3) We did a SACF of meantemp. The SACF decreases slowly with lags, therefore our time series is nonstationary.

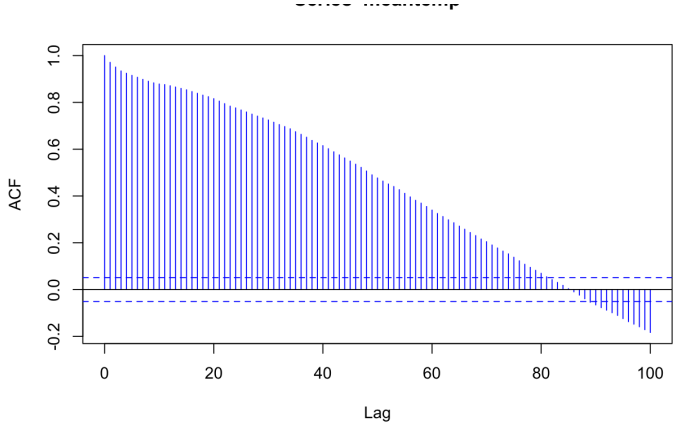
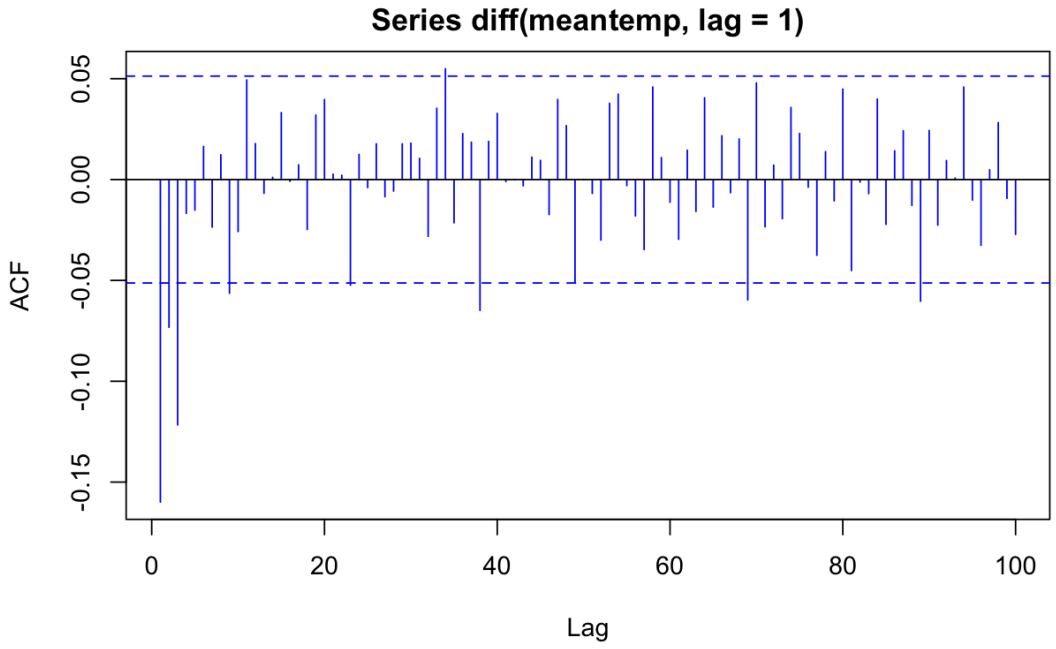


Fig 3

To make the time series stationary, we take the first difference of the time series and plot the SACF of the first difference (Fig 4), which appears to be stationary and has lags above 2 s.e.. From that, we can conclude that the first difference of our time series is not white noise and the series is therefore not a random walk.



Figs 4

(Further investigated in Section 4.3 with Integrated ARMA)

2.1 Deterministic Time Series Models (Seasonal Dummies and Trend, Cyclical Trend)

2.1.1 Seasonal Dummies and Trend Model

Monthly Dummies (summary of fit shown below in Fig 5): the p-value for F-statistics is smaller than 0.05, we reject the null hypothesis that all ßs are not statistically significant. Also, p-value of time and dummies coefficients are smaller than 0.05, so they are significant in this model.

```
Call:
lm(formula = n_meantemp ~ t + as.factor(n_month))

Residuals:
   Min     1Q Median     3Q    Max
-8.684 -1.763  0.106  1.740 10.026

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.232e+01  2.579e-01  47.782  < 2e-16 ***
t                   1.800e-03  2.036e-04   8.843  < 2e-16 ***
as.factor(n_month)2  4.220e+00  3.346e-01  12.611  < 2e-16 ***
as.factor(n_month)3  9.469e+00  3.269e-01  28.965  < 2e-16 ***
as.factor(n_month)4  1.588e+01  3.299e-01  48.125  < 2e-16 ***
as.factor(n_month)5  1.976e+01  3.276e-01  60.321  < 2e-16 ***
as.factor(n_month)6  2.013e+01  3.529e-01  57.039  < 2e-16 ***
as.factor(n_month)7  1.770e+01  3.529e-01  50.162  < 2e-16 ***
as.factor(n_month)8  1.704e+01  3.529e-01  48.280  < 2e-16 ***
as.factor(n_month)9  1.664e+01  3.564e-01  46.682  < 2e-16 ***
as.factor(n_month)10 1.300e+01  3.533e-01  36.779  < 2e-16 ***
as.factor(n_month)11 6.334e+00  3.571e-01  17.739  < 2e-16 ***
as.factor(n_month)12 1.386e+00  3.542e-01   3.914 9.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.572 on 1237 degrees of freedom
Multiple R-squared:  0.8832,    Adjusted R-squared:  0.8821
F-statistic: 779.6 on 12 and 1237 DF,  p-value: < 2.2e-16
```

