Bike Sharing Project

Group Member:

Qiang Wang G46077311

Abby Liu G44206031

Jamie Pan G23429598

Xinrong Chen G40861353

Qianying Diao G43856749

1. Introduction and Overview

1.1 Introduction

Bike Sharing System is an innovation of traditional vehicles rental, where the whole processes including member registration, rental, and ends become automatic so that users are able to rent sharing bikes and return them conveniently. Also, bike sharing industry is booming due to their important role in traffic, environmental and health issues.

This dataset captures the daily count of rental bikes from 2011 to 2012 in Capital Bikeshare System in Washington D.C., which is the dependent variable, and the corresponding weather information is also provided serving as the independent variables to conduct time series prediction.

1.2 Dataset Overview

dteday: Date key, starting from Jan.1st 2011 to Dec.31th 2012

atemp: Normalized feeling temperature in Celsius. The values are derived via

(t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

windspeed: Normalized wind speed. The values are divided to 67 (max)

cnt: Count of total rental bikes including both casual and registered

Prcp: Historical daily precipitation observations for DC area. Unit in mm.

PRCP_LAG: Observations with tomorrow's precipitation condition in TF model with regressors.

PRCP_STD: Standardized value based on PRCP_LAG.

Cnt_delta: the differenced bike trips counts.

cnt _delta_std: the standardized bike trips differenced counts used in vector model.

Start_holiday: Dummy variables for important holiday start date used in TF model with regressors.

End_holiday: Dummy variables for important holiday end date used in TF model with regressors.

Mnth: Month of each observation.

Jan_Feb: Dummy variables for January and February.

June_July_Aug_Sep: Dummy variables for June, July, August and September.

There are 731 records in the dataset and we use 100 records as our hold-out sample.

1.3 Preview in SAS

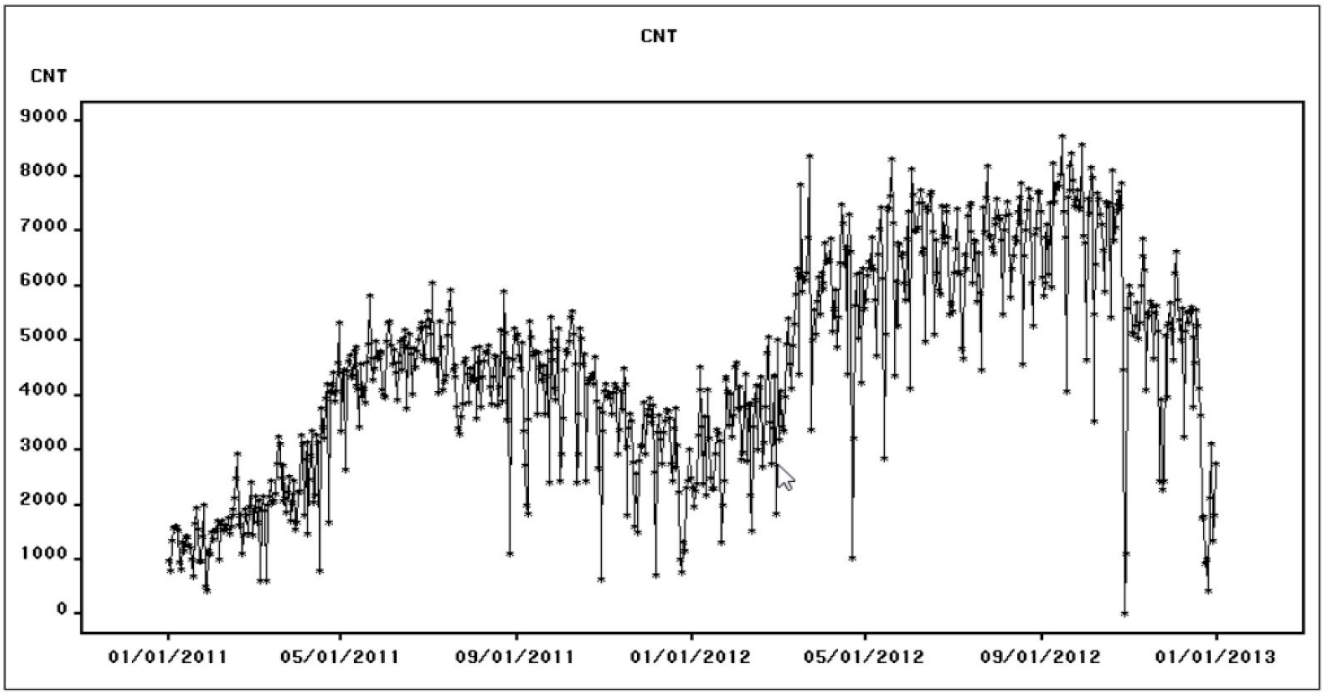

Figure 1

From Figure 1, we can find that it is nonstationary. In addition, a log transformation can be applied to the series, since there is an increasing change in the variability of the series over time. Once we take the first difference, it becomes stationary.

2. Univariate Time-series models.

2.1.1 Seasonal Dummies model

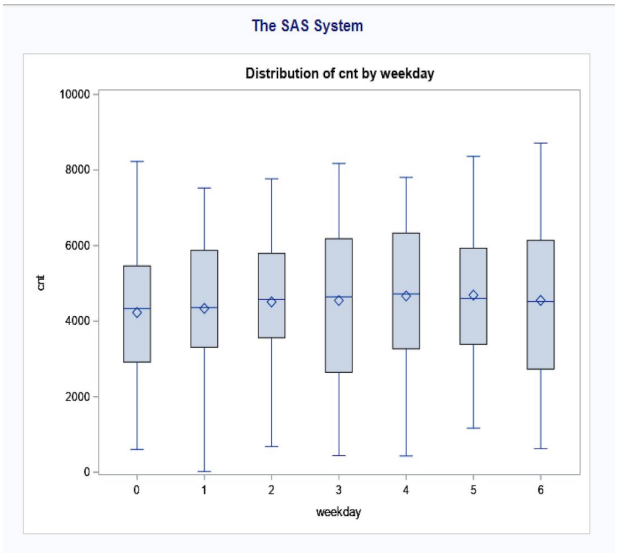

Figure 2

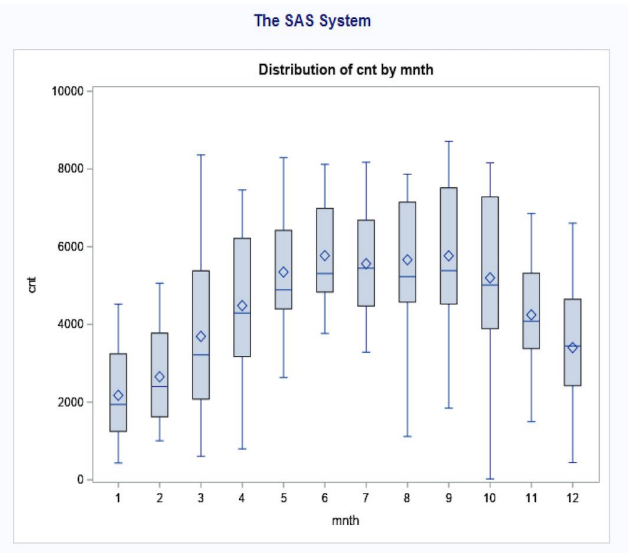

Figure 3

Figure 2 shows the distribution of the number of trips grouped by weekdays. The figure indicates no significant difference in terms of weekdays. Thus, using seasonal dummies by weekdays will be meaningless.

Figure 3 shows the distribution of the number of trips grouped by months. The figure indicates the number of trips will be influenced by months. During January and February, "cnt" reaches the minimum while from June to September, "cnt" reaches the peak. So we set two seasonal dummies to obtain the bottom and peak of "cnt". 'JAN_FEB' dummies distinguish Jan. and Feb with other months while 'JUNE_JULY_AUG_SEP' dummies distinguish June, July, August, and September with other months. The results are shown as follows.

