Team Name: MASK
Team Members:

1. Chinnasubbareddygari Mohan Reddy (A0163433L)
2. Anusuya Manickavasagam (A0163300Y)
3. Sridharan Kesavan (A0163207M)

Dataset used for input: https://data.gov.sg/dataset/mobile-data-usage

Dataset used for output: https://data.gov.sg/dataset/total-number-of-outgoing-retail-international-telephone-call-minutes

Input and output dataset description: The input dataset is the mobile data usage in Singapore which includes mobile data sent over the mobile network. The unit of measure of the data present is "Petabytes". The output dataset is the total number of retail international calls outgoing in minutes. The entire dataset is available quarterly and taken from 2004 to 2016.

Problem statement: To use the mobile data usage as input(X) and international outgoing calls as output(Y) and apply the transfer function and finally to derive at an equation. The problem statement was arrived on the assumption that people use mobile data(messenger/skype/viber) for making international calls and hence the mobile international calls are reduced.

Model used: Initially, differencing was done to make the data stationary. Later, a seasonal effect was found on data. So, seasonal ARIMA was applied to the dataset. After doing the pre-whitening the transfer function was applied and the best model was selected.

Tool used: **"JMP pro"** was used to build the transfer function model. Open source tool **"Gretl"** was used for cointegration test.

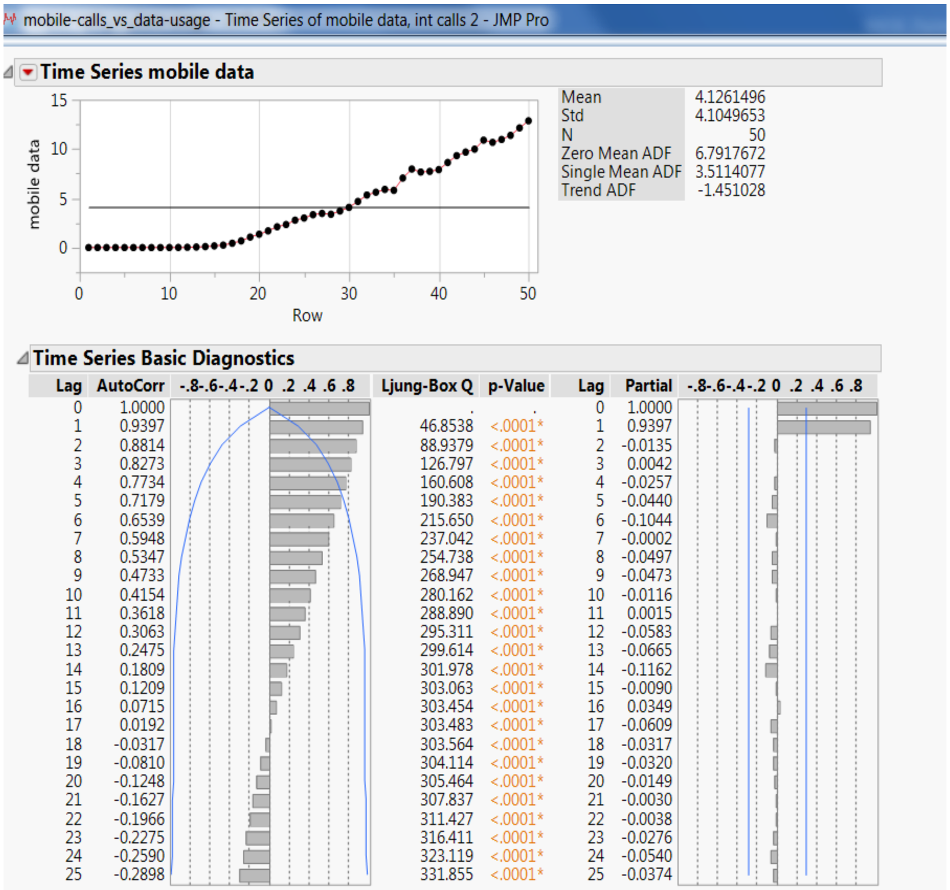
Summary:

1. After trial and error with various time series datasets, the best one was selected. This dataset has input data as the total mobile data usage in Petabytes and the total international calls in minutes as output.
2. The dataset was analyzed for stationarity and then differencing of first order was done to achieve stationarity.
3. After first order differencing the ADF results were checked for stationarity. Since, stationarity was not achieved second order differencing was done to achieve stationarity.
4. After differencing, there were significant lags observed at the 4,8,12,16 and 20 lags of the ACF plot. On analyzing the data further it was found that it has seasonal effect.
5. Next, the SARIMA model was fitted. Various orders of p,d,q,P,D,Q were fitted and model comparison was done. The model with least AIC and moderate MAPE was finally selected. The portmanteau test was done on the residual.
6. Cross-correlation was done on the input output data. As the correlation exists on both positive and negative side, Engle-Granger cointegration test was done using "Gretl" software.
7. As cointegration did not exist, the pre-whitening is done on the input series and the patterns suggest terms in the transfer function model.
8. The transfer model was applied with b and s values and the values of 1 and 2 were fitted for r for transfer function. Model comparison was done and the best model was selected.
9. The transfer function equation was noted down and interpretation was done.

Steps:

1. The dataset was fitted for time series in JMP pro by going to Analyze->Specialized Modelling->Time series. Initially, both the time series datasets were analyzed.

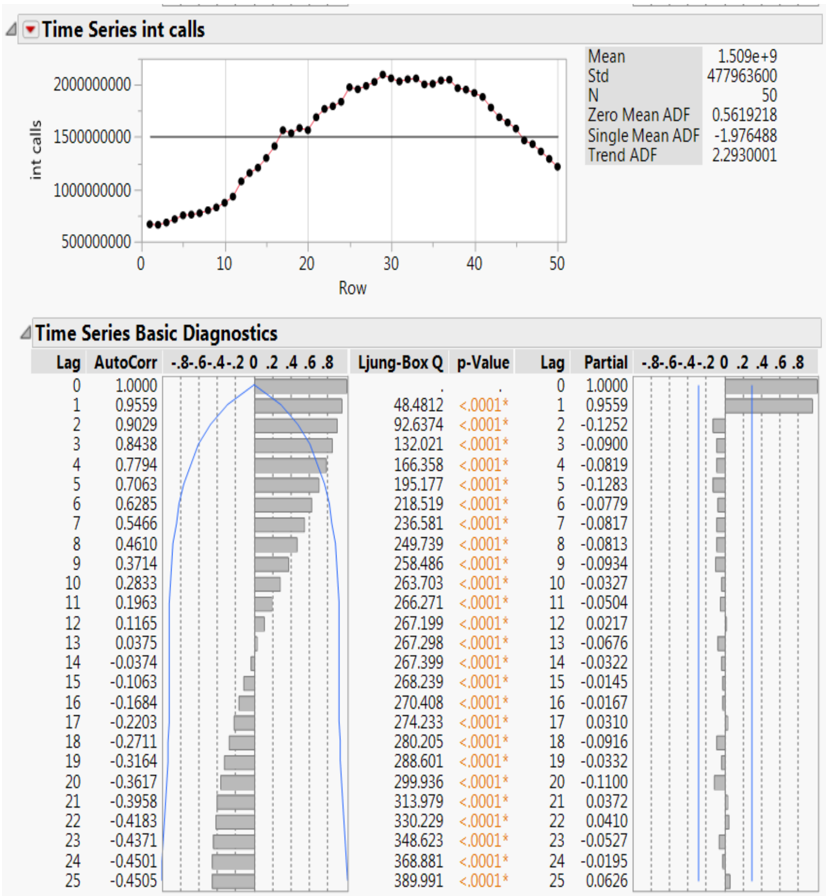
Input data



Mean	4.1261496
Std	4.1049653
N	50
Zero Mean ADF	6.7917672
Single Mean ADF	3.5114077
Trend ADF	-1.451028