Fig 5

2.1.2 Cyclical Trend Model

Based on the periodogram using n = 1250 (Fig 6) and the top 20 periodogram value shown below (Fig 7), we decided to use 3 pairs of sin and cos at harmonics 3, 4, 7 (period 416.67, 312.50, 178.57 accordingly). We also tried to include the harmonics 2, 5, and 6 into our model, which increased the MAPE of the hold-out sample. Therefore we decided to stick with the harmonics 3, 4, and 7 to fit our cyclical model.

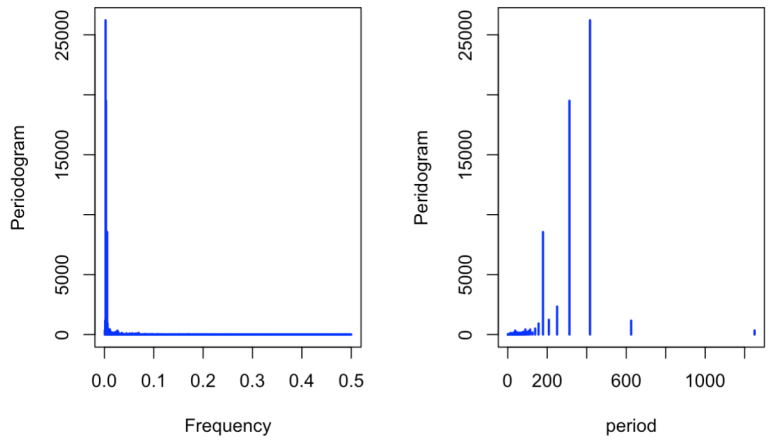


Fig 6

```
          period frequency   amplitude
 [1,] 1250.00000    0.0008   337.02197
 [2,]  625.00000    0.0016  1160.74101
 [3,]  416.66667    0.0024 26218.96157
 [4,]  312.50000    0.0032 19501.82850
 [5,]  250.00000    0.0040  2332.46573
 [6,]  208.33333    0.0048  1233.12650
 [7,]  178.57143    0.0056  8553.02436
 [8,]  156.25000    0.0064   920.03057
 [9,]  138.88889    0.0072   496.02651
[10,]  125.00000    0.0080   170.27766
[11,]  113.63636    0.0088   422.99046
[12,]  104.16667    0.0096   284.17959
[13,]   96.15385    0.0104   184.17943
[14,]   89.28571    0.0112   430.40460
[15,]   83.33333    0.0120   135.15580
[16,]   78.12500    0.0128   235.02102
[17,]   73.52941    0.0136   156.06005
[18,]   69.44444    0.0144   100.53287
[19,]   65.78947    0.0152   152.94860
[20,]   62.50000    0.0160    76.83169
```

Fig 7

Summary of fit for Cyclical Trend shown below (Fig 8): The p-value of F-statistics is less than 0.05 and we can reject the null hypothesis that all ßs are not statistically significant. Also, the coefficient p-values are less than 0.05 showing that all variables in the model are statistically significant.

```
Call:
lm(formula = n_meantemp ~ t + cos3 + sin3 + cos4 + sin4 + cos7 +
    sin7)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3616  -2.1588  -0.0435   2.1350  13.1331

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.8526224  0.2089403 114.160  < 2e-16 ***
t            0.0019323  0.0002937   6.579 6.95e-11 ***
cos3        -2.3202675  0.1407619 -16.484  < 2e-16 ***
sin3         6.0913818  0.1460517  41.707  < 2e-16 ***
cos4         1.7098171  0.1407619  12.147  < 2e-16 ***
sin4        -5.2845759  0.1437612 -36.759  < 2e-16 ***
cos7        -1.2281837  0.1407619  -8.725  < 2e-16 ***
sin7        -3.4706407  0.1417479 -24.485  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.519 on 1242 degrees of freedom
Multiple R-squared:  0.7805,    Adjusted R-squared:  0.7793
F-statistic: 631.1 on 7 and 1242 DF,  p-value: < 2.2e-16

                        ME     RMSE      MAE       MPE     MAPE      MASE
Training set -5.070652e-16 3.50776 2.660362 -2.821244 12.81777 0.4085753
```

