DNSC 6219 Time Series Final Report

Time Series Analysis of Delhi Climate

Team Member: Yihang Zhao, Qian Xie, Chengkai Ji, Leqi Yin

1. Introduction and Overview

Weather is important but hard to predict. The complexity of the weather gives us more predictability to this intriguing system. Many variables within the Earth's atmosphere, for example temperature, pressure, wind velocity, humidity, and precipitation, are not linear but interacting which aroused our interest in in-depth research and prediction.

We find the daily climate time series dataset on Kaggle and decide to use it as the main data for the project. It contains the general climate information from 1st January 2013 to 24th April 2017 for the city Delhi, India provided by the Weather Underground API. We will use this data to forecast the weather and gain further understanding on the time series models.

The data contains 1462 observations, including 4 variables that are the basic attributes of the climate, which are mean temperature, humidity, wind speed, and mean pressure. Our goal is to test the dataset through different models we learned in class and compare the models to find the best one in predicting the mean temperature for the city Delhi.

Figure1.1 Time Series Plot view

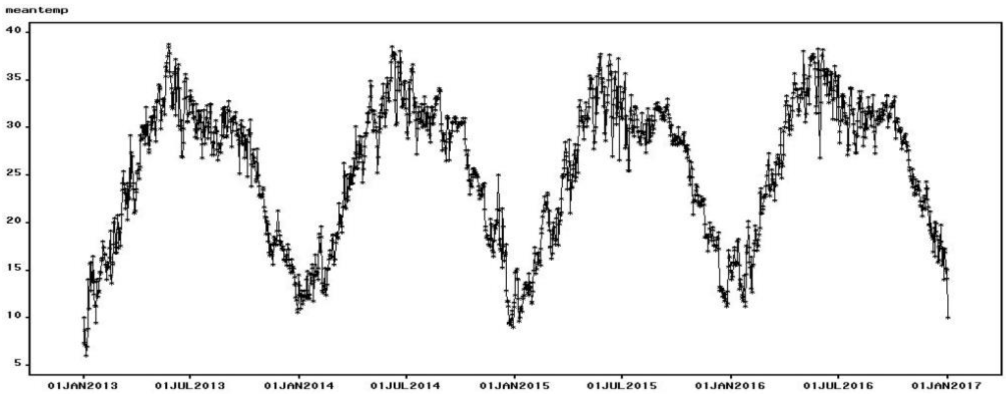


Figure 1.2.1 Box Plot View of the Mean Temperature in Year 2013

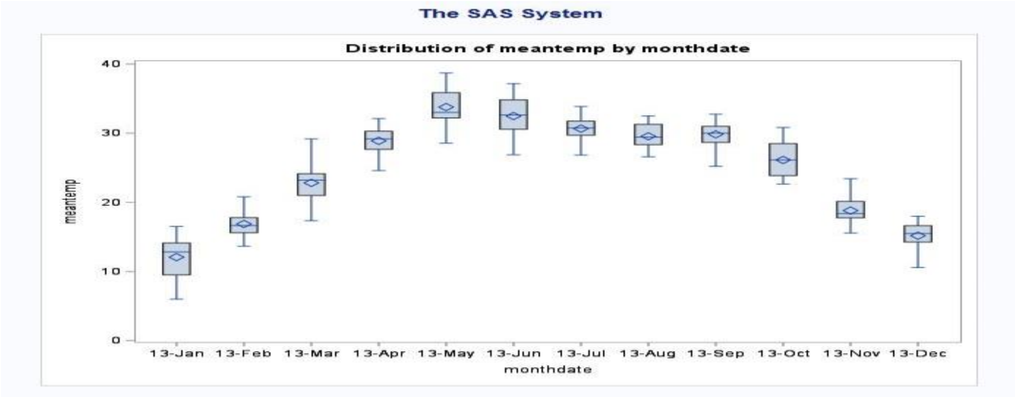
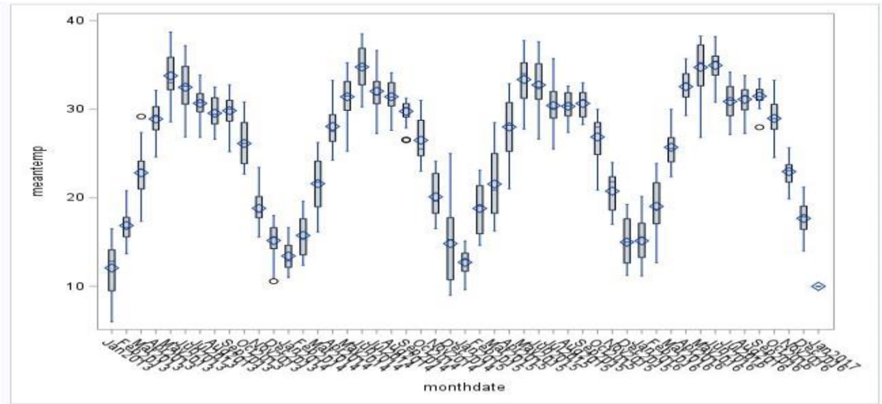


Figure 1.2.2 Box Plot View of the Mean Temperature in Year 2013-2017



Based on the Time Series Plot and the box plot above, there is a strong seasonality and cyclical behavior of the climate in Delhi. The seasonality for this dataset is that the mean temperature trend moves upward from January to July every year and goes downward from July to December. Also, The trend of the mean temperature repeats in regular intervals every year. Therefore, we can conclude cyclical behavior is an annual circle. Furthermore, we are using 200 hold-out samples in our model fitting procedures.

2. Univariate Time-series models

2.1 Deterministic Time Series Models and Error model

- Seasonal Dummies and Trend Model:

Figure 2.1.1 Fit of model

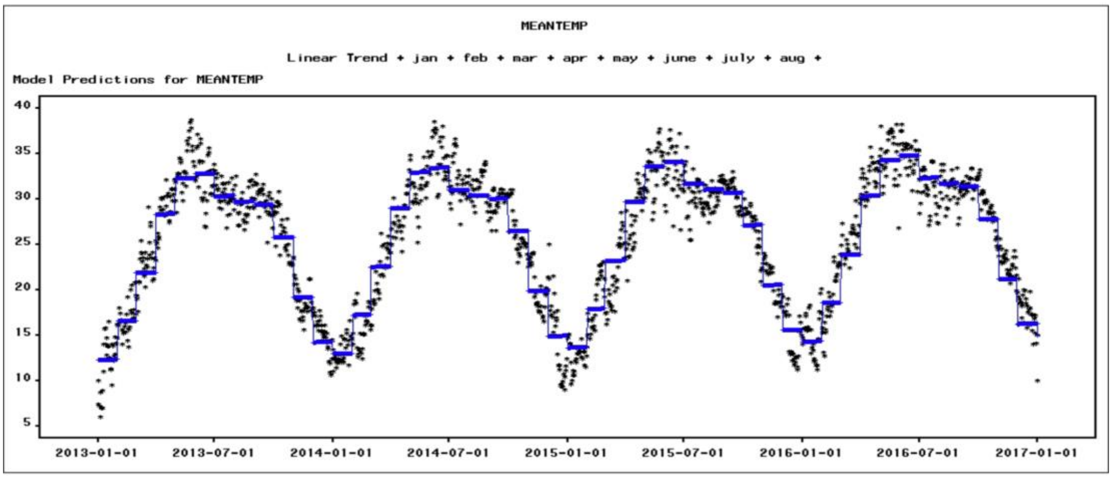


Figure 2.1.2 Parameters of seasonal dummies and trend model

Parameter Estimates

MEANTEMP

Linear Trend + jan + feb + mar + apr + may + june + july +

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	13.63358	0.3026	45.0515	<.0001
Linear Trend	0.00183	0.000199	9.1835	<.0001
jan	-1.32684	0.3536	-3.7526	0.0002
feb	2.89183	0.3603	8.0268	<.0001
mar	8.14014	0.3528	23.0753	<.0001
apr	14.54680	0.3550	40.9782	<.0001
may	18.43022	0.3523	52.3084	<.0001
june	18.87487	0.3660	51.5769	<.0001
july	16.37305	0.3777	43.3498	<.0001
aug	15.71069	0.3772	41.6456	<.0001
sep	15.30903	0.3800	40.2862	<.0001
oct	11.66579	0.3767	30.9719	<.0001
nov	5.00387	0.3786	13.2182	<.0001
Model Variance (sigma squared)	6.55445	.	.	.