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.9397	46.8538	<.0001*	1	0.9397
2	0.8814	88.9379	<.0001*	2	-0.0135
3	0.8273	126.797	<.0001*	3	0.0042
4	0.7734	160.608	<.0001*	4	-0.0257
5	0.7179	190.383	<.0001*	5	-0.0440
6	0.6539	215.650	<.0001*	6	-0.1044
7	0.5948	237.042	<.0001*	7	-0.0002
8	0.5347	254.738	<.0001*	8	-0.0497
9	0.4733	268.947	<.0001*	9	-0.0473
10	0.4154	280.162	<.0001*	10	-0.0116
11	0.3618	288.890	<.0001*	11	0.0015
12	0.3063	295.311	<.0001*	12	-0.0583
13	0.2475	299.614	<.0001*	13	-0.0665
14	0.1809	301.978	<.0001*	14	-0.1162
15	0.1209	303.063	<.0001*	15	-0.0090
16	0.0715	303.454	<.0001*	16	0.0349
17	0.0192	303.483	<.0001*	17	-0.0609
18	-0.0317	303.564	<.0001*	18	-0.0317
19	-0.0810	304.114	<.0001*	19	-0.0320
20	-0.1248	305.464	<.0001*	20	-0.0149
21	-0.1627	307.837	<.0001*	21	-0.0030
22	-0.1966	311.427	<.0001*	22	-0.0038
23	-0.2275	316.411	<.0001*	23	-0.0276
24	-0.2590	323.119	<.0001*	24	-0.0540
25	-0.2898	331.855	<.0001*	25	-0.0374

Output data



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.9559	48.4812	<.0001*	1	0.9559
2	0.9029	92.6374	<.0001*	2	-0.1252
3	0.8438	132.021	<.0001*	3	-0.0900
4	0.7794	166.358	<.0001*	4	-0.0819
5	0.7063	195.177	<.0001*	5	-0.1283
6	0.6285	218.519	<.0001*	6	-0.0779
7	0.5466	236.581	<.0001*	7	-0.0817
8	0.4610	249.739	<.0001*	8	-0.0813
9	0.3714	258.486	<.0001*	9	-0.0934
10	0.2833	263.703	<.0001*	10	-0.0327
11	0.1963	266.271	<.0001*	11	-0.0504
12	0.1165	267.199	<.0001*	12	0.0217
13	0.0375	267.298	<.0001*	13	-0.0676
14	-0.0374	267.399	<.0001*	14	-0.0322
15	-0.1063	268.239	<.0001*	15	-0.0145
16	-0.1684	270.408	<.0001*	16	-0.0167
17	-0.2203	274.233	<.0001*	17	0.0310
18	-0.2711	280.205	<.0001*	18	-0.0916
19	-0.3164	288.601	<.0001*	19	-0.0332
20	-0.3617	299.936	<.0001*	20	-0.1100
21	-0.3958	313.979	<.0001*	21	0.0372
22	-0.4183	330.229	<.0001*	22	0.0410
23	-0.4371	348.623	<.0001*	23	-0.0527
24	-0.4501	368.881	<.0001*	24	-0.0195
25	-0.4505	389.991	<.0001*	25	0.0626

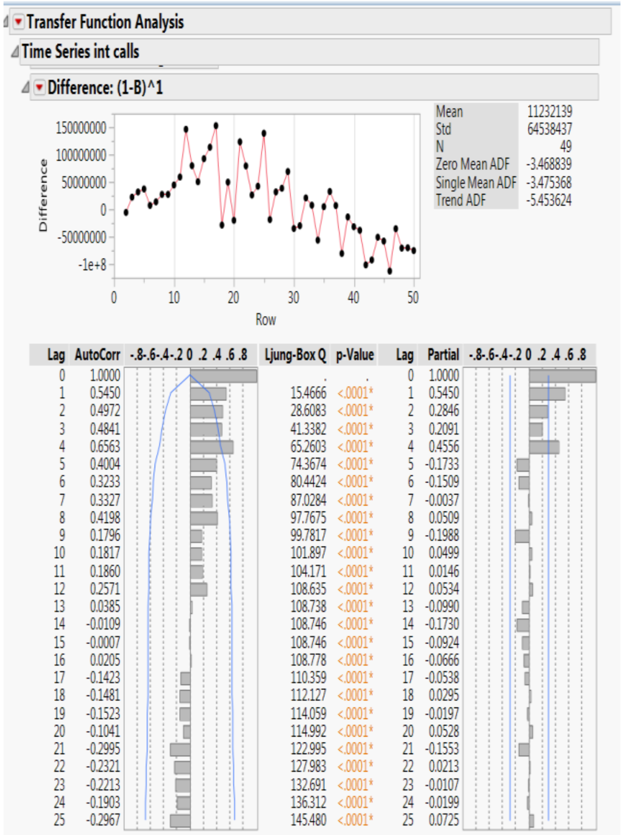
On analysis, we find that the input and output data have ACF in the decaying fashion up to a point beyond which the ACF increases on the negative side. The input and output show PACF at lag 1. On observation of the graphs of the input and the output we find that the data is not stationary. So, we do first order differencing on the input data first.