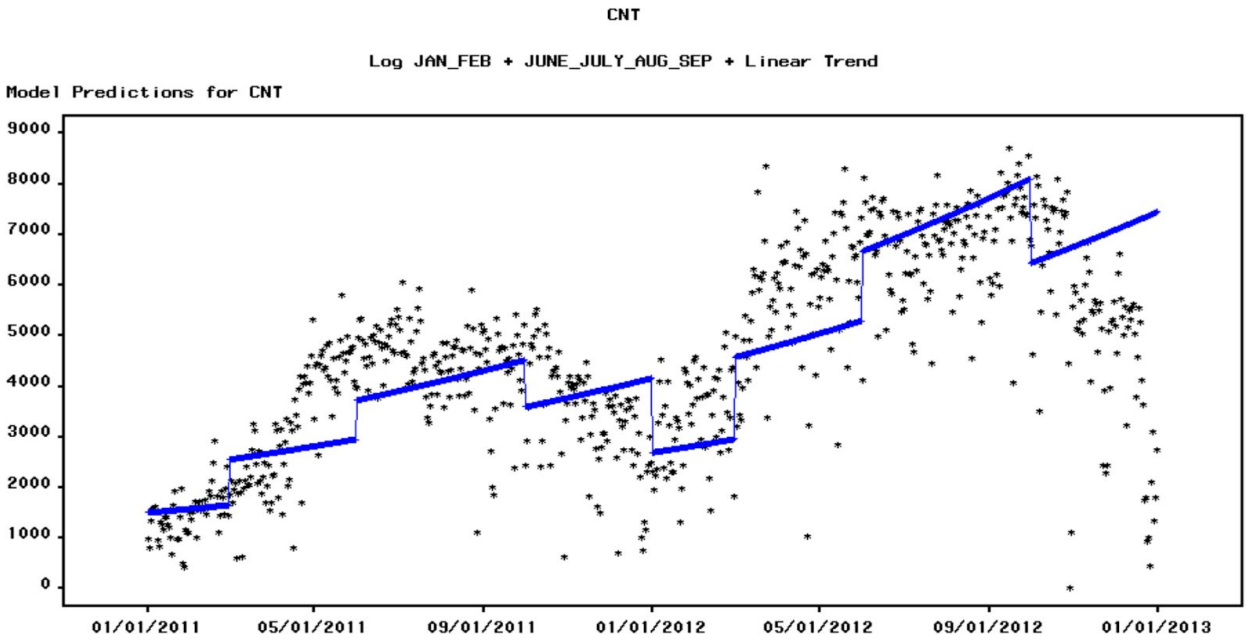

Figure 4

Parameter Estimates

CNT

Log JAN_FEB + JUNE_JULY_AUG_SEP + Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	7.75107	0.0311	249.4030	<.0001
JAN_FEB	-0.43567	0.0376	-11.5796	<.0001
JUNE_JULY_AUG_SEP	0.22996	0.0307	7.4814	<.0001
Linear Trend	0.00159	0.000079	20.2492	<.0001
Model Variance (sigma squared)	0.11403	.	.	.

Figure 5

Figure 4 indicates that the two dummies grasp generally grasp most of the ups and downs. The Parameter Estimates table from Figures 5 suggests both dummies are significant.

2.1.2 Cyclical Trend Model

2.1.2.1 Periodogram for cyclical trends model

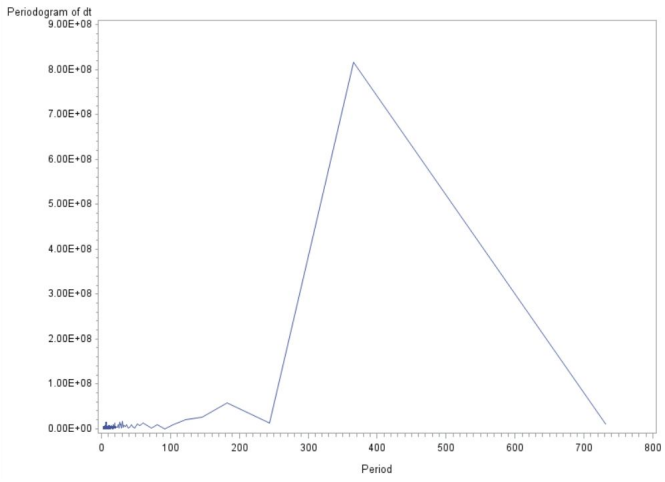

Figure 6

	Obs	FREQ	PERIOD	P_01
1				
2	3	0.01719	365.5	816186123
3	5	0.03438	182.75	58007683.1
4	6	0.04298	146.2	25420661.9
5	7	0.05157	121.833	20569437.7
6	25	0.20629	30.458	15529566
7	110	0.93689	6.706	15052975.3
8	29	0.24067	26.107	14403743.3
9	106	0.90251	6.962	14288650.9
10	13	0.10314	60.917	13474732.8
11	39	0.32662	19.237	12266775.6

Figure 7

As shown in Figure 6, when the harmonics are 2, 4, 5, 6, 24, 109, 28, 105, 13, 38, the value P_01(Figure 7) are highest, indicating the highest amplitudes and should be included in the cyclical trend model.

2.1.2.2 Cyclical trends model with log transformation and first difference

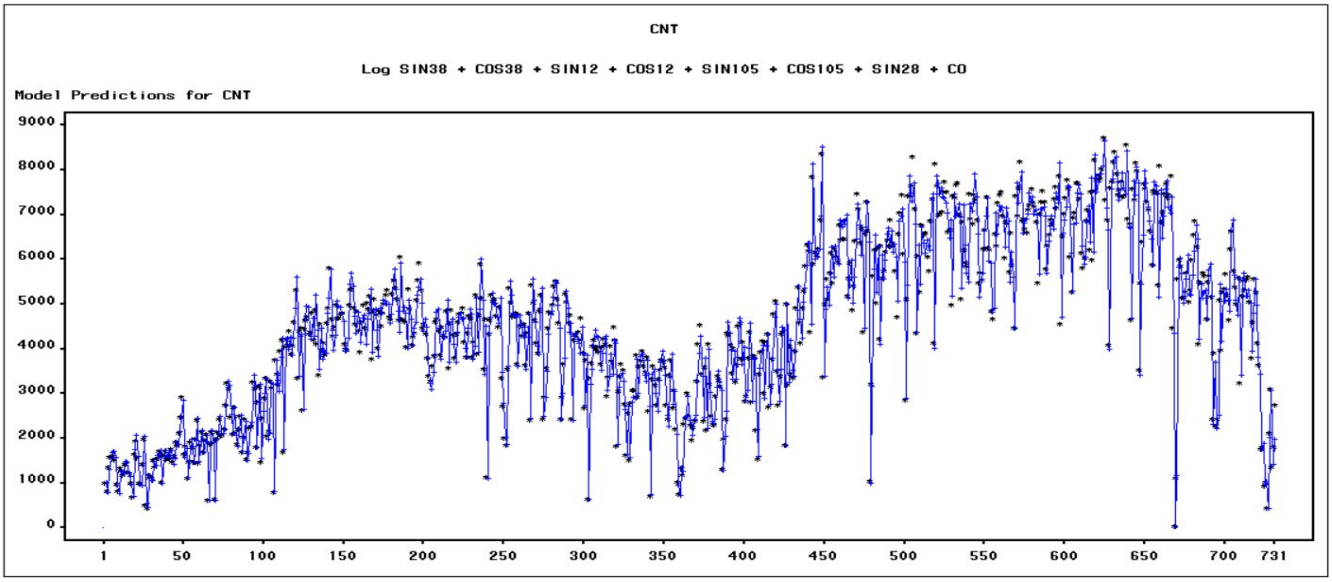

Figure 8

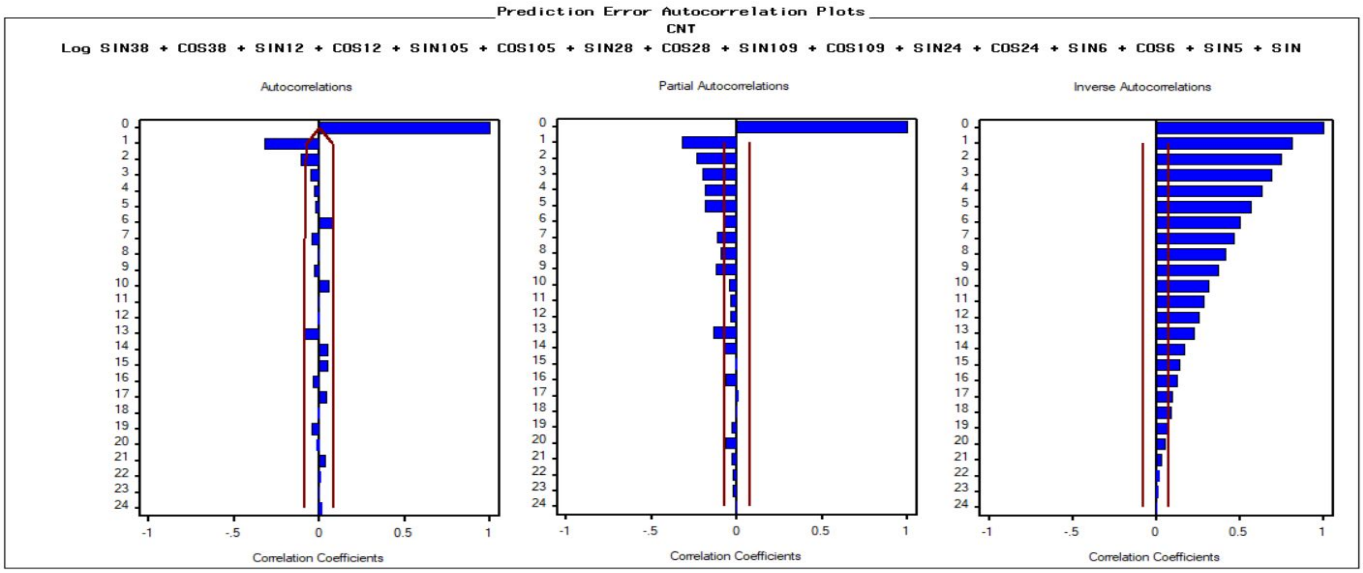

Figure 9

We first tried the basic cyclical model with log transformation and triangular regressors. The model has high RMSE of 2615.7 and the ACF plot shows that the errors in this model are nonstationary and not white noise. Thus, we take the difference of the series, since the first difference of series are stationary so that we can build the error model(Figure 9). The new model fits the series better(Figure 8) with much lower RMSE of 1310.7. Both ACF plot and PACF decays exponentially(Figure 9), which indicate that we may build an ARMA(1,1) model as the error model.