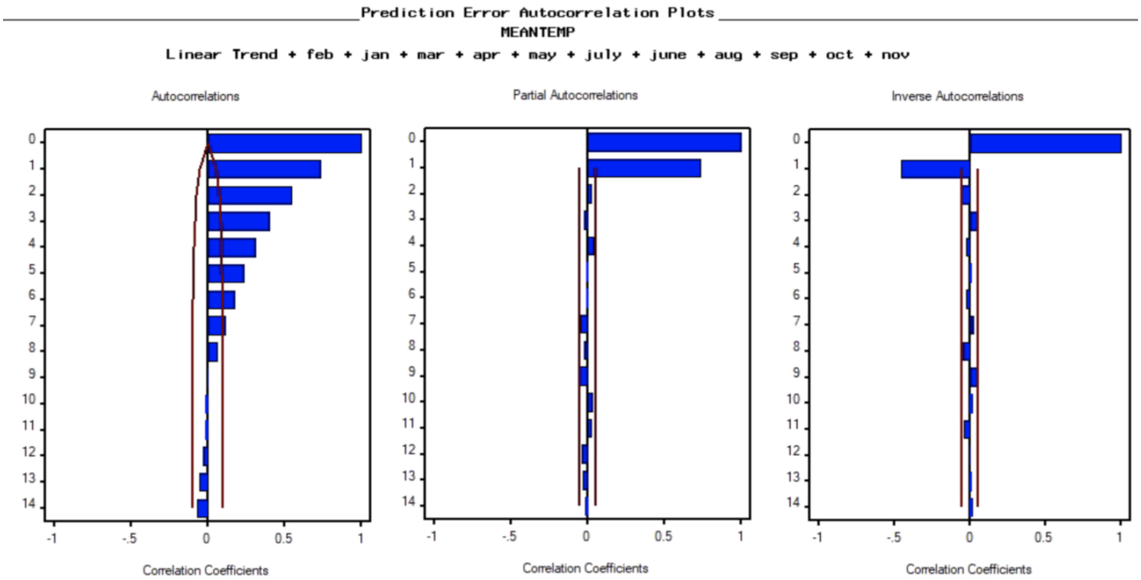
Figure 2.1.3 Statistic of fit

Statistic of Fit	Value
Mean Square Error	4.45313
Root Mean Square Error	2.11024
Mean Absolute Percent Error	6.59204
Mean Absolute Error	1.66299
R-Square	0.862

From the fit of the model(Figure 2.1.1), we can see that the model follows the pattern of the data, and the MAE and MAPE(Figure 2.1.3) of the model are 1.66 and 6.59, which is not too bad. The t-statistic of parameters of the model(Figure 2.1.2) are all less than 0.05, which indicates that all seasonal dummies and trend parameters are significant.

- Error Model of Seasonal Dummies and Trend Model

Figure 2.1.4 ACF of residual



According to figure 2.1.4, the ACF of residual decays exponentially and PACF drops to 0 after lag 1, so we believe AR(1) is a good model to model residuals.

Figure 2.1.5 ACF of residual of error model AR(1)

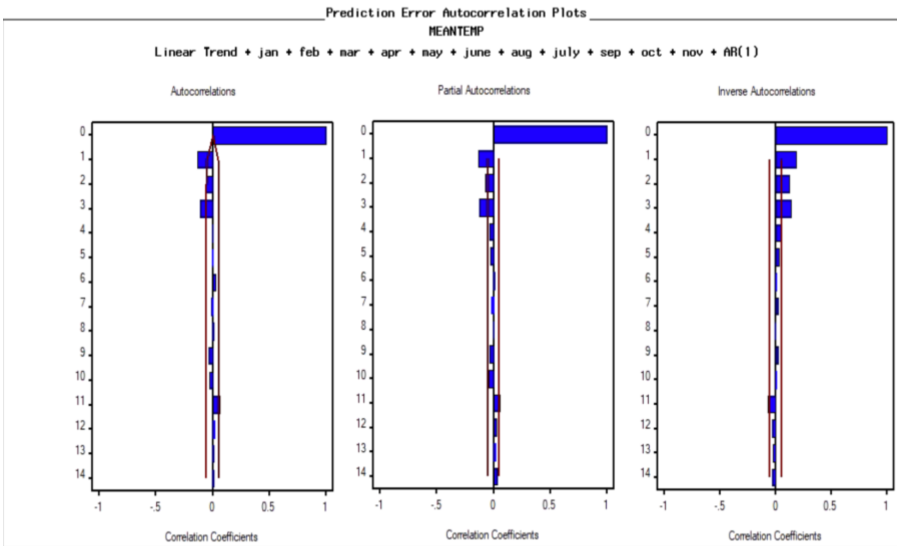


Figure 2.1.5 shows that the ACF drops to 0 after lag 3 and IACF decays, so we can fit an MA(3) to model the residuals of the error model.

Figure 2.1.6 ACF and white noise test of residual of error model ARMA(1,3)

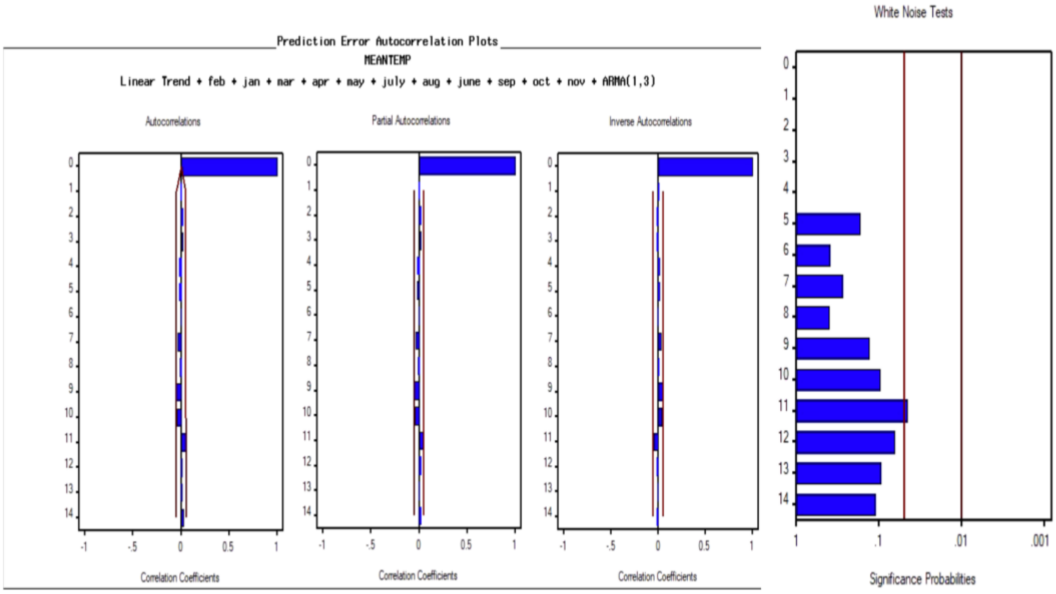


Figure 2.1.6 shows ACF, PACF and IACF drops to 0 after lag 0 and the white noise test indicates that the residuals are white noise, therefore we can s top fitting error at this point.