2. <u>ADF test</u>

   Null hypothesis: There is unit root

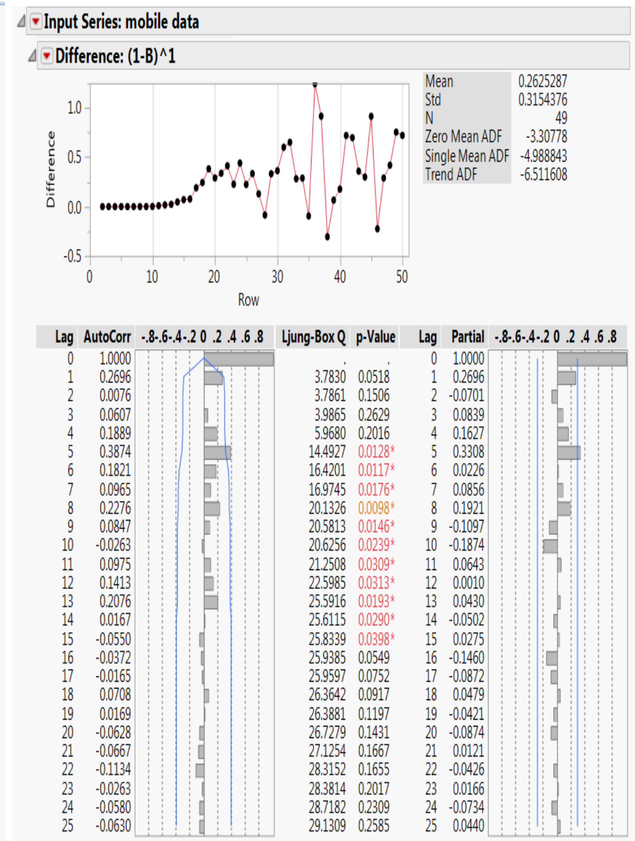
   Alternate hypothesis: The data is stationary.

   After doing the first order differencing we check the critical values of **Dickey Fuller t distribution**.



Mean	11232139
Std	64538437
N	49
Zero Mean ADF	-3.468839
Single Mean ADF	-3.475368
Trend ADF	-5.453624

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.5450	15.4666	<.0001*	1	0.5450
2	0.4972	28.6083	<.0001*	2	0.2846
3	0.4841	41.3382	<.0001*	3	0.2091
4	0.6563	65.2603	<.0001*	4	0.4556
5	0.4004	74.3674	<.0001*	5	-0.1733
6	0.3233	80.4424	<.0001*	6	-0.1509
7	0.3327	87.0284	<.0001*	7	-0.0037
8	0.4198	97.7675	<.0001*	8	0.0509
9	0.1796	99.7817	<.0001*	9	-0.1988
10	0.1817	101.897	<.0001*	10	0.0499
11	0.1860	104.171	<.0001*	11	0.0146
12	0.2571	108.635	<.0001*	12	0.0534
13	0.0385	108.738	<.0001*	13	-0.0990
14	-0.0109	108.746	<.0001*	14	-0.1730
15	-0.0007	108.746	<.0001*	15	-0.0924
16	0.0205	108.778	<.0001*	16	-0.0666
17	-0.1423	110.359	<.0001*	17	-0.0538
18	-0.1481	112.127	<.0001*	18	0.0295
19	-0.1523	114.059	<.0001*	19	-0.0197
20	-0.1041	114.992	<.0001*	20	0.0528
21	-0.2995	122.995	<.0001*	21	-0.1553
22	-0.2321	127.983	<.0001*	22	0.0213
23	-0.2213	132.691	<.0001*	23	-0.0107
24	-0.1903	136.312	<.0001*	24	-0.0199
25	-0.2967	145.480	<.0001*	25	0.0725