2.1.2.3 Cyclical trends model with log transformation and first difference and ARMA(1, 1) model

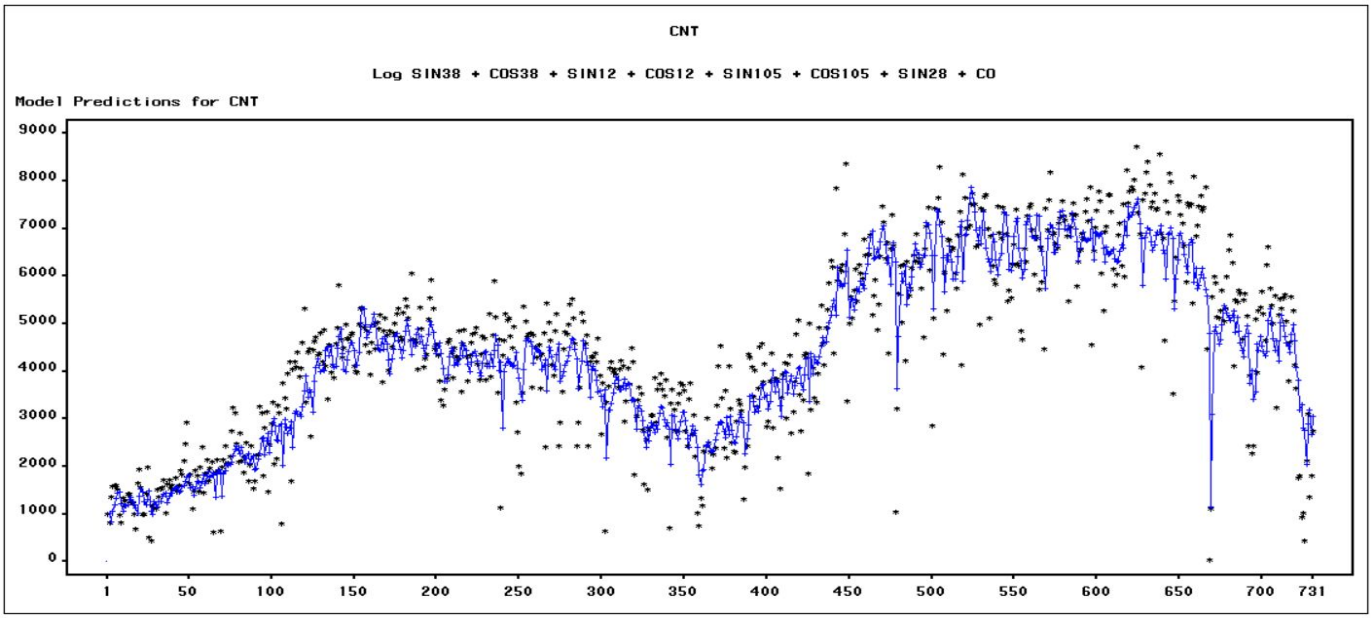

Figure 10

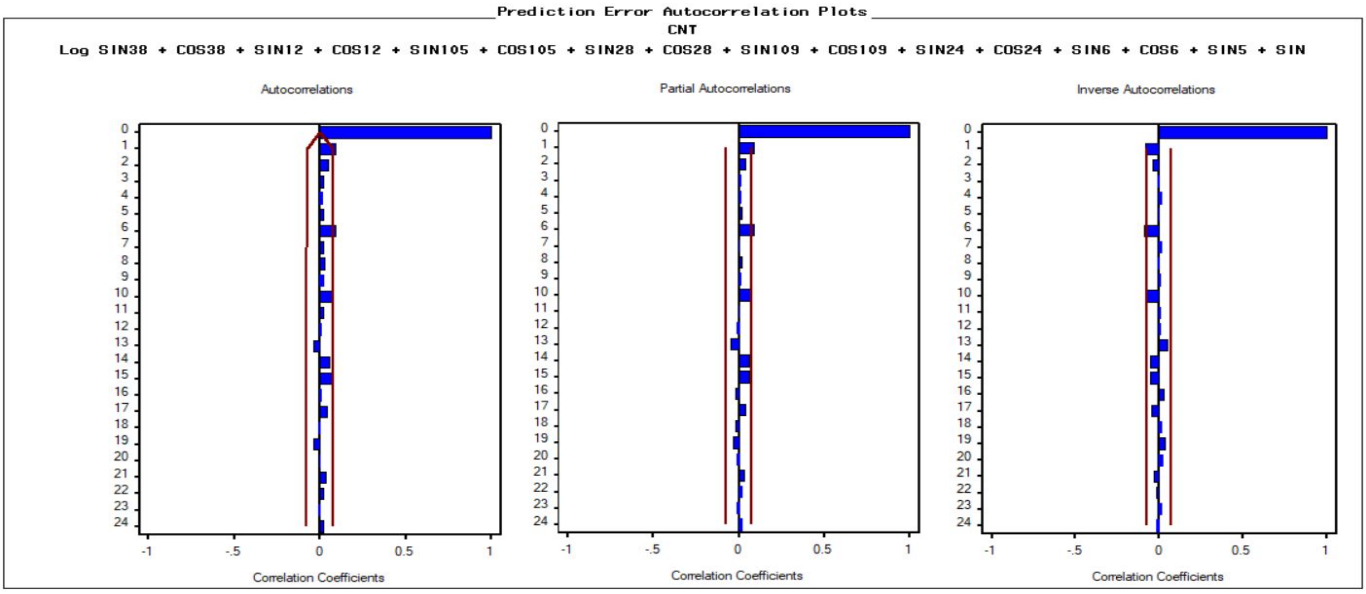

Figure 11

Then, we added the ARMA(1,1) as an error model to the previous model. The new model fits the series even better(Figure 10,11) with slightly lower RMSE of 1247.6(Figure 13). The ACF plot shows errors are white noise now, indicates the error model fits well.

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	0.00188	0.000512	3.6719	0.0004
Moving Average, Lag 1	0.96732	0.0113	85.3323	<.0001
Autoregressive, Lag 1	0.24225	0.0415	5.8306	<.0001
SIN38	0.03204	0.0194	1.6549	0.1020
COS38	-0.00183	0.0194	-0.0946	0.9249
SIN12	0.00992	0.0208	0.4777	0.6342
COS12	-0.02415	0.0208	-1.1625	0.2486
SIN105	0.02663	0.0172	1.5482	0.1257
COS105	-0.00443	0.0172	-0.2575	0.7975
SIN28	0.01024	0.0198	0.5179	0.6060
COS28	0.02786	0.0197	1.4126	0.1618
SIN109	0.03947	0.0171	2.3142	0.0233
COS109	0.02318	0.0171	1.3597	0.1779
SIN24	0.03396	0.0199	1.7088	0.0915
COS24	-0.02746	0.0199	-1.3816	0.1711
SIN6	0.03006	0.0251	1.1974	0.2348
COS6	-0.01871	0.0251	-0.7447	0.4587
SIN5	0.02608	0.0277	0.9434	0.3484
SIN4	-0.02891	0.0321	-0.9011	0.3703
COS5	0.05556	0.0270	2.0605	0.0427
COS4	-0.08702	0.0290	-2.9966	0.0037
SIN2	0.02967	0.0480	0.6177	0.5386
COS2	-0.40753	0.0483	-8.4422	<.0001
Model Variance (sigma squared)	0.07276	.	.	.

Figure 12

Statistic of Fit	Value
Mean Square Error	1556499.9
Root Mean Square Error	1247.6
Mean Absolute Percent Error	245.70072
Mean Absolute Error	992.69333
R-Square	0.595

Figure 13

Figure 13 states the estimated parameter for the best fit cyclical trends with ARMA(1, 1) error model. From the table, there are few regressors has very small p-values. which indicating statistically significant relationship to the count of count of bike trips. They're cos2, cos4, cos109, suggesting there exist seasonal patterns of period of 1 year, half year, and one week.

2.2 ARIMA models (with seasonal ARIMA components if relevant)

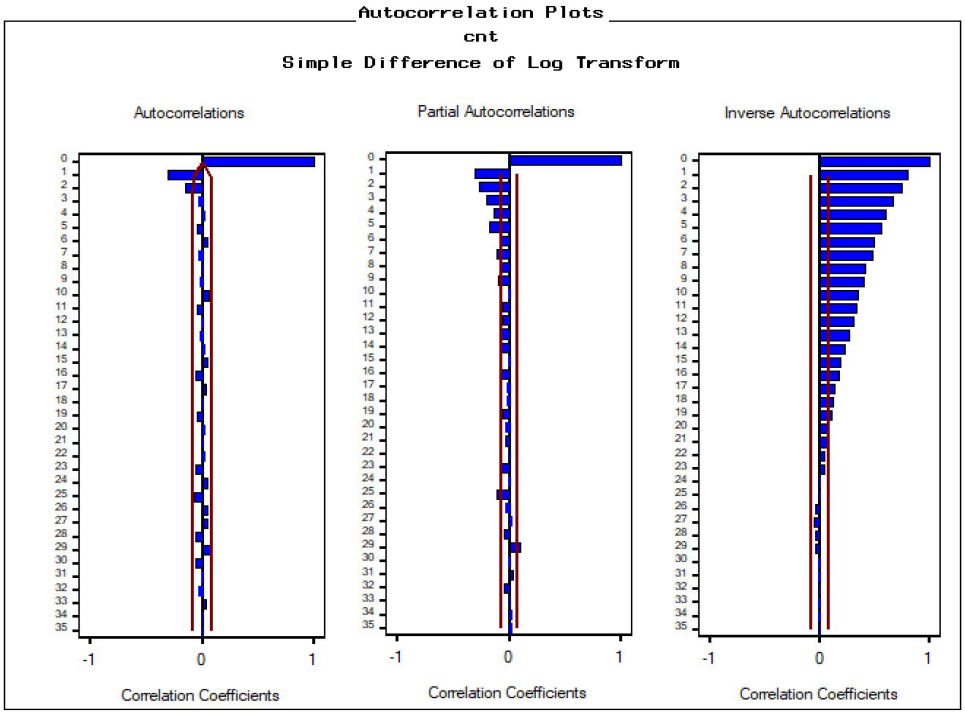

Figure 14

As we mentioned before, the original series is nonstationary. But when we take the first difference, the ACF decays quickly, which means the differencing is stationary. In addition, there is no obvious seasonality.