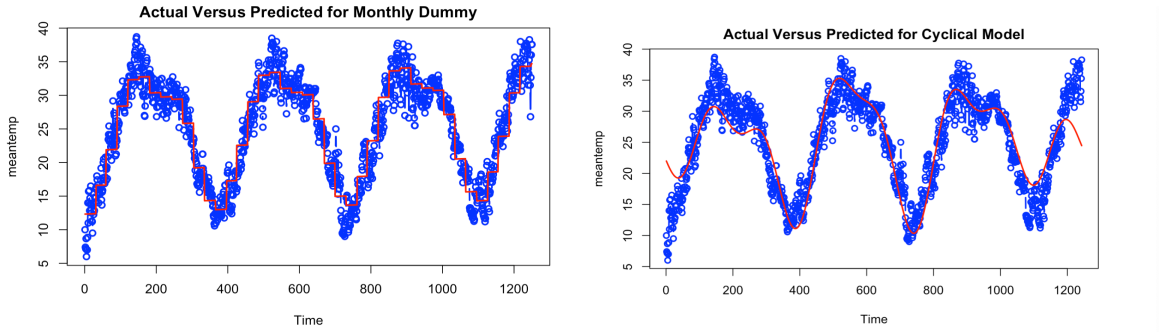
Fig 8

2.2 Comparison of "candidate" models in terms of fit and hold-out sample.

In terms of the fit, the monthly dummy model has an residual standard error of 2.572 and a MAPE of 9.63% on the training set, while the cyclical model has a residual standard error of 3.519 and a MAPE of 12.82% on the training set. Therefore, the monthly dummy model is a better model based on the fit.

In terms of the hold-out sample, the monthly dummy model has a MAPE of 6.41% while the cyclical model has a MAPE of 30.62%. Thus, the monthly dummy model is a better model based on the hold-out sample.

The actual versus predicted plot is shown below (Figs 9&10). Based on the plot, we are guessing that the monthly dummy appears to be a better model based on the MAPE because our cyclical model does not fully capture the upward trend thus further departing from the true value.



Figs 9 & 10

2.3 Looking at residuals of the model(s).

For the trend + monthly dummy model (ACF of the dummy model residuals shown below on the right of Fig 11), the Box-Pierce test yielded a p-value result of 2.2e-16 < 0.05 (Fig 12). We can reject that its residual is a WN, thus our fit is poor.

For the trend + cyclical model (ACF of cyclical model residuals shown below on the left of Fig 11), the Box-Pierce test yielded a p-value result of 2.2e-16 < 0.05 (Fig 12). We again can reject that its residual is a WN, thus our fit is poor.

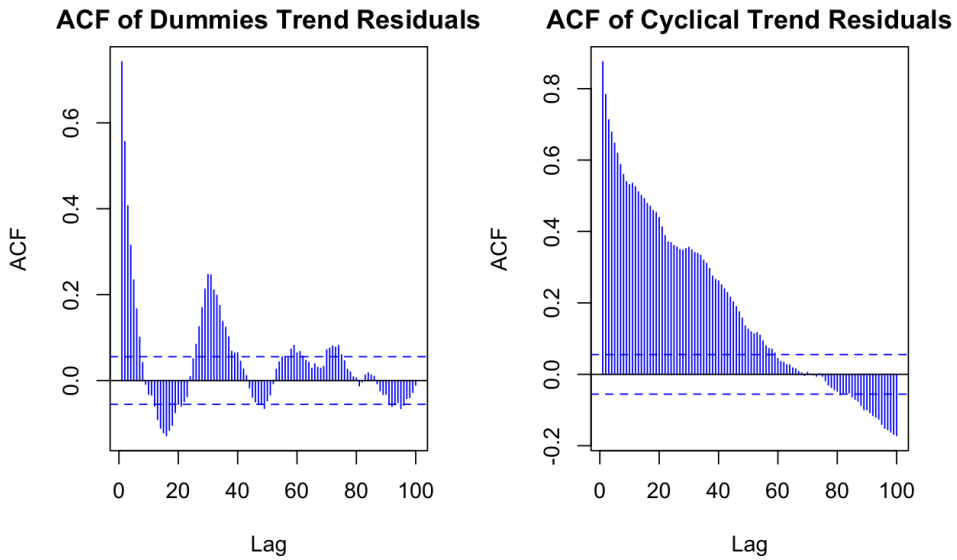


Fig 11

```
        Box-Pierce test

data:  fitDummy$residuals
X-squared = 2299.7, df = 100, p-value < 2.2e-16
```

```
        Box-Pierce test

data:  fitCyclical$residuals
X-squared = 12253, df = 100, p-value < 2.2e-16
```

Fig 12

3. Time Series Regression Models

3.1 Discussion of independent variables. Correlation analysis and scatter plots.

Independent variables: Our first predictor variable is the humidity percentage for that day, which will be represented by a number 0-100. Our second predictor variable is the wind speed in m/s for the day. Our third predictor variable is the mean pressure in hectoPascals (hPa) for the day.

Our scatter plot is shown below (Fig 13). Based on the correlation scatter plot we found that humidity shows a negative relationship with meantemp and wind speed appears to have a positive relationship with meantemp. The meantemp and the mean pressure variables are not very correlated so we decided to exclude it from the regression model. We will do further analysis on mean pressure when we move to section 4.2.