Figure 2.1.7 Fit of error model ARMA(1,3)

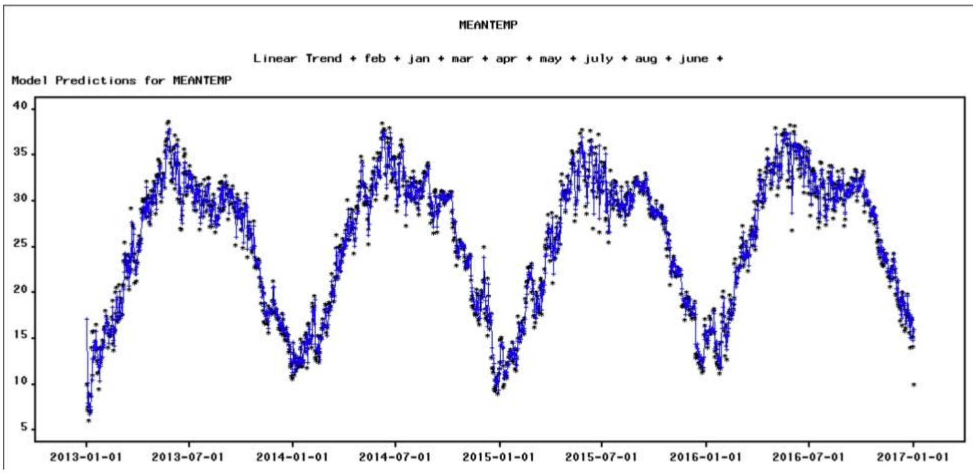


Figure 2.1.8 Parameters of ARMA(1,3)

Parameter Estimates

MEANTEMP

Linear Trend + feb + jan + mar + apr + may + july + aug + june + sep + oct + nov + ARMA(1,3)

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	16.75776	5.9164	2.8324	0.0051
Moving Average, Lag 1	0.20695	0.0284	7.2924	<.0001
Moving Average, Lag 2	0.13217	0.0289	4.5720	<.0001
Moving Average, Lag 3	0.15887	0.0284	5.5971	<.0001
Autoregressive, Lag 1	0.99405	0.0033	305.1713	<.0001
Linear Trend	0.01084	0.0076	1.4298	0.1545
feb	0.60433	1.0891	0.5549	0.5797
jan	0.38639	0.8488	0.4552	0.6495
mar	-0.09294	1.2433	-0.0748	0.9405
apr	1.26704	1.3383	0.9467	0.3450
may	2.04806	1.4005	1.4624	0.1453
july	1.27960	1.4338	0.8925	0.3733
aug	1.21324	1.3930	0.8710	0.3849
june	1.24220	1.4285	0.8696	0.3857
sep	1.54353	1.2952	1.1917	0.2349
oct	1.16471	1.1275	1.0330	0.3030
nov	0.27906	0.8399	0.3322	0.7401
Model Variance (sigma squared)	2.63587	.	.	.

Figure 2.1.9 Statistic of fit of ARMA(1,3)

Statistic of Fit	Value
Mean Square Error	2.38520
Root Mean Square Error	1.54441
Mean Absolute Percent Error	4.69438
Mean Absolute Error	1.15630
R-Square	0.926

The MAE and MAPE of error model(Figure 2.1.9) are 1.15 and 4.69, which is lower than simple seasonal dummies and trend model at 1.66 and 6.59. On the other hand, the t-statistics of parameters of error model(Figure 2.1.8) show that only the parameters of ARMA are significant. Seasonal dummies and trend are no longer showing significance. Therefore, the performance may be better to use a simple ARMA or ARIMA model.

- Cyclical Model

  Hold-out sample 200

Figure 2.1.10 Periodogram

Obs	FREQ	PERIOD	P_01
1	0.00000	.	0.00
2	0.00498	1262.00	521.52
3	0.00996	631.00	1252.13
4	0.01494	420.67	23331.99
5	0.01992	315.50	22988.16
6	0.02489	252.40	2714.14
7	0.02987	210.33	1597.86
8	0.03485	180.29	8446.42
9	0.03983	157.75	606.64
10	0.04481	140.22	433.84
11	0.04979	126.20	263.08
12	0.05477	114.73	392.55
13	0.05975	105.17	166.04
14	0.06472	97.08	211.94
15	0.06970	90.14	458.67

Figure 2.1.10 shows that the 10 harmonics with highest amplitude are: 1,2,3,4,5,6,7,8,9,14, so we created 20 sin and cos terms.

Figure 2.1.11 Fit of cyclical model

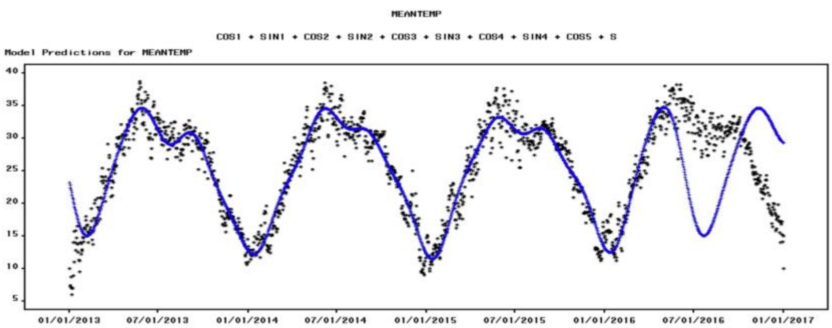


Figure 2.1.12 Parameters of cyclical model

Parameter Estimates

MEANTEMP

1 + SIN1 + COS2 + SIN2 + COS3 + SIN3 + COS4 + SIN4 + COS5 + SIN5 + COS6 + SIN6 + COS7 +

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	25.15950	0.0761	330.4785	<.0001
COS1	0.21017	0.1077	1.9520	0.0525
SIN1	0.07513	0.1077	0.6978	0.4862
COS2	0.12204	0.1077	1.1335	0.2585
SIN2	0.99863	0.1077	9.2753	<.0001
COS3	-1.54655	0.1077	-14.3645	<.0001
SIN3	5.61034	0.1077	52.1094	<.0001
COS4	1.27141	0.1077	11.8089	<.0001
SIN4	-6.10334	0.1077	-56.6883	<.0001
COS5	0.25435	0.1077	2.3624	0.0192
SIN5	-2.22053	0.1077	-20.6245	<.0001
COS6	0.11535	0.1077	1.0714	0.2854
SIN6	-1.72224	0.1077	-15.9964	<.0001
COS7	-1.78539	0.1077	-16.5829	<.0001
SIN7	-3.30801	0.1077	-30.7251	<.0001
COS8	-0.34346	0.1077	-3.1901	0.0017
SIN8	-1.01883	0.1077	-9.4629	<.0001
COS9	0.12370	0.1077	1.1489	0.2521
SIN9	-0.91016	0.1077	-8.4537	<.0001
COS14	0.05514	0.1077	0.5121	0.6092
SIN14	-0.90874	0.1077	-8.4405	<.0001
Model Variance (sigma squared)	7.31437	.	.	.

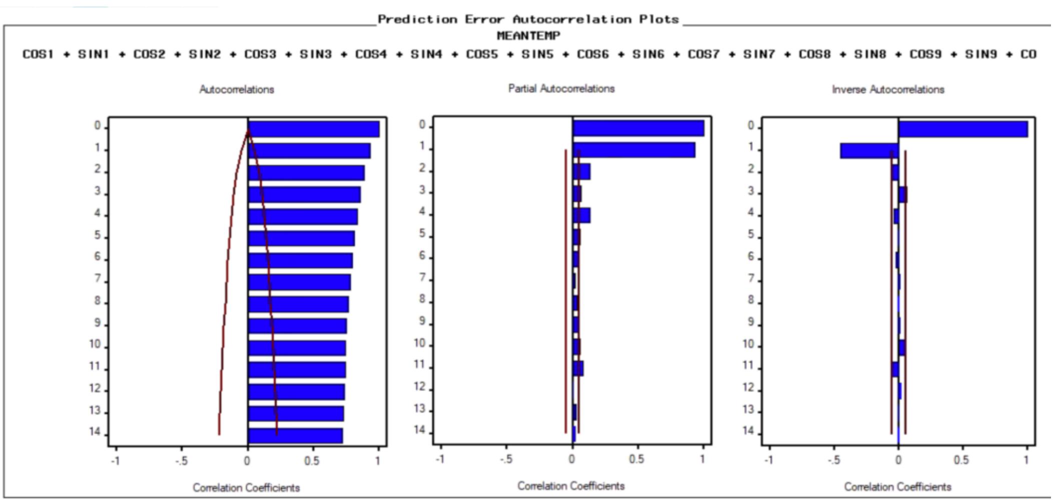
Figure 2.1.13 Statistic of cyclical model

Statistic of Fit	Value
Mean Square Error	132.45250
Root Mean Square Error	11.50880
Mean Absolute Percent Error	41.71666
Mean Absolute Error	10.53840
R-Square	-3.096

From the fit of the cyclical model(Figure 2.1.11) and the statistic of fit (Figure 2.1.13), we can see that the cyclical model fits well on existing data but doesn't perform well on prediction. Most of the parameters of the model(Figure 2.1.12) are significant except for cos1, sin1, cos2, cos6 and cos14.