From the plot of ACF and PACF, we try to build an ARIMA(0,1,2) model and ARIMA(1,1,1) model. Comparison of these two models are as below:

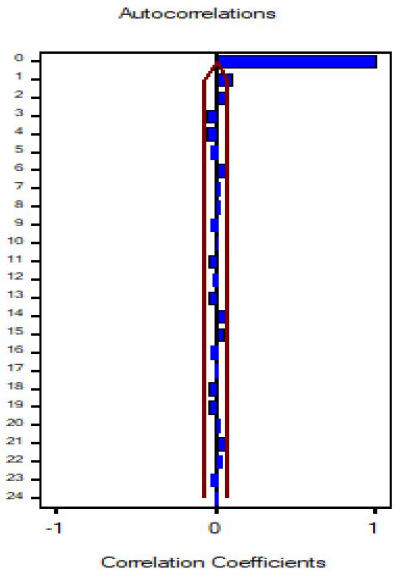

Figure 15

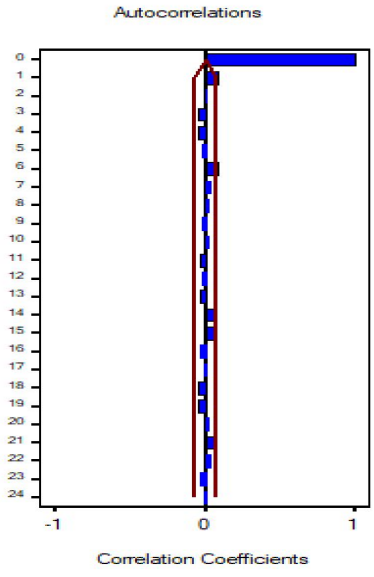

Figure 16

Statistic of Fit	Value
Mean Square Error	1756347.5
Root Mean Square Error	1325.3
Mean Absolute Percent Error	302.30539
Mean Absolute Error	953.34912
R-Square	0.543

Figure 17

Statistic of Fit	Value
Mean Square Error	1736875.0
Root Mean Square Error	1317.9
Mean Absolute Percent Error	302.29158
Mean Absolute Error	940.46107
R-Square	0.548

Figure 18

Figure 15 and Figure 16 are the ACF of errors for ARIMA(0,1,2) and ARIMA(1,1,1). Figure 17 and Figure 18 are the errors of ARIMA(0,1,2) and ARIMA(1,1,1). We can see that they are both white noise. However, the error of ARIMA(1,1,1) is less than that of ARIMA(0,1,2).

Based on the information above, we decided to use ARIMA(1,1,1). Figure 19 shows the model parameters, and they are both significant. So ARIMA(1,1,1) is appropriate.

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.89461	0.0216	41.4699	<.0001
Autoregressive, Lag 1	0.24745	0.0465	5.3253	<.0001
Model Variance (sigma squared)	0.07666	.	.	.

Figure 19

2.3 Comparison of models (in terms of fit and validation)

Model	Model Variance	Root Mean Square Error	Mean Absolute Percent Error
Seasonal Dummy	**0.11403**	**2651.9**	**388.38455**
Cyclical + ARIMA(1,1,1)	**0.07276**	**1247.6**	**245.70072**
ARIMA(0,1,2)	**0.07633**	**1325.3**	**302.30539**
ARIMA(1,1,1)	**0.07666**	**1317.9**	**302.29158**

*All models here take the log-transformation.

Table 20

In our exploration, we analyze two deterministic models, seasonal dummies with a linear trend model and cyclical model. We choose the cyclical model because it has a better performance in fitting the data. Then we fit the error term with an ARIMA(1,1,1) process. Another point needs to mention is, according to the behavior of ACF, our series does not display significant seasonality after taking the first difference, so we didn't include seasonal ARIMA component into our model.

In the third part, we fit the differenced, log-transformed series with an MA(2) and an ARMA(1,1) process. For each process, the modeling improvement is magnificent in terms of both fit and validation. When it comes to the comparison between two error models, the performance of ARIMA(1,1,1) is slightly better than ARIMA(0,1,2).

Here we put them together in table 1 for comparison in terms of fit and validation. Model Variance indicates how the model fits in with our model-training dataset, in terms of this, Cyclical + ARIMA(1,1,1) has the best output. RMSE and MAPE indicate how the models perform in predicting, Cyclical + ARIMA(1,1,1) is still the best model. Hence, in our following steps, we choose to use the ARMA(1,1) process to model the error term.

3. Transform Function Model

In our Multivariate Time Series Models, we identify 3 significant variables: Average Temperature, Humidity, Windspeed, correlated with the dependent variable: Count of Bike Trips.

3.1 Average Temperature

The ACF of atemp in the autocorrelation plot (Figure 21) decays slowly, implying the series is not stationary, and need differencing. Figure 22 shows the first differencing is stationary.

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	3747654	1.00000	|********************|	0
1	3171192	0.84618	|.|**************** |	0.036986
2	2920854	0.77938	|. |**************** |	0.057680
3	2794882	0.74577	|. |*************** |	0.070632
4	2760183	0.73651	|. |*************** |	0.080688
5	2780711	0.74199	|. |*************** |	0.089413
6	2827939	0.75459	|. |*************** |	0.097473
7	2769836	0.73909	|. |*************** |	0.105161
8	2719966	0.72578	|. |*************** |	0.112042
9	2656939	0.70896	|. |************** |	0.118299
10	2638394	0.70401	|. |************** |	0.123975
11	2594892	0.69240	|. |************** |	0.129328
12	2583964	0.68949	|. |************** |	0.134304
13	2581869	0.68893	|. |************** |	0.139062
14	2650477	0.70724	|. |************** |	0.143655
15	2631020	0.70204	|. |************** |	0.148342
16	2546974	0.67962	|. |************** |	0.152819
17	2521631	0.67286	|. |************* |	0.156899
18	2477313	0.66103	|. |************* |	0.160798

Figure 21

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.0035273	1.00000	|********************|	0
1	-0.0001051	-.02980	| *|. |	0.037012
2	-0.0008708	-.24687	| *****|. |	0.037045
3	-0.0006844	-.19401	| ****| . |	0.039234
4	-0.0002647	-.07504	| **| . |	0.040526
5	0.00016201	0.04593	| . |*. |	0.040716
6	0.00007950	0.02254	| . | . |	0.040787
7	0.00020827	0.05904	| . |*. |	0.040804
8	-0.0000171	-.00486	| . | . |	0.040921
9	0.00005495	0.01558	| . | . |	0.040922
10	-0.0001084	-.03072	| .*| . |	0.040930
11	-0.0002522	-.07149	| .*| . |	0.040962
12	0.00003732	0.01058	| . | . |	0.041132
13	0.00005937	0.01683	| . | . |	0.041136
14	0.00020384	0.05779	| . |*. |	0.041145
15	-0.0002413	-.06840	| .*| . |	0.041256
16	0.00013729	0.03892	| . |*. |	0.041412
17	0.00020401	0.05784	| . |*. |	0.041462
18	-0.0002971	-.08424	| **| . |	0.041572

Figure 22

3.1.1 Prewhitening Process

After applying ARIMA(1,1,2), the ACF of residuals in Figure 23 and Figure 24 are not significant and we cannot reject the hypothesis the residuals are white noise.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.83	3	0.0316	0.010	0.008	-0.087	-0.032	0.054	0.019
12	16.49	9	0.0572	0.074	0.000	0.024	-0.019	-0.060	0.019
18	28.39	15	0.0193	-0.005	0.068	-0.054	0.042	0.052	-0.062
24	33.94	21	0.0367	-0.002	0.026	0.043	0.052	0.028	-0.037
30	47.26	27	0.0093	0.019	0.022	0.104	0.069	0.029	-0.018
36	49.53	33	0.0323	-0.017	0.008	0.038	0.000	-0.018	0.029
42	54.40	39	0.0516	0.064	-0.033	0.006	-0.007	-0.004	0.033
48	59.05	45	0.0780	0.036	0.026	0.046	0.034	-0.004	0.027

Figure 23

Autocorrelation Plot of Residuals				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.0029838	1.00000	|********************|	0
1	0.00002897	0.00971	| .|. |	0.037012
2	0.00002407	0.00807	| .|. |	0.037015
3	-0.0002585	-.08665	| **|. |	0.037018
4	-0.0000969	-.03247	| *|. |	0.037294
5	0.00016127	0.05405	| .|* |	0.037333
6	0.00005671	0.01900	| .|. |	0.037440
7	0.00022105	0.07408	| .|* |	0.037453
8	2.09312E-7	0.00007	| . | . |	0.037654
9	0.00007045	0.02361	| . | . |	0.037654
10	-0.0000574	-.01925	| . | . |	0.037674
11	-0.0001782	-.05971	| .*| . |	0.037687
12	0.00005677	0.01903	| . | . |	0.037817

Figure 24

3.1.2 CCF and Parameters Identification

-4	-0.095448	-.00178	. | .
-3	-0.731280	-.01366	. | .
-2	2.001698	0.03739	. | *
-1	3.303776	0.06172	. | *
0	15.009403	0.28040	. | ******
1	-8.429890	-.15748	*** | .
2	1.154310	0.02156	. | .
3	-0.997582	-.01864	. | .
4	-0.581076	-.01086	. | .