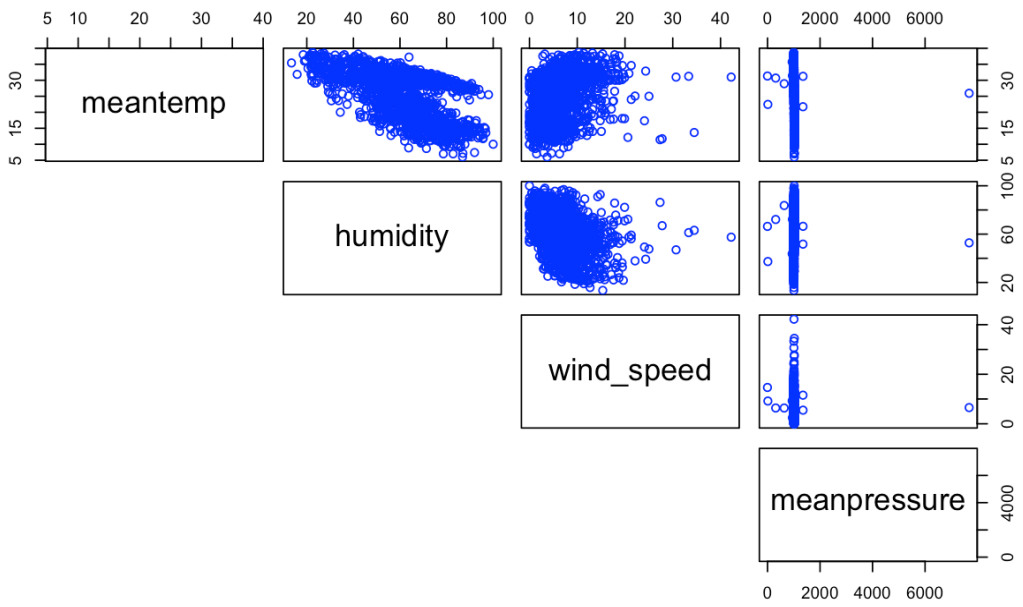


Fig 13

3.2 Comparison of "candidate" models in terms of fit and hold-out sample.

We fit the model using meantemp as our dependent variable and humidity, wind speed and mean pressure as our predictors (Fig 14). Our F-statistic has a very small p-value which leads us to reject the null saying all ßs are not different from 0, which says that at least one or more variables in the model are statistically significant. Looking at the individual p-value for coefficients, all of them are statistically significant. The residual standard error on fit is 5.813 and the MAPE of the training set is 23.76%.

```
Call:
lm(formula = n_meantemp ~ n_humidity + n_wind_speed)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3884  -4.3885  -0.1302   4.7149  13.1409

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.47657    0.77105  51.198  < 2e-16 ***
n_humidity   -0.25569    0.01025 -24.938  < 2e-16 ***
n_wind_speed  0.14842    0.03792   3.914 9.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.813 on 1247 degrees of freedom
Multiple R-squared:  0.3989,    Adjusted R-squared:  0.3979
F-statistic: 413.7 on 2 and 1247 DF,  p-value: < 2.2e-16

[1] "Regression model MAPE for training sample: 23.7632 %"
[1] "Regression model MAPE for hold-out sample: 22.8816 %"
```

Fig 14

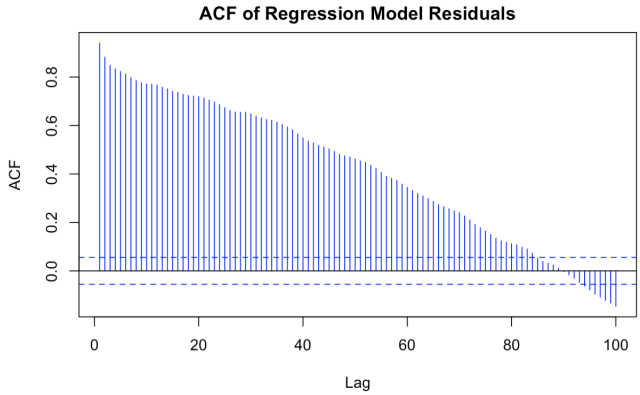
We performed VIF for all our predictors (Fig 15), with all VIFs lower than 10. We can conclude that there is no evidence of multicollinearity. Since they are all lower than 3, we regard them as very good and produce trustworthy p-values.

```
n_humidity n_wind_speed
  1.164419     1.164419
```

Fig 15

3.3 Looking at residuals of the model(s).

The MAPE of the hold-out sample for the regression model is 22.89%. The p-value of the Box-Pierce test for the residuals is 2.2e-16 which is smaller than 0.05 (Box-Pierce test shown below on the right as Fig 17) and the ACF is not stationary given it is decreasing slowly (ACF shown below on the left as Fig 16). We reject the null hypothesis. The residuals are not White Noise and thus our models are not good at this stage.




```
        Box-Pierce test

data:  fitReg$residuals
X-squared = 32348, df = 100, p-value < 2.2e-16
```

Figs 16 & 17

The actual versus predicted plot of the regression model is shown below (Fig 18).

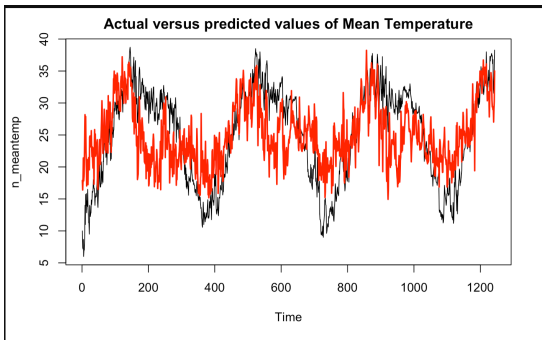


Fig 18

4. Stochastic Time Series Models

4.1 Analysis and modeling of deterministic time series model residuals

4.1.1 Analysis and Modeling of Monthly Model Residuals

```
acf(fitDummy$residuals,col="blue",lag=100)
pacf(fitDummy$residuals,col="blue",lag=100)
```