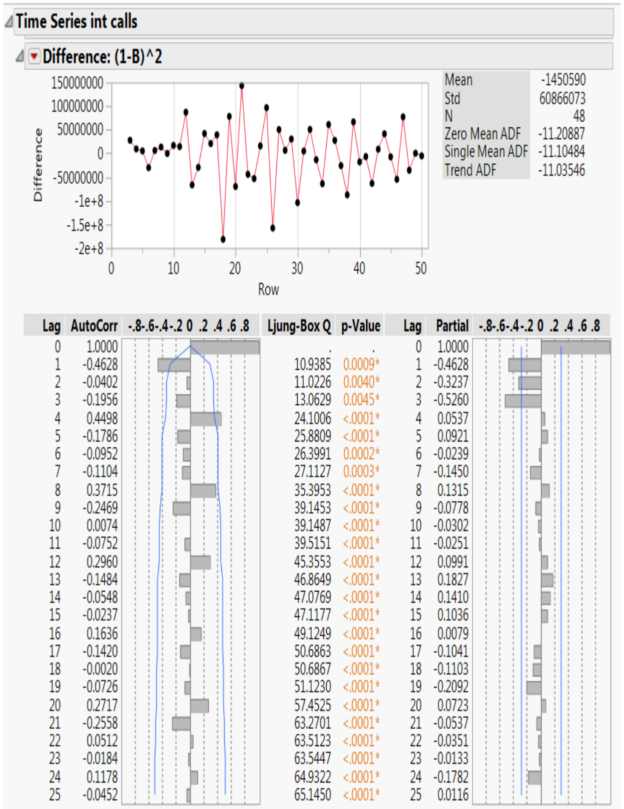
Mean	0.2625287
Std	0.3154376
N	49
Zero Mean ADF	-3.30778
Single Mean ADF	-4.988843
Trend ADF	-6.511608

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.2696	3.7830	0.0518	1	0.2696
2	0.0076	3.7861	0.1506	2	-0.0701
3	0.0607	3.9865	0.2629	3	0.0839
4	0.1889	5.9680	0.2016	4	0.1627
5	0.3874	14.4927	0.0128*	5	0.3308
6	0.1821	16.4201	0.0117*	6	0.0226
7	0.0965	16.9745	0.0176*	7	0.0856
8	0.2276	20.1326	0.0098*	8	0.1921
9	0.0847	20.5813	0.0146*	9	-0.1097
10	-0.0263	20.6256	0.0239*	10	-0.1874
11	0.0975	21.2508	0.0309*	11	0.0643
12	0.1413	22.5985	0.0313*	12	0.0010
13	0.2076	25.5916	0.0193*	13	0.0430
14	0.0167	25.6115	0.0290*	14	-0.0502
15	-0.0550	25.8339	0.0398*	15	0.0275
16	-0.0372	25.9385	0.0549	16	-0.1460
17	-0.0165	25.9597	0.0752	17	-0.0872
18	0.0708	26.3642	0.0917	18	0.0479
19	0.0169	26.3881	0.1197	19	-0.0421
20	-0.0628	26.7279	0.1431	20	-0.0874
21	-0.0667	27.1254	0.1667	21	0.0121
22	-0.1134	28.3152	0.1655	22	-0.0426
23	-0.0263	28.3814	0.2017	23	0.0166
24	-0.0580	28.7182	0.2309	24	-0.0734
25	-0.0630	29.1309	0.2585	25	0.0440