Figure 25

The CCF bewteen first difference of CNT and first difference of ATEMP after prewhitening in Figure 25 is chopped off after lag 1, exhibiting significant cross-correlation at lag 0 and lat 1. Thus the TF parameters for the differenced ATEMP are r = 0, s = 1, b = 0.

3.1.3 Adequacy Check

We fit the TF model with r = 0, s = 1, b = 0 and a noise model of ARIMA(1,1,1). The high p-value in Figure 26 indicates the residuals are white noise. Figure 27 checks the correlations of atemp(1) and the residuals of the model, respectively. The high p-value suggests we cannot reject the null hypothesis that the two series are not cross correlated. So the transfer function model is appropriate.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.07	4	0.0595	0.010	-0.004	-0.046	-0.046	-0.019	0.087
12	11.90	10	0.2918	0.019	0.033	-0.036	0.010	-0.020	-0.024
18	20.48	16	0.1994	-0.027	0.055	0.073	-0.025	-0.006	-0.042
24	27.56	22	0.1907	-0.046	0.012	0.082	-0.011	-0.014	-0.003
30	48.17	28	0.0103	-0.051	0.034	0.049	0.063	0.114	-0.063
36	55.02	34	0.0127	0.010	-0.081	-0.006	0.037	0.030	0.002
42	72.18	40	0.0014	0.074	-0.081	0.026	-0.061	0.073	0.020
48	75.20	46	0.0042	-0.025	-0.020	-0.045	0.001	0.029	-0.001

Figure 26

Crosscorrelation Check of Residuals with Input atemp									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	1.64	5	0.8966	0.000	-0.017	0.023	0.025	-0.006	-0.028
11	2.95	11	0.9914	0.004	-0.009	0.034	-0.011	-0.012	-0.016
17	6.11	17	0.9924	0.013	0.001	0.021	0.028	0.024	0.048
23	17.20	23	0.7992	0.027	0.034	0.010	-0.000	0.105	0.048
29	24.83	29	0.6871	0.022	0.008	0.068	-0.069	0.024	-0.002
35	34.68	35	0.4833	0.006	-0.068	0.081	0.006	-0.048	-0.005
41	38.27	41	0.5926	0.003	-0.049	-0.001	0.040	-0.019	-0.022
47	41.52	47	0.6983	-0.016	0.028	0.043	-0.012	0.038	-0.003

Figure 27

The parameter estimate is as follows,

Parameter Estimates

CNT

Log atemp[Dif(1) N(1)] + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88566	0.0228	38.7638	<.0001
Autoregressive, Lag 1	0.23897	0.0474	5.0440	<.0001
ATEMP[Dif(1) N(1)]	1.25883	0.1630	7.7234	<.0001
ATEMP[Dif(1) N(1)] Num1	-0.29004	0.1629	-1.7803	0.0782
Model Variance (sigma squared)	0.07007	.	.	.

Figure 28

3.2 Humidity

The ACF of Humidity in the autocorrelation plot (Figure 29) decays quickly, implying the series is stationary.

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.020258	1.00000	|********************|	0
1	0.010694	0.52789	.|********** |	0.036986
2	0.0051439	0.25392	. |***** |	0.046156
3	0.0024513	0.12100	. |** |	0.048029
4	0.0010970	0.05415	. |*. |	0.048445
5	0.0017081	0.08431	. |** |	0.048527
6	0.0029745	0.14683	. |*** |	0.048727
7	0.0032087	0.15839	. |*** |	0.049329
8	0.0017417	0.08598	. |** |	0.050020
9	0.0010275	0.05072	. |*. |	0.050222
10	0.0016376	0.08083	. |** |	0.050292
11	0.0016682	0.08235	. |** |	0.050469
12	0.0020570	0.10154	. |** |	0.050652
13	0.00097137	0.04795	. |*. |	0.050930

Figure 29

3.2.1 Prewhitening Process

After applying ARIMA(1,0,1)(0,0,1)s, the ACF of residuals in Figure 30 and Figure 31 are not significant and we cannot reject the hypothesis the residuals are white noise.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	6.79	3	0.0787	0.000	0.006	0.002	-0.046	-0.004	0.084
12	15.93	9	0.0683	-0.004	-0.005	-0.039	0.052	0.014	0.088
18	26.98	15	0.0289	-0.020	-0.042	0.056	0.092	-0.021	0.020
24	28.71	21	0.1212	-0.009	0.013	-0.004	0.012	0.031	-0.030
30	34.24	27	0.1593	0.024	-0.010	0.072	0.033	0.011	-0.012
36	39.93	33	0.1892	-0.002	0.022	0.009	0.055	0.062	-0.000
42	48.00	39	0.1530	-0.033	0.044	0.028	-0.054	0.061	-0.008
48	53.71	45	0.1750	-0.022	0.029	0.014	0.033	0.029	-0.062

Figure 30

Autocorrelation Plot of Residuals				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.014465	1.00000	|********************|	0
1	1.59268E-7	0.00001	. |. |	0.036986
2	0.00008505	0.00588	. |. |	0.036986
3	0.00002663	0.00184	. |. |	0.036988
4	-0.0006662	-.04606	* |. |	0.036988
5	-0.0000532	-.00368	. |. |	0.037066
6	0.0012126	0.08383	. |** |	0.037067
7	-0.0000627	-.00434	. |. |	0.037325
8	-0.0000667	-.00461	. |. |	0.037326
9	-0.0005632	-.03894	* |. |	0.037327
10	0.00075920	0.05248	. |* |	0.037382
11	0.00020179	0.01395	. |. |	0.037483
12	0.0012749	0.08814	. |** |	0.037490
13	-0.0002959	-.02046	. | . |	0.037772

Figure 31

3.2.2 CCF and Parameters Identification

-3	-3.765657	-.02298	.|.
-2	3.589291	0.02190	.|.
-1	-0.289988	-.00177	.|.
0	-48.182018	-.29401	******|.
1	54.653864	0.33350	.|*******
2	-12.561408	-.07665	**|.
3	2.580246	0.01574	.|.
4	-4.892855	-.02986	*|.
5	2.072093	0.01264	.|.

Figure 32

The CCF of Humidity in Figure 32 is chopped off after lag 2, exhibiting significant cross-correlation at lag 0, lat 1 and lag 2. Thus the TF parameters for Humidity are r = 0, s = 2, b = 0.

3.2.3 Adequacy Check

We fit the TF model with r = 0, s = 2, b = 0 with a noise model of ARMA(1,1). The high p-value in Figure 33 indicates the residuals are white noise. Figure 34 checks the correlations of HUM and the residuals of the model, respectively. The high p-value suggests we cannot reject the null hypothesis that the two series are not cross correlated. So the transfer function model is appropriate.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.68	4	0.0697	0.011	-0.005	-0.047	-0.043	-0.008	0.087
12	14.55	10	0.1492	0.051	-0.005	0.004	-0.006	-0.034	-0.064
18	26.47	16	0.0478	-0.026	0.081	0.055	-0.024	0.030	-0.065
24	29.63	22	0.1277	-0.012	0.004	0.040	0.028	0.008	-0.040
30	49.13	28	0.0081	-0.053	0.037	0.041	0.069	0.111	-0.052
36	56.87	34	0.0083	-0.048	-0.059	-0.041	0.006	0.048	-0.018
42	64.52	40	0.0083	0.059	-0.062	-0.023	-0.014	0.003	0.043
48	66.97	46	0.0234	0.041	-0.014	-0.029	0.000	-0.001	0.020

Figure 33

Crosscorrelation Check of Residuals with Input hum									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	1.79	4	0.7736	-0.008	0.003	0.011	0.032	-0.034	0.008
11	7.17	10	0.7091	-0.033	0.060	-0.043	0.009	-0.015	0.023
17	10.64	16	0.8309	-0.035	0.038	0.019	0.006	0.022	0.036
23	15.85	22	0.8233	-0.026	-0.007	-0.009	-0.079	-0.001	0.002
29	21.37	28	0.8096	-0.069	-0.021	-0.022	-0.022	0.002	-0.038
35	25.90	34	0.8390	-0.016	-0.019	0.014	-0.070	0.017	0.013
41	34.65	40	0.7093	-0.029	-0.031	0.020	-0.056	0.068	-0.046
47	38.51	46	0.7754	-0.012	-0.004	0.028	0.021	0.035	-0.052

Figure 34

The parameter estimate is as follows,

Parameter Estimates

CNT

Log hum[N(2)] + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88425	0.0234	37.7280	<.0001
Autoregressive, Lag 1	0.27339	0.0479	5.7023	<.0001
HUM[N(2)]	-0.79837	0.0838	-9.5217	<.0001
HUM[N(2)] Num1	-0.14301	0.0845	-1.6934	0.0937
HUM[N(2)] Num2	-0.03816	0.0833	-0.4584	0.6477
Model Variance (sigma squared)	0.06710	.	.	.