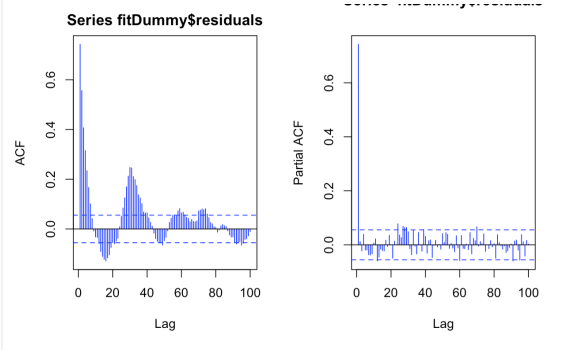


Fig 19

Our monthly dummy model had residuals that were not white noise (shown above as Fig 19). Our ACF plot was decreasing relatively fast and the PACF plot chopped off after lag 1, so we model the error using the AR(1). All coefficients except intercept term and ar1 are not statistically significant. The MAPE of training is 5.50%

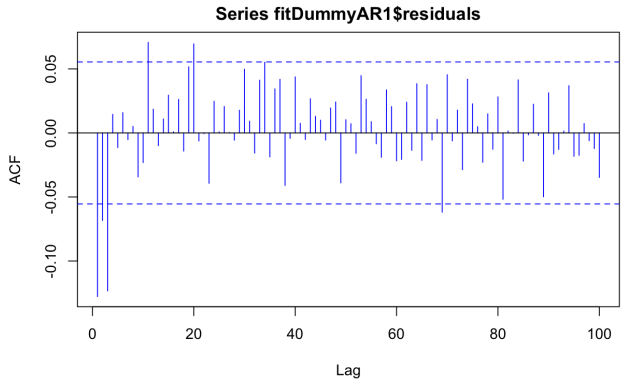
```
Series: n_meantemp
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept       t     n_m2     n_m3    n_m4    n_m5    n_m6    n_m7    n_m8    n_m9   n_m10   n_m11
      0.9732    22.3380  0.0038  -0.1222  -0.9038  0.5934  1.4974  0.3548  0.7355  0.8275  1.3998  0.9973  0.2519
s.e.  0.0077     3.3679  0.0045   0.8172   1.1422  1.3919  1.5796  1.6945  1.7403  1.7072  1.6067  1.4464  1.2296
        n_m12
      -0.1147
s.e.   0.9169

sigma^2 = 2.81:  log likelihood = -2413.95
AIC=4857.9   AICc=4858.28   BIC=4934.86
                      ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 0.01380688 1.667031 1.244375 -0.5392594 5.498817 0.9985652 -0.1273121
```

Figs 20

The p-value of the box pierce test is 0.009348 so we can reject the null hypothesis and say that our model residuals are not white noise (Fig 22). However, we believe that AC values smaller than 0.10-0.15 are not practically significant for modeling purposes (ACF plot of model residual shown below on Fig 21), especially when we have more than a thousand observations (n = 1250 in the training set). Also, the MAPE on the hold-out set is calculated to be 4.82%. Thus, our model is appropriate and a fair fit.




```
        Box-Pierce test

data:  fitDummyAR1$residuals
X-squared = 136.24, df = 100, p-value = 0.009348
```

Figs 21 & 22

4.1.2 Analysis and Modeling of Cyclical Models Residuals

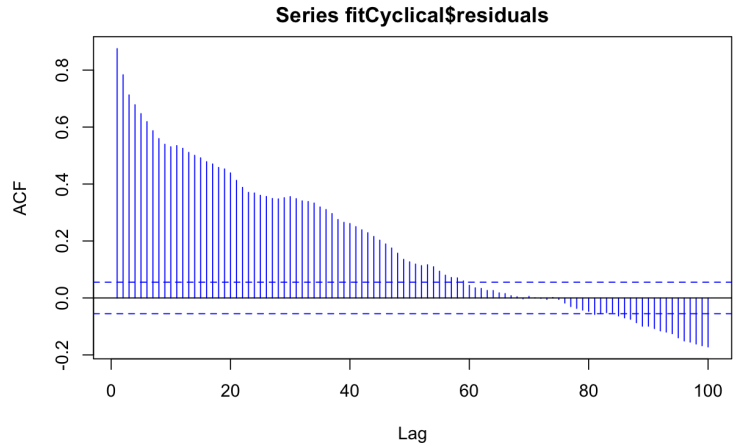


Fig 23

From section 2.3, we know that the cyclical models have residuals that are not white noise and nonstationary, we can see from the plot above that ACF is decreasing slowly up to 100 lags (Fig 23).