- Error Model of Cyclical Model

Figure 2.1.14 ACF of residual of cyclical model



From figure 2.1.14, we can see that the ACF decays slowly, PACF drops to 0 after lag 4 and IACF drops quickly. So we decided to fit an AR(4) error model.

Figure 2.1.15 Fit of AR(4) error model

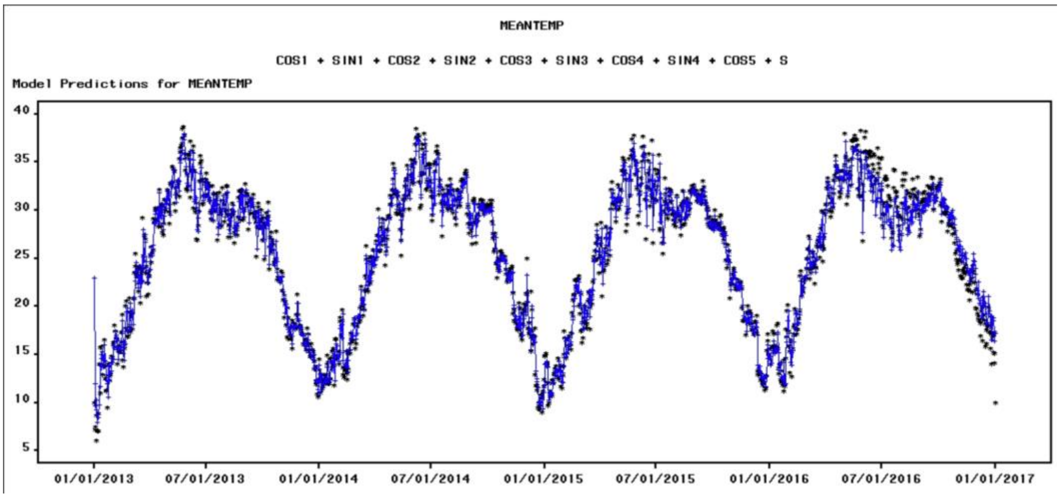


Figure 2.1.16 Parameters of error model

Parameter Estimates

MEANTEMP

SIN1 + COS2 + SIN2 + COS3 + SIN3 + COS4 + SIN4 + COS5 + SIN5 + COS6 + SIN6 + COS7 + S

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
Intercept	25.14582	0.3193	78.7563	<.0001
Autoregressive, Lag 1	0.76315	0.0283	26.9591	<.0001
Autoregressive, Lag 2	0.03442	0.0357	0.9645	0.3361
Autoregressive, Lag 3	-0.04986	0.0357	-1.3951	0.1647
Autoregressive, Lag 4	0.11115	0.0286	3.8863	0.0001
COS1	0.18283	0.4490	0.4072	0.6844
SIN1	0.08889	0.4532	0.1962	0.8447
COS2	0.09477	0.4477	0.2117	0.8326
SIN2	1.02598	0.4519	2.2706	0.0244
COS3	-1.57372	0.4456	-3.5320	0.0005
SIN3	5.65097	0.4497	12.5663	<.0001
COS4	1.24439	0.4426	2.8115	0.0055
SIN4	-6.04990	0.4467	-13.5430	<.0001
COS5	0.22751	0.4389	0.5184	0.6048
SIN5	-2.15488	0.4430	-4.8645	<.0001
COS6	0.08873	0.4345	0.2042	0.8384
SIN6	-1.64508	0.4385	-3.7514	0.0002
COS7	-1.81177	0.4294	-4.2191	<.0001
SIN7	-3.22012	0.4334	-7.4292	<.0001
COS8	-0.36957	0.4238	-0.8720	0.3844
SIN8	-0.92105	0.4278	-2.1531	0.0327
COS9	0.09787	0.4177	0.2343	0.8150
SIN9	-0.80339	0.4216	-1.9054	0.0584
COS14	0.03088	0.3827	0.0807	0.9358
SIN14	-0.77010	0.3860	-1.9951	0.0476
Model Variance (sigma squared)	2.58862	.	.	.

Figure 2.1.17 ACF of residual of error model

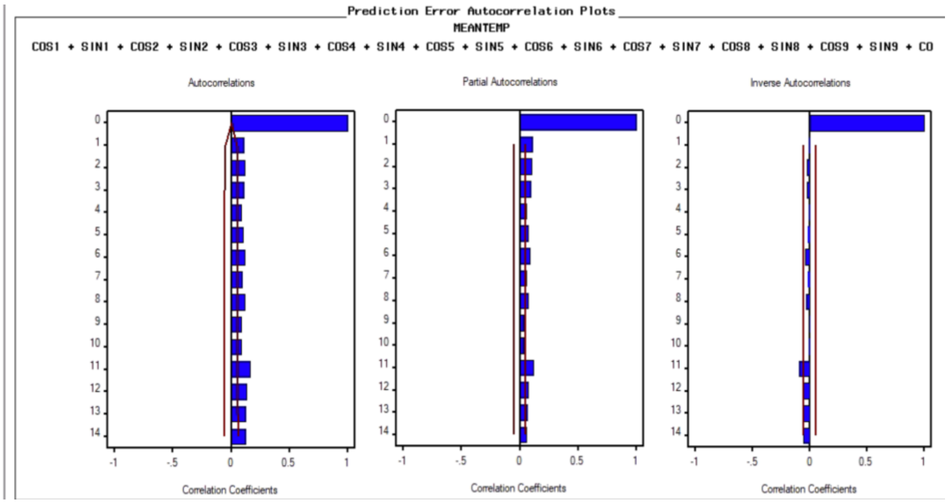


Figure 2.1.18 Statistic of fit of error model

Statistics of Fit

MEANTEMP

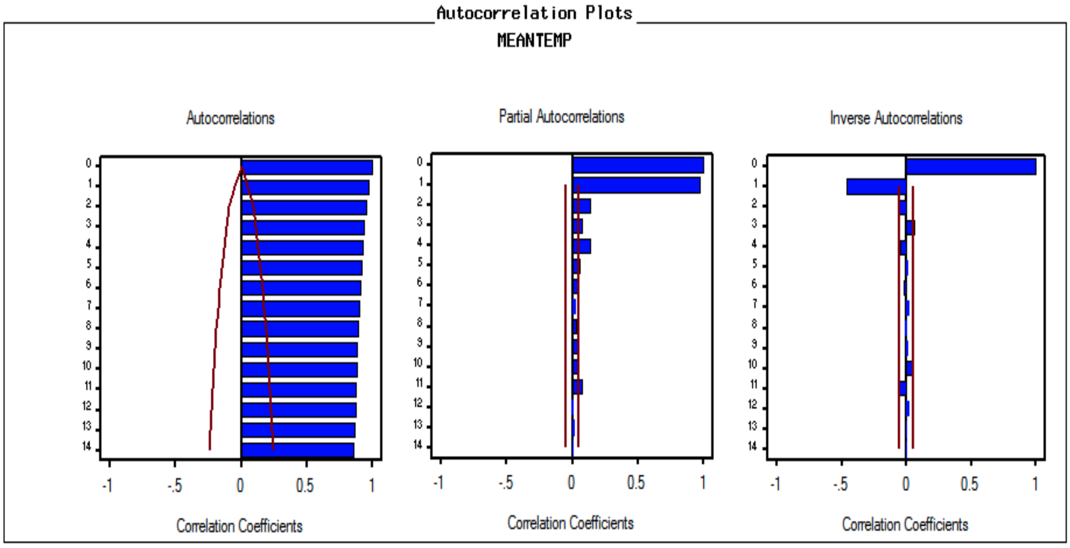
COS2 + SIN2 + COS3 + SIN3 + COS4 + SIN4 + COS5 + SIN5 + COS6 + SIN6 +

Statistic of Fit	Value
Mean Square Error	4.91133
Root Mean Square Error	2.21615
Mean Absolute Percent Error	7.22072
Mean Absolute Error	1.79893
R-Square	0.848

From the fit of the model(Figure 2.1.15) and statistics of fit(Figure 2.1.18), we can see that the error model has a much better performance than the original model. Only a few parameters remain significant in the model(Figure 2.1.16). The ACF, PACF and IACF of residuals of the error model becomes very small after lag0, so we assume they're white noise.