Figure 35

3.3 Wind speed

The ACF of Wind speed in the autocorrelation plot (Figure 36) decays quickly, implying the series is stationary.

Autocorrelations				
Lag	Covariance	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.0059977	1.00000	|********************|	0
1	0.0019578	0.32642	.|******* |	0.036986
2	0.00033846	0.05643	. |*. |	0.040737
3	0.00033202	0.05536	. |*. |	0.040844
4	0.00025164	0.04196	. |*. |	0.040946
5	0.00047168	0.07864	. |** |	0.041005
6	0.00038130	0.06357	. |*. |	0.041211
7	0.00007717	0.01287	. | . |	0.041345
8	-0.0000713	-.01189	. | . |	0.041350
9	0.00020112	0.03353	. |*. |	0.041355
10	0.00056620	0.09440	. |** |	0.041392
11	0.00061524	0.10258	. |** |	0.041686

Figure 36

3.3.1 Prewhitening Process

After applying Factor Model, Q = (1) (16), the ACF of residuals in Figure 37 and Figure 38 are not significant and we cannot reject the hypothesis the residuals are white noise.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.19	4	0.3804	0.007	0.031	0.034	0.023	0.045	0.032
12	12.16	10	0.2743	0.010	-0.020	0.017	0.054	0.081	-0.020
18	18.13	16	0.3165	0.055	0.022	0.061	0.005	0.003	0.027
24	22.99	22	0.4022	-0.008	-0.050	0.035	0.005	-0.037	0.035
30	27.21	28	0.5070	0.014	0.009	-0.002	0.055	0.022	0.042
36	31.86	34	0.5728	-0.011	0.035	0.036	-0.033	0.032	0.035
42	39.86	40	0.4763	-0.021	0.041	0.027	0.045	0.058	0.045
48	45.18	46	0.5065	0.061	0.036	0.020	0.019	0.026	0.018

Figure 37

Autocorrelation Plot of Residuals				
Lag	Covariance	Correlation	−1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.0052311	1.00000	|********************|	0
1	0.00003778	0.00722	. | . |	0.036986
2	0.00016011	0.03061	. |* |	0.036988
3	0.00017747	0.03393	. |* |	0.037023
4	0.00011970	0.02288	. | . |	0.037065
5	0.00023341	0.04462	. |* |	0.037085
6	0.00016828	0.03217	. |* |	0.037158
7	0.00005168	0.00988	. | . |	0.037196
8	-0.0001043	-.01995	. | . |	0.037200
9	0.00009060	0.01732	. | . |	0.037214

Figure 38

3.3.2 CCF and Parameters Identification

-4	-4.122035	-.04570	I	*I . I
-3	2.962472	0.03284	I	. I* I
-2	2.472865	0.02741	I	. I* I
-1	-5.477368	-.06072	I	*I . I
0	-9.109581	-.10099	I	**I . I
1	4.740751	0.05255	I	. I* I
2	8.424005	0.09339	I	. I** I
3	-2.198354	-.02437	I	. I . I
4	3.459657	0.03835	I	. I* I

Figure 39

The CCF between Wind Speed and Count of Bike trips in Figure 39 is chopped off after lag 2, exhibiting significant cross-correlation at lag 0, lat 1, lag2. But when applying TF Model with r = 0, s = 2, b = 0, the Wind_Speed NUM2 Parameter is not statistically significant. Thus the TF parameters for Wind speed are r = 0, s = 1, b = 0.

3.3.3 Adequacy Check:

We fit the TF model with r = 0, s = 1, b = 0 with a noise model of ARMA(1,1). Figure 40 check the correlation of the residuals of TF model with WINDSPEED. The high p-value in Figure 40 indicates the residuals are white noise. Figure 41 checks the correlations of WINDSPEED and the residuals of the model, respectively. The high p-value suggests we cannot reject the null hypothesis that the two series are not cross correlated. So the transfer function model is appropriate.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	10.17	4	0.0376	0.003	0.010	-0.048	-0.048	-0.006	0.095
12	12.72	10	0.2397	0.023	0.009	-0.031	0.015	-0.036	-0.018
18	21.04	16	0.1771	-0.046	0.062	0.056	-0.023	-0.004	-0.039
24	25.41	22	0.2779	-0.036	0.006	0.055	0.032	0.002	-0.022
30	43.18	28	0.0334	-0.034	0.038	0.032	0.062	0.118	-0.043
36	46.68	34	0.0724	-0.026	-0.052	0.014	0.012	0.027	-0.009
42	55.37	40	0.0537	0.050	-0.070	0.034	-0.038	0.035	-0.004
48	57.64	46	0.1166	0.021	-0.011	-0.043	0.006	0.014	0.015

Figure 40

Crosscorrelation Check of Residuals with Input windspeed									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	11.50	5	0.0424	-0.007	-0.114	0.025	-0.000	0.042	0.017
11	14.71	11	0.1962	0.019	0.020	-0.018	0.003	0.050	-0.029
17	22.71	17	0.1588	-0.015	0.043	-0.041	-0.024	0.060	-0.056
23	28.39	23	0.2014	-0.007	0.060	-0.030	0.055	0.016	0.004
29	39.86	29	0.0863	-0.025	0.103	0.019	0.018	0.041	0.046
35	48.33	35	0.0663	0.039	0.087	0.037	0.033	-0.000	-0.004
41	52.14	41	0.1139	0.042	0.036	0.004	0.018	0.011	-0.041
47	54.26	47	0.2174	-0.008	0.035	-0.035	0.010	0.005	0.016

Figure 41

The parameter estimate is as follows,

Parameter Estimates

CNT

Log windspeed[N(1)] + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.88667	0.0226	39.2958	<.0001
Autoregressive, Lag 1	0.22713	0.0472	4.8085	<.0001
WINDSPEED[N(1)]	-0.74726	0.1488	-5.0222	<.0001
WINDSPEED[N(1)] Num1	0.26839	0.1489	1.8024	0.0746
Model Variance (sigma squared)	0.07366	.	.	.

Figure 42

3.4 Precipitation

The ACF of Precipitation in the autocorrelation plot (Figure 43) decays quickly, implying the series is stationary.

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.155738	1.00000	|********************|	0
1	0.019998	0.12841	.|*** |	0.036986
2	0.0048944	0.03143	. |*. |	0.037591
3	-0.0058701	-.03769	.*| . |	0.037627
4	-0.0024553	-.01577	. | . |	0.037679
5	-0.0035367	-.02271	. | . |	0.037688
6	0.0069971	0.04493	. |*. |	0.037707
7	0.0020894	0.01342	. | . |	0.037780
8	-0.0033320	-.02139	. | . |	0.037786
9	0.0032160	0.02065	. | . |	0.037803
10	0.0036919	0.02371	. | . |	0.037818

Figure 43

3.4.1 Prewhitening Process

After applying Factor Model, Q = (1) (11), the ACF of residuals in Figure 44 and Figure 45 are not significant and we cannot reject the hypothesis the residuals are white noise.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.08	4	0.3957	0.003	0.021	-0.039	-0.016	-0.033	0.047
12	5.18	10	0.8787	0.017	-0.015	0.016	-0.003	0.004	-0.027
18	15.22	16	0.5086	0.098	-0.003	0.043	0.003	0.033	-0.029
24	23.82	22	0.3569	-0.039	0.035	0.032	0.015	-0.024	0.083
30	27.52	28	0.4901	-0.023	-0.034	0.028	-0.038	-0.010	-0.030
36	29.75	34	0.6759	-0.034	0.032	-0.002	-0.024	0.012	-0.003
42	35.20	40	0.6860	0.069	-0.033	-0.011	-0.001	0.032	-0.009
48	37.15	46	0.8210	-0.018	0.000	-0.030	-0.025	0.018	-0.016

Figure 44

Autocorrelation Plot of Residuals

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.150816	1.00000	|********************|	0
1	0.00050741	0.00336	. | . |	0.036986
2	0.0031482	0.02087	. | . |	0.036987
3	-0.0058838	-.03901	* | . |	0.037003
4	-0.0023834	-.01580	. | . |	0.037059
5	-0.0050262	-.03333	* | . |	0.037068
6	0.0070705	0.04688	. | * |	0.037109
7	0.0025171	0.01669	. | . |	0.037190
8	-0.0022886	-.01517	. | . |	0.037200
9	0.0023712	0.01572	. | . |	0.037209
10	-0.0003916	-.00260	. | . |	0.037218
11	0.00063640	0.00422	. | . |	0.037218

Figure 45

3.4.2 CCF and Parameters Identification

-4	4.458571	0.01021	. | .	
-3	-13.586131	-.03111	* | .	
-2	-1.076865	-.00247	. | .	
-1	-171.457	-.39262	******** | .	
0	103.796	0.23768	. | *****	
1	46.412504	0.10628	. | **	
2	20.116708	0.04607	. | *	
3	-12.652386	-.02897	* | .	
4	7.717735	0.01767	. | .	
5	-10.037174	-.02298	. | .	