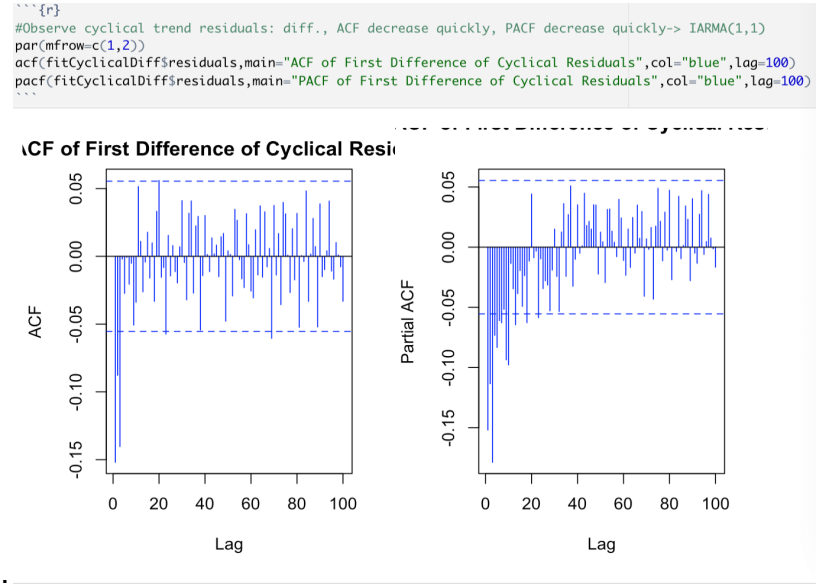


Fig 24

We apply the ARIMA function order = (0,1,0) to take the first difference of the original series to make it stationary (Fig 24). The ACF on residuals of the differenced cyclical model cuts off at around lag 3 and the PACF cuts off at around lag 3. Therefore, we

decided to fit a cyclical trend model with Integrated ARMA(1,1) rather than a high order AR or MA function. Our fit of cyclical terms with IARMA(1,1) in the error term is shown below (Fig 25).

```{r}
#Cyclical trend + IARMA(1,1)
x = cbind(cos3,sin3,cos4,sin4,cos7,sin7)
fitARIMACyclicalDiff = Arima(n_meantemp, order = c(1,1,1), xreg = x, include.constant = T)
fitARIMACyclicalDiff
accuracy(fitARIMACyclicalDiff)
```

```
Series: n_meantemp
Regression with ARIMA(1,1,1) errors

Coefficients:
         ar1      ma1   drift     cos3    sin3    cos4     sin4     cos7     sin7
      0.6607  -0.9207  0.0195  -2.3325  8.4206  1.6974  -3.5376  -1.2403  -2.4717
s.e.  0.0296   0.0145  0.0106   0.9985  1.0176  0.7578   0.7720   0.4561   0.4638

sigma^2 = 2.531:  log likelihood = -2347.86
AIC=4715.73   AICc=4715.9   BIC=4767.03
                      ME     RMSE      MAE        MPE     MAPE      MASE        ACF1
Training set 0.003686319 1.584428 1.191985 -0.4247931 5.257419 0.9565239 0.007639342
```

Fig 25

All coefficients divide by their s.e. accordingly have an absolute value larger than 2 expect drift term, and are thus significant. MAPE of the training set is 5.26%. We do the Box-Pierce test p-value 0.8035 is larger than 0.05 (Fig 26). So we cannot reject the null hypothesis and the residual of cyclical terms with IARMA(1,1) errors is white noise. We also calculate the hold-out sample MAPE, which is 4.66%. We can conclude that cyclical terms with the IARMA(1,1) errors model is a fair fit for this data set.

```
        Box-Pierce test

data:  fitARIMACyclicalDiff$residuals
X-squared = 87.783, df = 100, p-value = 0.8035

                     ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set -0.594521 1.571602 1.170048 -2.545895 4.667895 0.9739894 -0.1716421
```

Fig 26

The ACF of cyclical terms with IARMA(1,1) residuals are shown below and have no ACs practically beyond 2 s.e. (Fig 27).

```{r}
acf(fitARIMACyclicalDiff$residuals,main="ACF of Cyclical Residuals with IARMA(1,1)",col="blue",lag=100)
```

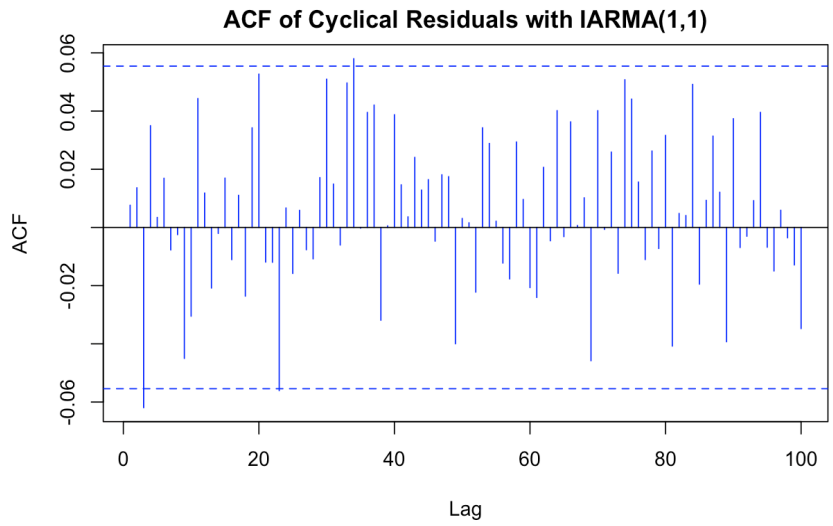


Fig 27

4.2 Analysis and modeling of regression model residuals

From section 3.3, we know that our regression models have residuals that are not white noise and not stationary (refer to Fig 16). We attempt to mitigate this by applying an ARIMA function of order = (0,1,0). This function takes the first difference of the series.

```
Series: n_meantemp
Regression with ARIMA(0,1,0) errors

Coefficients:
       drift  n_humidity  n_wind_speed
      0.0175     -0.1330       -0.0327
s.e.  0.0365      0.0045        0.0075

sigma^2 = 1.668:  log likelihood = -2090.08
AIC=4188.16   AICc=4188.19   BIC=4208.68
```

Fig 28

An analysis of this model shows that all coefficients are significant (Fig 28). However, the ACF plot of model residuals has lags outside of 2 s.e. (Fig 29 on the left). Therefore, the residuals are not white noise. After differencing, our model still needs improvement.