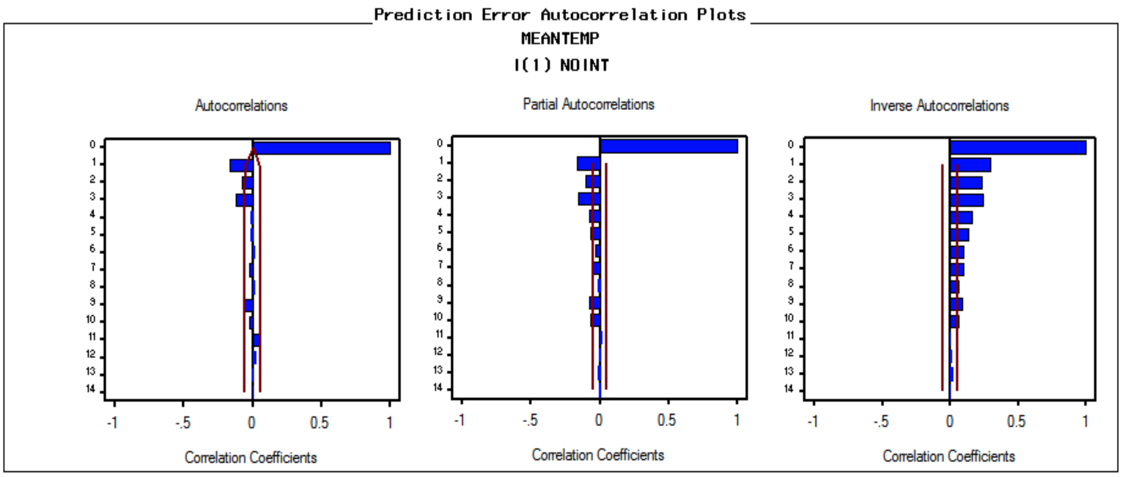
2.2 ARIMA models

Figure 2.2.1 (ACF, PACF and Inverse Autocorrelation)



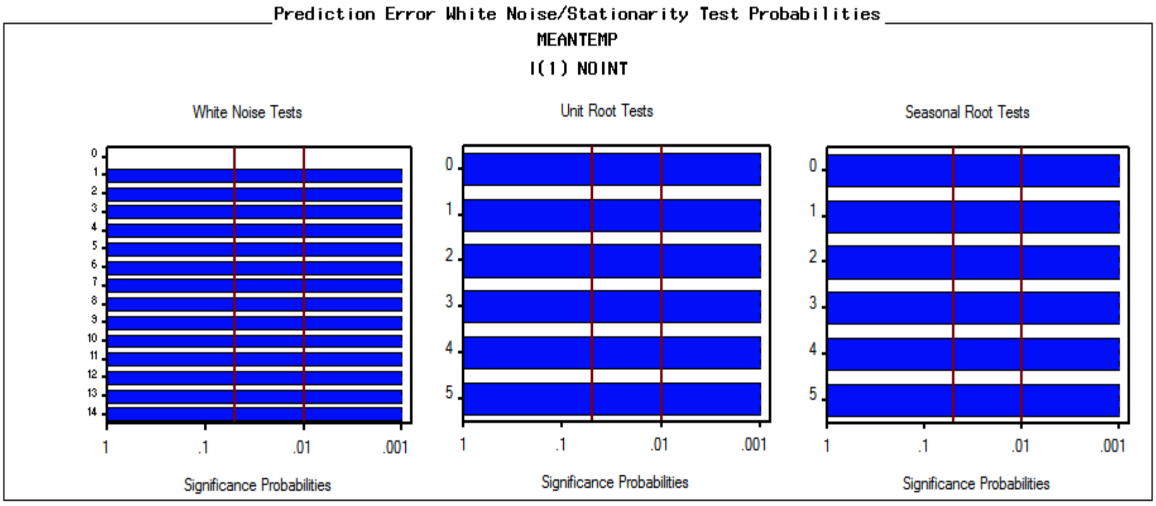
After graphing the series, we can see the obvious pattern of seasonality within the series. Based on autocorrelation plots, we found that ACF decays slowly which indicates that the series is nonstationary. Therefore, we fit the differencing model first.

Figure 2.2.2 (ACF, PACF and Inverse Autocorrelation using non-seasonal differencing)



From figure (2.2.2), both ACF and PACF decays quickly, the first difference makes the series stationary. There is no seasonal pattern after the first difference, so we don't need to include seasonal ARIMA components.

Figure 2.2.3 (White Noise Test and Unit Root Test using non-seasonal differencing)



From figure (2.2.3), the white noise test indicates all lags have values less than 0.05, we can reject the null and conclude with reasonable confidence that the series is not White Noise. Also, the unit root test indicates the values are less than 0.05, we can reject the null and conclude with reasonable confidence that this series is stationary (it doesn' t have a unit root).

Figure 2.2.4 (Fitting ARIMA model)

Forecast Model	Model Title	Mean Absolute Percent Error
☐	ARIMA(1,1,1) NOINT	4.58286
☐	ARIMA(2,1,1) NOINT	4.59388
☐	ARIMA(2,1,2) NOINT	4.43923
☐	ARIMA(1,1,2)	4.61855
☐	ARIMA(1,1,3) NOINT	4.58465
☑	ARIMA(2,1,3) NOINT	4.60311

Based on the Mean Absolute Percent Error on figure (2.2.4), we found these models have similar predictive performance since their Mean Absolute Percent Error are close to each other and all around 4.5. We pick ARIMA(2,1,2) which has the lowest error to look into other information.

Figure 2.2.5 ARIMA(2, 1, 2) Error

Statistics of Fit
MEANTEMP
ARIMA(2,1,2) NOINT

Statistic of Fit	Value
Mean Square Error	2.22715
Root Mean Square Error	1.49236
Mean Absolute Percent Error	4.43923
Mean Absolute Error	1.10930
R-Square	0.931

Figure 2.2.6 ARIMA(2, 1, 2) parameters

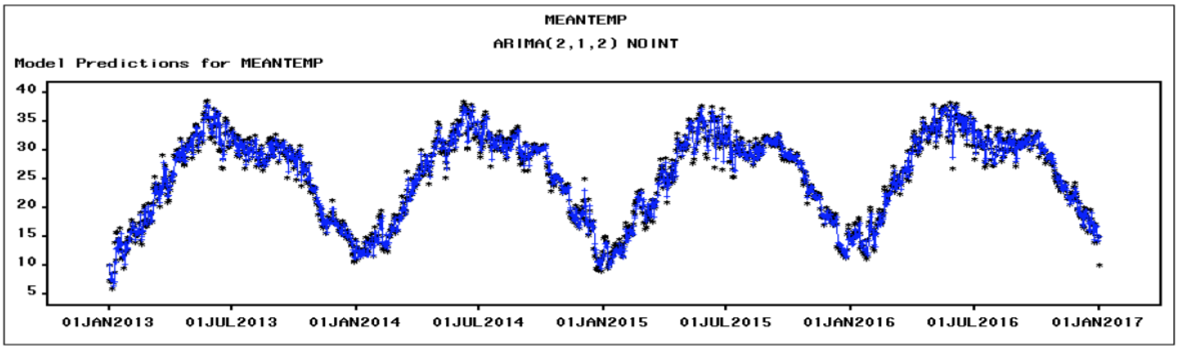
Parameter Estimates
MEANTEMP
ARIMA(2,1,2) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	1.92756	0.0156	123.5798	<.0001
Moving Average, Lag 2	-0.93340	0.0156	-59.7554	<.0001
Autoregressive, Lag 1	1.70951	0.0301	56.8728	<.0001
Autoregressive, Lag 2	-0.71526	0.0301	-23.7740	<.0001
Model Variance (sigma squared)	2.58755	.	.	.