Figure 46

The CCF between Bike Users Count and Precipitation in Figure 46 exhibits both positive and negative significant lags, the interpretation of which is that bike users can check tomorrow's weather situations. The decision today will depend on precipitation volumes of yesterday, today and tomorrow. Thus, TF Model is not applicable in this feedback relationship.

3.5 Transfer Function model

In the TF Model aforementioned, the parameters identified corresponding to Differenced Average Temperature, Wind speed, Humidity are TF(0,0,1), TF(0,0,1), TF(0,0,2) respectively(r,b,s). Then we use SAS Forecast system to fit the model.

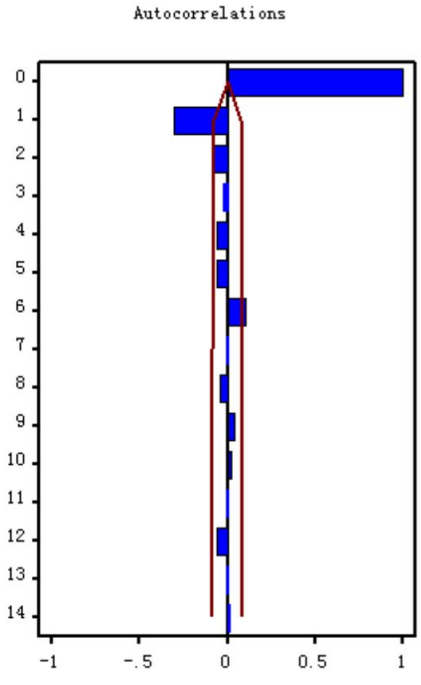

Figure 47

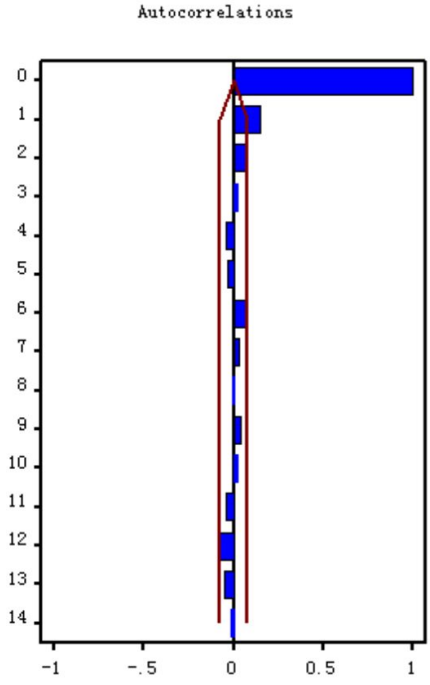

Figure 48

After fitting the model, As shown in Figure 47, the residuals are not white noise so an error model is needed. We decided to use ARIMA(1,1,1) to fit the residual. After fitting the error model, the residuals behave close to white noise(a little beyond the 2-standard-error bound at lag 2), as Figure 48 indicates.

The Parameter Estimations are showed as follows in Figure 49. It is clear that most of the parameters are significant. The model has a Root Mean Square Error of 1114.8 and Mean Absolute Percent Error of 202.25135.

Parameter Estimates

CNT

Log atemp[Dif(1) N(1)] + windspeed[N(1)] + hum[N(2)] + ARIMA(1,1,1) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.84890	0.0275	30.8383	<.0001
Autoregressive, Lag 1	0.20554	0.0505	4.0682	0.0001
ATEMP[Dif(1) N(1)]	1.08876	0.1513	7.1942	<.0001
ATEMP[Dif(1) N(1)] Num1	-0.39755	0.1473	-2.6984	0.0083
WINDSPEED[N(1)]	-0.86002	0.1338	-6.4267	<.0001
WINDSPEED[N(1)] Num1	0.49965	0.1425	3.5053	0.0007
HUM[N(2)]	-1.00334	0.0794	-12.6372	<.0001
HUM[N(2)] Num1	-0.18655	0.0783	-2.3819	0.0193
HUM[N(2)] Num2	-0.11360	0.0765	-1.4854	0.1409
Model Variance (sigma squared)	0.05477	.	.	.

Figure 49

Statistics of Fit

CNT

Log atemp[Dif(1) N(1)] + windspeed[N(1)] + hum[N(2)] + ARIMA(1,1,1) NOINT

Statistic of Fit	Value
Mean Square Error	1242800.8
Root Mean Square Error	1114.8
Mean Absolute Percent Error	202.25135
Mean Absolute Error	857.75356
R-Square	0.677

Figure 50

4. Transform Function Model with Regressors

4.1 Feedback Relationship with Precipitation

In the TF Model aforementioned, the parameters identified corresponding to Differenced Average Temperature, Wind speed, Humidity are TF(0,0,1), TF(0,0,1), TF(0,0,2) respectively (TF(r,b,s). In the meantime, the CCF plot between trips count and precipitation exhibits both positive and negative significant lags, the interpretation of which is bike users can check tomorrow's weather situations. The decision that if ride bike today will depend on precipitation volumes of yesterday, today and tomorrow.

-5	0.00063535	0.00061	.|.
-4	0.010599	0.01021	.|.
-3	-0.032296	-.03111	*|.
-2	-0.0025598	-.00247	.|.
-1	-0.407576	-.39262	********|.
0	0.246737	0.23768	.|*****
1	0.110328	0.10628	.|**
2	0.047820	0.04607	.|*
3	-0.030076	-.02897	*|.
4	0.018346	0.01767	.|.

Figure 51

4.2 Holiday Effect

Besides, we can detect some notable outliers substantially higher or lower than predicted values in former ARIMA or TF models in the time series plot, some of which are due to special events and the pattern of these will last in the following years. We assume that on the break starting day, people tend to stay at home with families and refresh themselves after tiring work and study and the change of Users Count will drop to obvious negative values. On the first working day then, the difference would be notably positive. So, dummy regressors are added to capture this impact. The Starting date and working start date of the specific holiday break are as below:

Holiday	**2011**		**2012**	
Martin Luther King Jr. Day	Jan. 15th	Jan. 18th	Jan. 14th	Jan. 17th
Valentine's Day		Feb. 14th		Feb. 14th
Presidents' Day	Feb. 19th	Feb. 22th	Feb. 18th	Feb. 21th
Labor Day	Sep. 3rd	Sep. 6th	Sep. 1st	Sep. 4th
Thanksgiving Day	Nov. 24th	Nov. 25th	Nov. 22th	Nov. 23th
Black Friday		Nov. 26th		Nov. 23th
Christmas Eve	Dec. 24th		Dec. 24th	

Table 52

The regressors of precipitation and tomorrow's precipitation situation as well as holiday break starting dummy are statistically significant, while the first working date after the holiday break is not, implying the prediction for the commute condition on that day is elusive and hard to infer as the result of the fact that days during the holiday break experience bike rental peaks and the difference is not significant then. However, peaks of the bike user count during the holiday are not stable and not the same in every holiday break, imposing lasting effects.

4.3 Model Identification

In sum, regressors of contemporary and tomorrow's Precipitation conditions, and starting date of holiday break dummy variable apart from the independent variables in the above TF model are applied in the combined TF + Regression model.

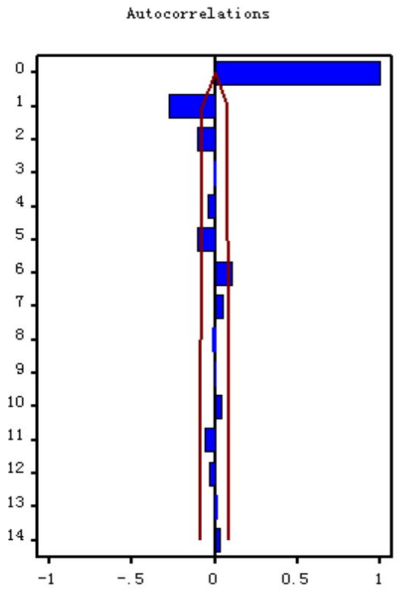

Figure 53

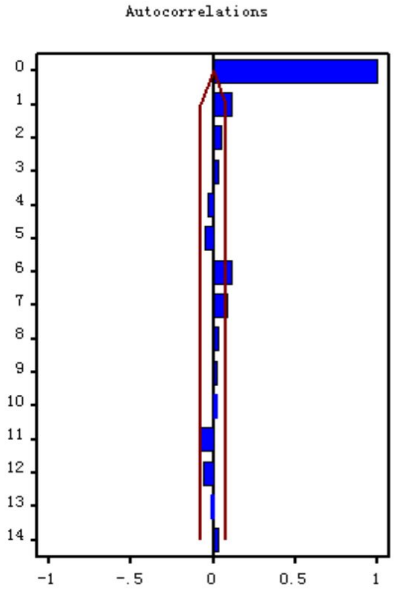

Figure 54

After fitting the model, As shown in Figure 53, the residuals are not white noise so an error model is needed. We decided to use ARIMA(1,1,1) to fit the residual. After fitting the error model, the residuals behave like white noise, as Figure 54 indicates.