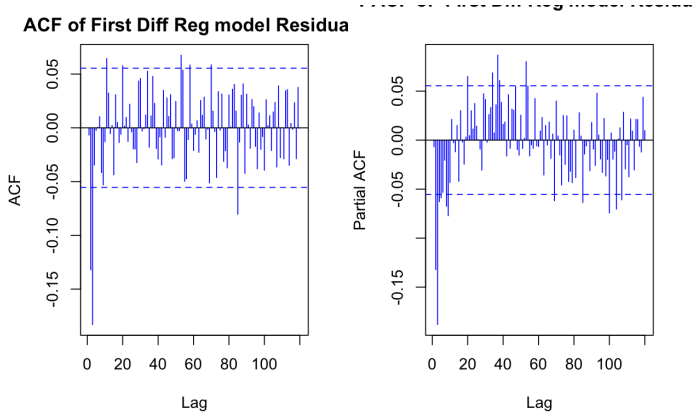


Fig 29

Based on the ACF and PACF of the first differenced regression model shown above (Fig 29), which both decay quickly, we decided to apply an Arima function of order = (2,1,1) after attempts. This applies the autoregressive process of order 2, alongside the first differencing and moving average process of order 1. We tried to include the predictor variable mean pressure which was not significant, so that we remove it from our model.

```
Series: n_meantemp
Regression with ARIMA(2,1,1) errors

Coefficients:
         ar1      ar2      ma1  n_humidity  n_wind_speed
      0.7322  -0.1642  -0.7829     -0.1370       -0.0304
s.e.  0.0464   0.0308   0.0391      0.0046        0.0076

sigma^2 = 1.566:  log likelihood = -2049.8
AIC=4111.6   AICc=4111.67   BIC=4142.38
                      ME     RMSE       MAE        MPE    MAPE      MASE         ACF1
Training set 0.03311908 1.248225 0.9324781 -0.1595318 4.32637 0.7482794 -0.007058125

        Box-Pierce test

data:  fitARIMAReg$residuals
X-squared = 132.73, df = 100, p-value = 0.01588
```

Fig 30

All coefficients are found to be significant in this model and the MAPE of the training set is 4.33% (Fig 30). A Box-Pierce test on this model yields a p-value 0.016 > 0.01 under

1% confidence interval. Looking at the ACF of model residuals (Fig 31), AC values smaller than 0.10 are not practically significant for modeling purposes. From these, we can conclude that our residuals are white noise. MAPE of the hold-out sample is 3.04%. Therefore, the model is a good fit and the residuals are not correlated.