Fit Range: 01JAN2013 to 15JUN2016

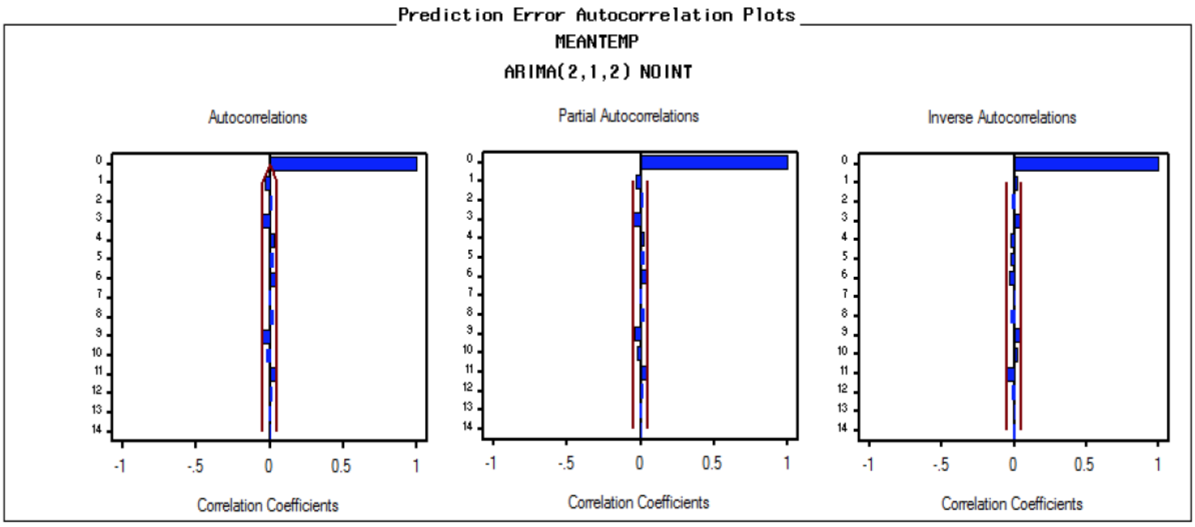
Based on the figure (2.2.6), we can find that both MA and AR coefficients at lag 1 and lag 2 have p-values less than 0.05, which indicate they are significant. Also, because the intercept is not significant, we delete the intercept.

Figure 2.2.7 (Actual vs. predicted values of the series) --- ARIMA(2,1,2)



Looking at the Figure (2.2.7) actual vs. predicted graph, we can see ARIMA (2,1,2) model has a good performance in fitting the actual values and the moving trend.

Figure 2.2.8 (ACF, PACF and Inverse Autocorrelation) --- ARIMA(2,1,2)



Based on figure 2.2.8, after fitting the ARIMA(2,1,2) model, the series becomes stationary. Both ACF and PACF decays quickly and within the bound after lag 0.

2.3 Comparison of models

Comparing the models above, the mean absolute percent error of Cyclical Model with AR(4) and seasonal dummies and trend model with MA(3) is 7.22 and 6.59. It is higher than the ARIMA model which has the mean absolute percent error of 4.44. Although the difference between the values is not very significant which means they all did a good job in predicting the values, the ARIMA model still has better performance than the other. Thus, the ARIMA(2,1,2) can predict the data more accurately than seasonal dummies and trend models with MA(3).

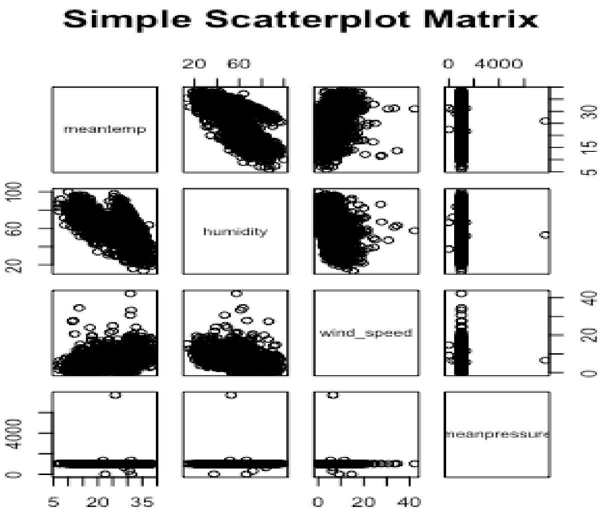
Model	Root Mean Square Error	Mean Absolute Percent Error
Seasonal Dummies and trend	2.11	6.59
Seasonal Dummies and trend + ARMA(1,3)	1.54	4.69
ARIMA(2,1,2)	1.49	4.44
Cyclical Model	11.50	41.71
Cyclical Model with AR(4)	2.21	7.22

3. Multivariate Time Series Models

3.1 Regression model and analysis of regression residuals

Because we use 3 variables: humidity, wind_speed and mean pressure to predict the mean temperature in Delhi, we decided to fit the model with these variables as predictors. Like the last draft, we choose the hold-out sample as 200 observations.

Figure 3.1.1 (Scatter Plot of correlations between variables)



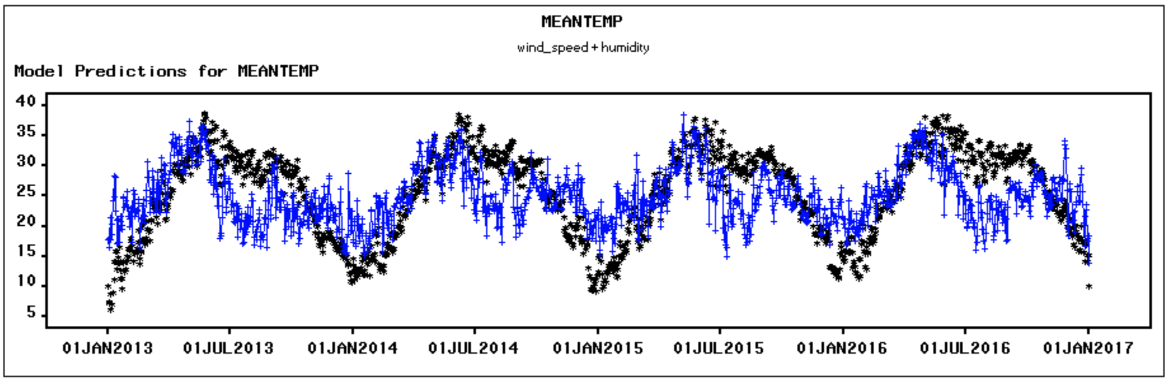
Based on the scatterplot of dependent variable ‘meantemp’ and other predictors, we can see that predictor ‘meanpressure’ doesn’t have much correlation with the ‘meantime’. later, we look into more information about the regression outputs.