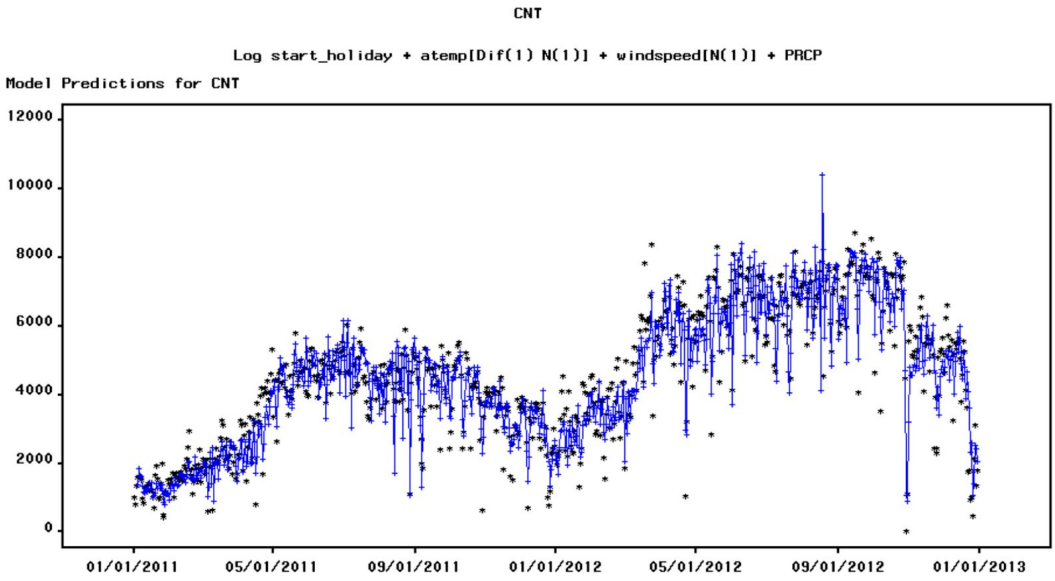

Figure 55

Parameter Estimates

CNT

Log start_holiday + atemp[Dif(1) N(1)] + windspeed[N(1)] + PRCP + PRCP_LAG + hum[N(2)] +

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	0.00283	0.0016	1.7485	0.0839
Moving Average, Lag 1	0.85473	0.0274	31.2373	<.0001
Autoregressive, Lag 1	0.24369	0.0508	4.7986	<.0001
start_holiday	-0.21953	0.0688	-3.1915	0.0020
ATEMP[Dif(1) N(1)]	1.01249	0.1352	7.4876	<.0001
ATEMP[Dif(1) N(1)] Num1	-0.40382	0.1318	-3.0632	0.0029
WINDSPEED[N(1)]	-0.61818	0.1222	-5.0581	<.0001
WINDSPEED[N(1)] Num1	0.33574	0.1299	2.5855	0.0114
PRCP	-0.05024	0.0244	-2.0563	0.0428
PRCP_LAG	-0.29191	0.0243	-12.0110	<.0001
HUM[N(2)]	-0.69635	0.0767	-9.0834	<.0001
HUM[N(2)] Num1	-0.18074	0.0730	-2.4753	0.0153
HUM[N(2)] Num2	-0.13229	0.0686	-1.9274	0.0572
Model Variance (sigma squared)	0.04379	.	.	.

Figure 56

Statistics of Fit

CNT

Log start_holiday + atemp[Dif(1) N(1)] + windspeed[N(1)] + PRCP + PRCP_LAG + hum[N(2)]

Statistic of Fit	Value
Mean Square Error	795430.9
Root Mean Square Error	891.86935
Mean Absolute Percent Error	68.48967
Mean Absolute Error	699.33603
R-Square	0.791

Figure 57

All the parameter estimates are quite significant, and ACF for errors in Figure 56are white noise. Also, the Root Mean Square Error and Mean Absolute Percent Error is the smallest among the several models.

5. Vector Model

The CCF plot between bike users count and precipitation aforementioned exhibits both positive and negative significant lags, which makes sense in the real world. Thus we applied vector model between these two variables which have been normalized first.

Schematic Representation of Cross Correlations													
Variable/Lag	0	1	2	3	4	5	6	7	8	9	10	11	12
cnt_delta_std	++	--	-.	-.	..	..	+.	..	..	..	..	..	..
PRCP1	++	++	..	..	..	..	..	..	..	..	..	.+	..
+ is > 2*std error, - is < -2*std error, . is between													

Figure 58

Schematic Representation of Partial Autoregression												
Variable/Lag	1	2	3	4	5	6	7	8	9	10	11	12
cnt_delta_std	-+	-+	-.	-+	-+	.+	..	..	..	..	..	..
PRCP1	-+	-+	-.	-.	-.	-+	-.	..	..	..	.+	..
+ is > 2*std error, - is < -2*std error, . is between												

Figure 59

Differenced bike users count exhibits two significant lags:

$\rho_1 = -0.291, \rho_2 = -0.109, \rho_3 = -0.076$ (ρ_3 *statistically significant but not very high*) in the diagonal autocorrelation in Figure 58 and the decaying pattern in partial autocorrelation in Figure 59. Precipitation exhibits one significant lags: $\rho_1 = -0.128$. Besides, the off-diagonal cell at lag 1 experienced significant feedback relationship. Thus VMA(2) is applied here and the model parameter estimates are as follows:

Model Parameter Estimates						
Equation	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Variable
cnt_delta_std	MA1_1_1	0.55096	0.03353	16.43	0.0001	e1(t-1)
	MA1_1_2	-0.05861	0.03596	-1.63	0.1036	e2(t-1)
	MA2_1_1	0.25936	0.03202	8.10	0.0001	e1(t-2)
	MA2_1_2	-0.00144	0.03470	-0.04	0.9669	e2(t-2)
PRCP1	MA1_2_1	0.55967	0.03711	15.08	0.0001	e1(t-1)
	MA1_2_2	-0.09792	0.03719	-2.63	0.0086	e2(t-1)
	MA2_2_1	0.06901	0.04136	1.67	0.0957	e1(t-2)
	MA2_2_2	-0.03966	0.04289	-0.92	0.3555	e2(t-2)

Figure 60

The matrix function is as below:

$$\begin{bmatrix}\Delta CNT\_STD \\ PRCP\_STD\end{bmatrix} = \begin{bmatrix}e_{1t} \\ e_{2t}\end{bmatrix} - \begin{bmatrix}0.55096 & -0.05861 \\ 0.55967 & -0.09792\end{bmatrix} \times \begin{bmatrix}e_{1t-1} \\ e_{2t-1}\end{bmatrix} - \begin{bmatrix}0.25936 & -0.00144 \\ 0.06901 & -0.03966\end{bmatrix} \times \begin{bmatrix}e_{1t-2} \\ e_{2t-2}\end{bmatrix}$$

In the diagnose of the error term, the interaction pattern disappears and the statistically significant cells at following lag3, lag6, lag11 lag 12, lag 13 are not very high, not worthy to whiten these lags.

Schematic Representation of Cross Correlations of Residuals																				
Variable/Lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
cnt_delta_std	+.	..	..	-.	..	..	+.	..	..	..	..	..	.-	..	..	..	..	..	..	..	..
PRCP1	.+	..	..	..	..	..	.+	..	..	..	..	.+	..	++	..	..	..	..	..	..	..
+ is > 2*std error, - is < -2*std error, . is between																					

Figure 61

Response to Comments

- **Draft 1**

1. Comments: Specify the start date of the dataset.

Response: We add the start date and end date of the dataset in 'dataset overview' on P1.

2. Comments: Specify the hold-out sample.

Response: We add the number of hold-out sample and total records in 'dataset overview' on P1.

3. Comments: Check the box-plot for seasonality and see whether a seasonal dummies model necessary.

Response: We add two box-plots to show the distribution grouped by weekdays and months. Since there is no significant seasonality for weekdays but obvious seasonality for months, we create two dummies and fit a seasonal dummies model to grasp the peak and bottom months. (P2-P3)

4. Comments: Add more harmonics for Cyclical Model.

Response: We add 4 more harmonics (up to 10 in total) on P4.

5. Comments: Try ARIMA (1,1,1) for ARIMA Model,

Response: We build ARIMA (1,1,1) and make a comparison with ARIMA (0,1,2) on P7-P9.

6. Comments: Add Mean Absolute Percent Error for Comparison Of Models.

Response: We add it and make a more detailed conclusion on P9.

- **Draft 2**

1. Comments: Show the ACF after differencing.

Response: Show the ACF after differencing for atemp on P10.

2. Comments: Point out this is between difference of cnt and difference of atemp after prewhitening.

Response: For the CCF of atemp, we point out that this is between difference of cnt and difference of atemp after prewhitening (P11).

3. Comments/Response: Show the estimated TF model.

4. Comment: How about ACF of residuals of the estimated error TF model?

Response: After fitting the error with an ARIMA(1,1,1) process for the initial TF model, the ACF behaves more close to white noise, with a little beyond 2-standard-error bound ACF at lag 2. And we observed that the ACF of error TF model with regressors is white noise.