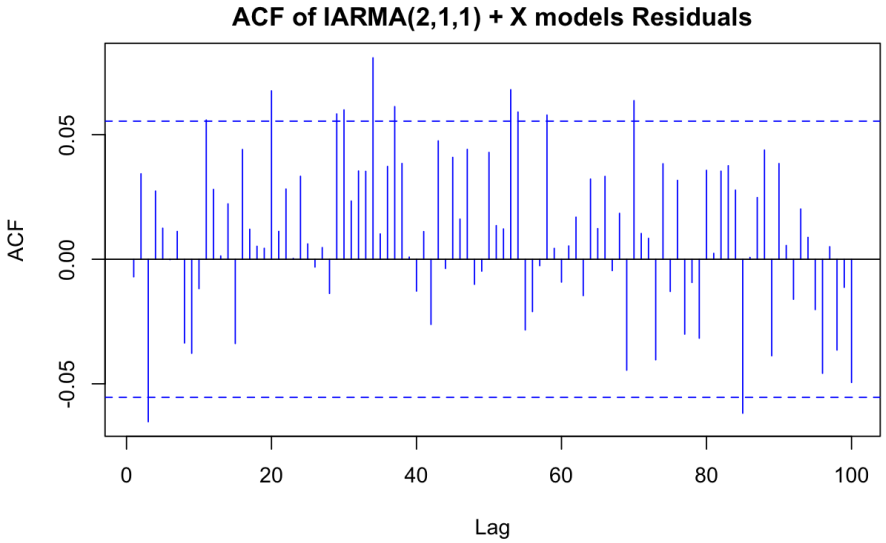


Fig 31

4.3 ARIMA models (for the variable of interest)

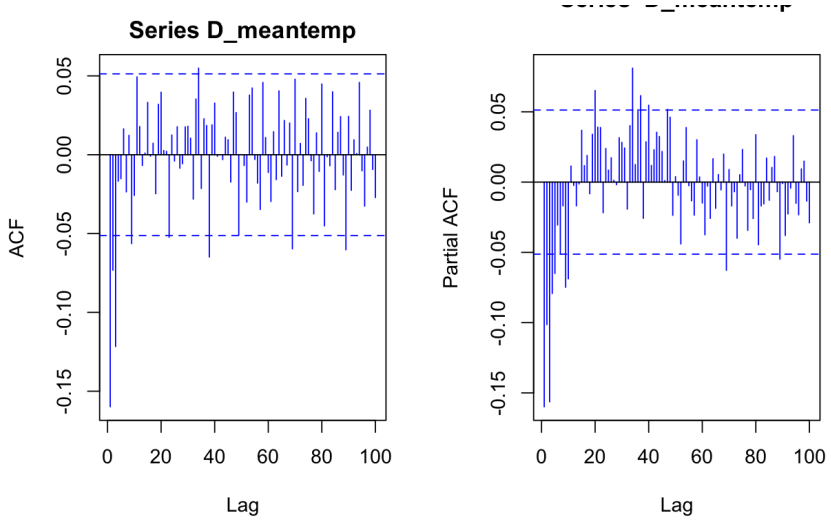


Fig 32

Based on the ACF and PACF of the first difference of series shown above (Fig 32), which both decay quickly, we decided to fit an Integrated ARMA(1,1). The fit of IARMA(1) is shown below (Fig 33). ar1 and ma1 are both significant given |0.5902/0.0454| > 2 and |-0.8075/0.0315| > 2. The RMSE of the model is 1.62 and the MAPE of the training set is 5.43% which shows that the model is a good fit.

```
Series: n_meantemp
ARIMA(1,1,1)

Coefficients:
         ar1      ma1
      0.5902  -0.8075
s.e.  0.0454   0.0315

sigma^2 = 2.644:  log likelihood = -2378.5
AIC=4763   AICc=4763.02   BIC=4778.39
                      ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 0.04482255 1.624026 1.230607 -0.2128901 5.429803 0.9875166 0.01321861
```

Fig 33

Our Box-Pierce test p-value has a value of 0.2524 > 0.05 up to 100 lags, we cannot reject the null hypothesis that the residual is a White Noise. The model residuals are not auto-correlated. Thus, IARMA(1,1) is a good fit with MAPE = 4.53% on the hold-out sample.

```
        Box-Pierce test

data:  fitARIMA$residuals
X-squared = 109.03, df = 100, p-value = 0.2524

                      ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set -0.2468188 1.519041 1.142349 -1.369414 4.532449 0.9509314 -0.1501772
```

Fig 34

The ACF of model residuals are shown below (Fig 35). It shows that no ACs are practically beyond 2 s.e., which reinforces that the residual is white noise and IARMA(1,1) is a good fit for our data set.

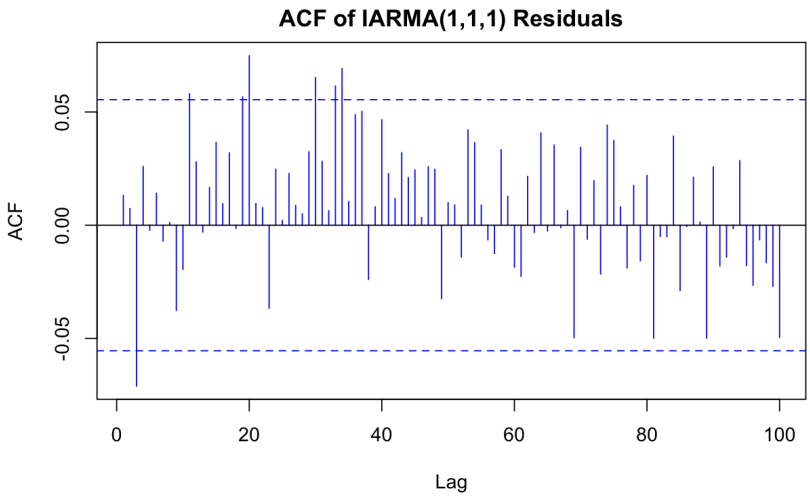


Fig 35

5. Conclusion Summary of Findings and Comparison of Deterministic and Stochastic Models Performance

Model	Fit / training	Hold-out (MAPE)	Residuals
Seasonal dummy + trend	RSE = 2.572 MAPE = 9.63%	6.41%	Not White Noise
Cyclical + trend	RSE = 3.519 MAPE = 12.82%	30.62%	Not WN
Regression	RSE = 5.813 MAPE = 23.76%	22.88%	Not WN
Seasonal dummy + AR(1)	MAPE = 5.50%	4.76%	Practically WN
Cyclical + IARMA(1,1)	MAPE = 5.26%	4.67%	WN under 0.05% CI
Regression + IARMA(2,1)	MAPE = 4.33%	3.04%	WN under 0.01% CI
IARMA(1,1)	MAPE = 5.43%	4.53%	WN under 0.05% CI

Fig 36

We fitted models to see what effects that time has on Delhi's weather trends. By this data we were able to fit dummy, cyclical, regression, ARIMA models and improve them

based on our observations and tests on the model residuals. Through the process, we see that not only our predictor variables, but also time have an effect on our target variable mean temperature. After running multiple different models (summary table shown above on Fig 36), we found that stochastic models in general have better performance compared to deterministic models in terms of fit, hold-out, and whether residuals are white noise. We believe that it is because the data we used is the daily data on Delhi's weather, which is not stationary, and stochastic models allow us to model it via taking the first difference. Also, stochastic models allow us to make improvements building on deterministic models by recapturing the autocorrelations on the residuals.

To make a final conclusion, the Regression with IARMA(2,1) error model is the best model, based on the comparison of their MAPEs on the hold-out sample. It has the lowest value at 3.04% and its residuals are white noise. For these reasons, it is the best among all models to predict Delhi's daily weather.