Figure 3.1.2 (Error of Regression model)

Forecast Model	Model Title	Mean Absolute Percent Error
☐	wind_speed + humidity	23.14179
☑	wind_speed + humidity + meanpressure	23.09708

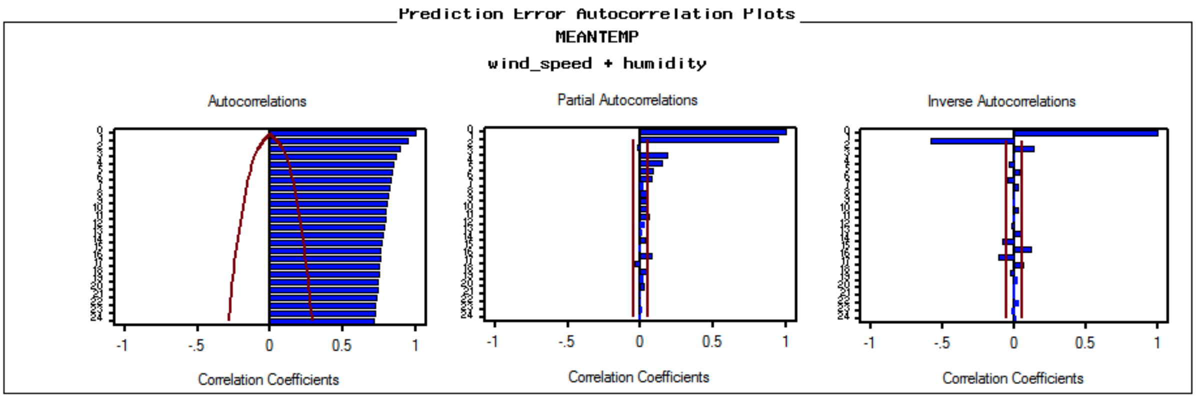
According to Figure 3.1.2, although ‘meanpressure’ is not a great predictor, the two models above have very similar predictive performance and Mean Absolute Percent Error. Then, we will look at predictive vs. actual values graphs.

Figure 3.1.3 (Predicted values vs. actual values of regression model)



Based on the figure (3.1.3), we can see that the predicted values have a very similar trend compared to the actual values. But there are some errors between these two values.

Figure 3.1.4 (ACF and PACF of regression model)



From the figure (3.1.4), we can find the ACF of the regression model is not stationary, so we take the first difference of this regression model to make it stationary.

Figure 3.1.5 (First difference on regression model)

Forecast Model	Model Title	Mean Absolute Percent Error
☐	wind_speed + humidity	23.14179
☑	wind_speed + humidity + I(1)	3.11296

After taking the first difference, we can see the Mean Absolute Percent Error is reduced from 23.14179 to 3.11296. Next, we are going to develop error models based on the regression model with first differencing.

3.2 Error model using regression residuals

Figure 3.2.1 (ACF and PACF of differencing model)

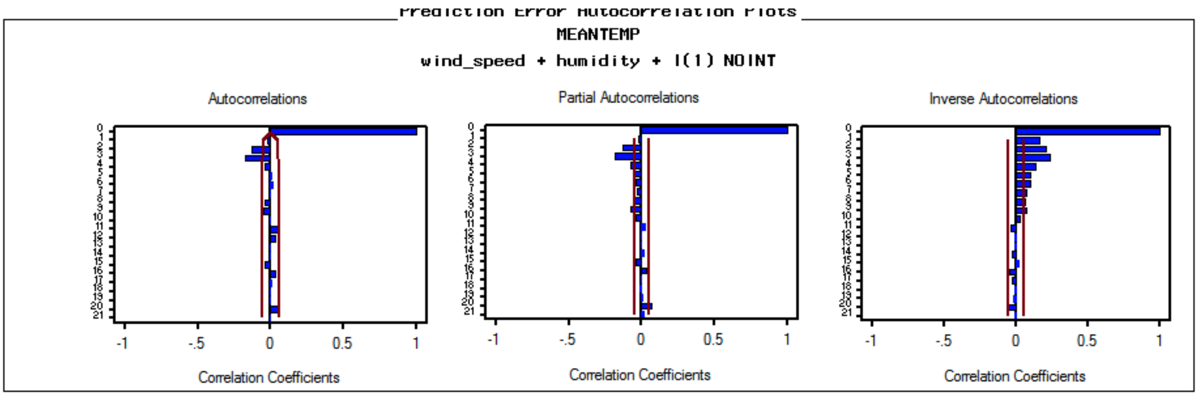


Figure 3.2.2 (Parameters of differencing model)

Parameter Estimates

MEANTEMP

wind_speed + humidity + I(1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
wind_speed	-0.03181	0.0074	-4.2783	<.0001
humidity	-0.13291	0.0045	-29.7241	<.0001
Model Variance (sigma squared)	1.66315	.	.	.

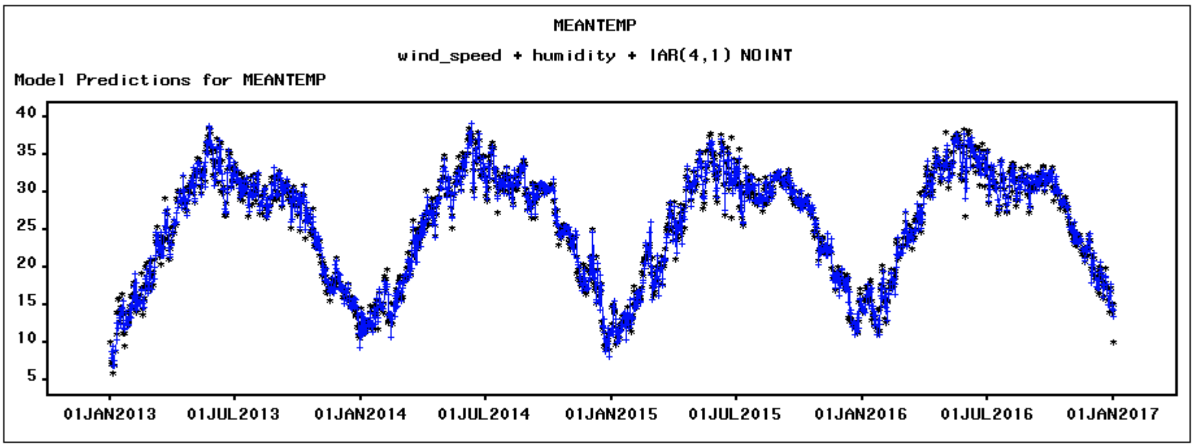
Fit Range: 01JAN2013 to 15JUN2016

After taking the first difference, the IACF decays quickly, ACF chops off after lag 3 and PACF chops off after lag 4. Therefore, This is very likely to apply to MA(4), AR(3) or ARIMA (1,1,1) models. Because the p-value of the intercept is not significant, so we decide not to add the intercept to our model.

Figure 3.2.3 (Performance of ARIMA models)

Forecast Model	Model Title	Mean Absolute Percent Error
☐	wind_speed + humidity	23.14179
☐	wind_speed + humidity + I(1) NOINT	3.11171
☐	wind_speed + humidity + ARIMA(1,1,1) NOINT	3.11722
☐	wind_speed + humidity + IMA(1,3) NOINT	3.08585
☑	wind_speed + humidity + IAR(4,1) NOINT	3.03959

Figure 3.2.4 (Actual values vs. predicted values of IAR(4,1))



After fitting the above models, we found they have similar Mean Absolute Percent Error. Then, we pick the IAR(4,1) model to find the graph of actual vs. predicted values. Figure (3.2.4) indicates the predicted values have a similar pattern as the actual values.

Figure 3.2.5 (Parameters of IAR (4,1))

Parameter Estimates

MEANTEMP

wind_speed + humidity + IAR(4,1) NOINT

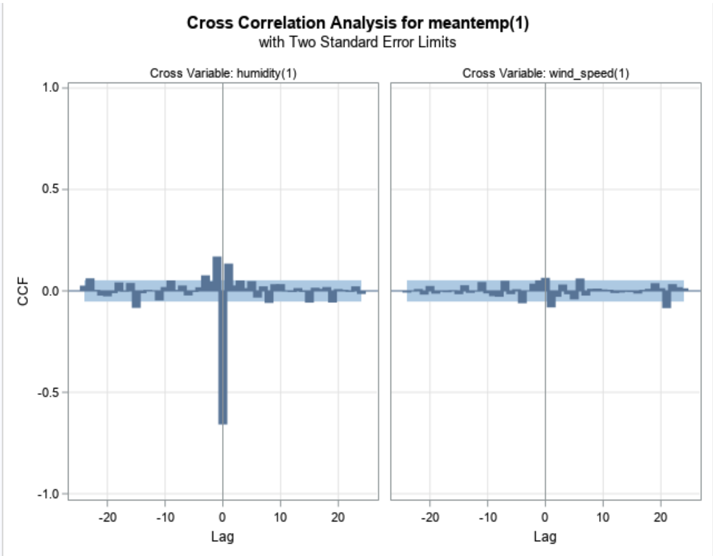
Model Parameter	Estimate	Std. Error	T	Prob>|T|
Autoregressive, Lag 1	-0.04316	0.0284	-1.5214	0.1298
Autoregressive, Lag 2	-0.14502	0.0278	-5.2102	<.0001
Autoregressive, Lag 3	-0.18659	0.0278	-6.7230	<.0001
Autoregressive, Lag 4	-0.06596	0.0282	-2.3358	0.0205
wind_speed	-0.02969	0.0075	-3.9699	0.0001
humidity	-0.13597	0.0045	-30.0171	<.0001
Model Variance (sigma squared)	1.57610	.	.	.

Based on these parameters in the figure (3.2.5), we can be certain that predictors ‘wind_speed’ and ‘humidity’ are useful in predicting the dependent variable ‘meantemp’.

3.3 Cross correlation analysis

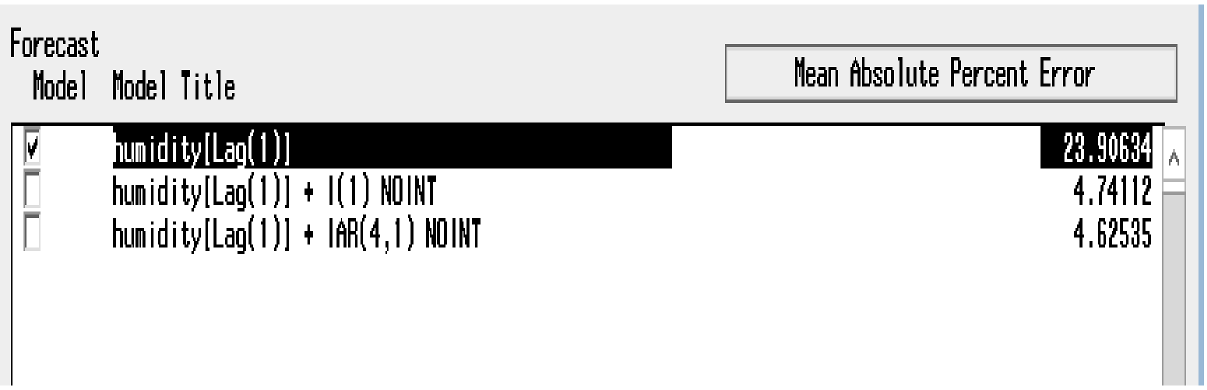
Since only predictors ‘wind_speed’ and ‘humidity’ are useful in predicting dependent variable ‘meantemp’ and these predictors are not stationary before first differencing, we look at their Cross Correlation Analysis after the first differencing.

Figure 3.3.1 (Cross Correlation plot)



Based on the figure (3.3.1), we can find that variable ‘humidity’ at lag 0, lag 1 and lag -1 have some correlations with the ‘meantemp’ after the first difference. Moreover, we can conclude that we can use the ‘humidity’ value at today and yesterday to predict the dependent variable ‘meantemp’ at today. Later, we are going to fit a cross correlation model based on the results.

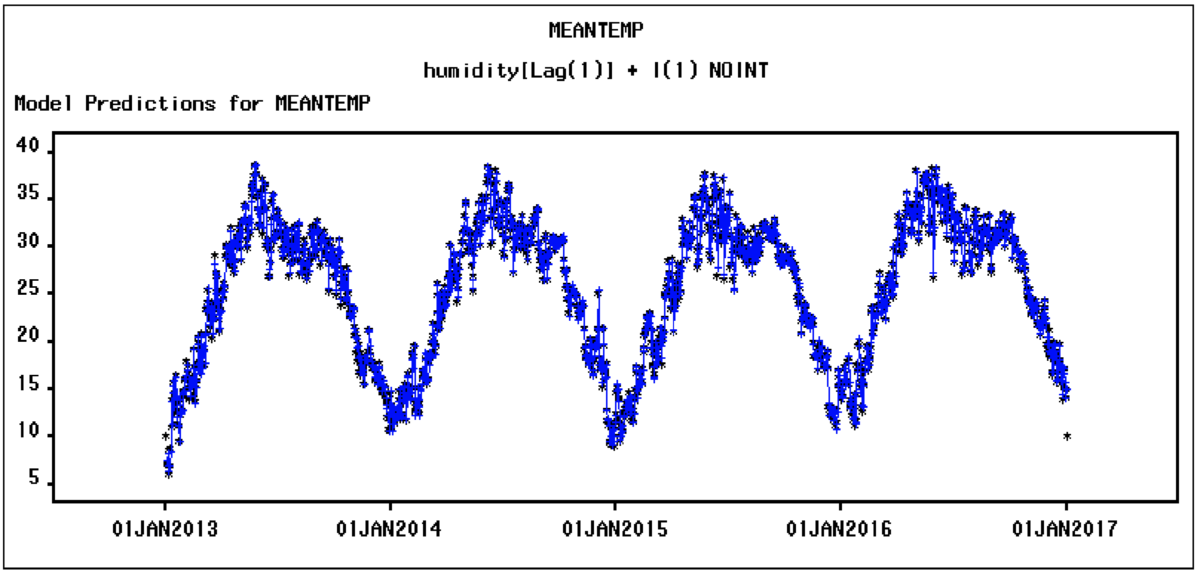
Figure 3.3.2(Performance of error models)



Forecast Model	Model Title	Mean Absolute Percent Error
☑	humidity[Lag(1)]	23.90634
☐	humidity[Lag(1)] + I(1) NOINT	4.74112
☐	humidity[Lag(1)] + IAR(4,1) NOINT	4.62535

In figure (3.3.2), we fitted 3 cross correlation models. Since only one predictor 'humidity' at lag 1 is useful in predicting the dependent variable, we apply 'humidity' in the model. Moreover, we added the error terms to the model, but the predicting performance MAPE does not improve much.

Figure 3.3.3 (Predicted values vs. actual values of first difference)



Based on the figure (3.3.3), we can see the predicted values and actual values are close and their trends also match each other.

Figure 3.3.4 (Parameters of first difference)

Parameter Estimates

MEANTEMP

humidity[Lag(1)] + I(1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
HUMIDITY[Lag(1)] Lag1	0.02747	0.0056	4.8650	<.0001
Model Variance (sigma squared)	2.77871	.	.	.

On the figure 3.3.4, the predator 'humidity' at lag 1 has a significant p-value (less than 0.05). Therefore, we can conclude 'humidity' at lag 1 is useful in predicting the dependent variable 'menatemp'.

4. Conclusion

Finally, we compare the model fit using the square root of model variance estimate. And we compare model predictive performance using MAPE.

Models	Square root of model variance	MAPE
Seasonal dummies and trend	2.5602	6.59204
Seasonal dummies and trend + ARMA(1,3)	1.62354	4.69438
Cyclical Model	2.7045	41.7666
Cyclical Model + AR(4)	1.6089	7.22072
ARIMA (2,1,2)	1.6085	4.43923
Regression Model	5.8189	23.09708
Regression Model + I(1)	1.2896	3.11296
Regression Model + IAR(4,1)	1.2554	3.03959
Cross-correlation (humidity)	6.0035	21.74557

Cross-correlation (humidity) + I(1)	1.6669	4.74112

Based on the square root of variance, the regression model with variables 'wind_speed' and 'humidity' after differencing has the best model fit. Besides the regression model; cyclical model + AR(4), ARIMA(2,1,2) and Cross-correlation (humidity) after differencing all have pretty good model fit. Moreover, the regression model after differencing has the best predictive performance with MAPE (3.03959), which indicates we could receive the best results in predicting mean temperature in Delhi India by using variables 'wind_speed' and 'humidity'.