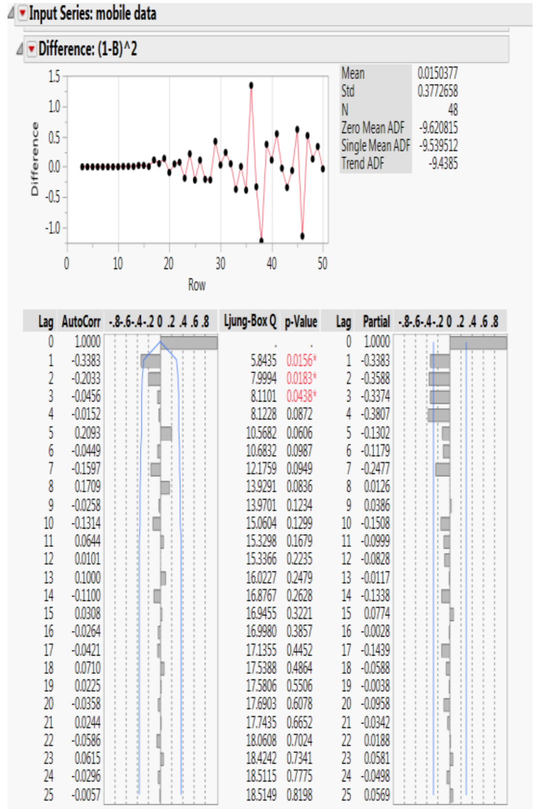
Since the data does not follow trend we compare the single mean ADF with the critical value of Dickey Fuller test. We find that single mean ADF value < critical value for input data. Hence, null hypothesis is rejected for input data. However, for output data the ADF value >critical value.

So, we do second order differencing on both input and output and observe the results.

3. Second order differencing



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.4628	10.9385	0.0009*	1	-0.4628
2	-0.0402	11.0226	0.0040*	2	-0.3237
3	-0.1956	13.0629	0.0045*	3	-0.5260
4	0.4498	24.1006	<.0001*	4	0.0537
5	-0.1786	25.8809	<.0001*	5	0.0921
6	-0.0952	26.3991	0.0002*	6	-0.0239
7	-0.1104	27.1127	0.0003*	7	-0.1450
8	0.3715	35.3953	<.0001*	8	0.1315
9	-0.2469	39.1453	<.0001*	9	-0.0778
10	0.0074	39.1487	<.0001*	10	-0.0302
11	-0.0752	39.5151	<.0001*	11	-0.0251
12	0.2960	45.3553	<.0001*	12	0.0991
13	-0.1484	46.8649	<.0001*	13	0.1827
14	-0.0548	47.0769	<.0001*	14	0.1410
15	-0.0237	47.1177	<.0001*	15	0.1036
16	0.1636	49.1249	<.0001*	16	0.0079
17	-0.1420	50.6863	<.0001*	17	-0.1041
18	-0.0020	50.6867	<.0001*	18	-0.1103
19	-0.0726	51.1230	<.0001*	19	-0.2092
20	0.2717	57.4525	<.0001*	20	0.0723
21	-0.2558	63.2701	<.0001*	21	-0.0537
22	0.0512	63.5123	<.0001*	22	-0.0351
23	-0.0184	63.5447	<.0001*	23	-0.0133
24	0.1178	64.9322	<.0001*	24	-0.1782
25	-0.0452	65.1450	<.0001*	25	0.0116



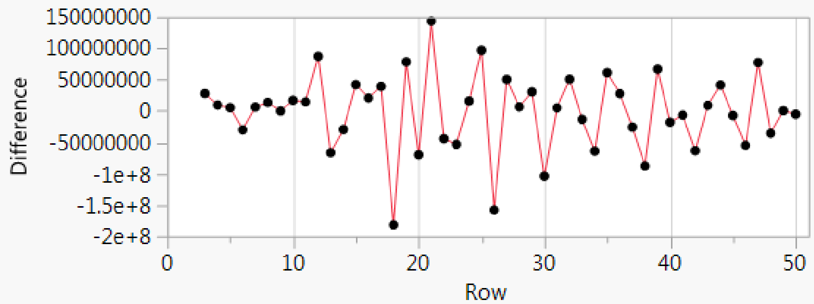
Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.3383	5.8435	0.0156*	1	-0.3383
2	-0.2033	7.9994	0.0183*	2	-0.3588
3	-0.0456	8.1101	0.0438*	3	-0.3374
4	-0.0152	8.1228	0.0872	4	-0.3807
5	0.2093	10.5682	0.0606	5	-0.1302
6	-0.0449	10.6832	0.0987	6	-0.1179
7	-0.1597	12.1759	0.0949	7	-0.2477
8	0.1709	13.9291	0.0836	8	0.0126
9	-0.0258	13.9701	0.1234	9	0.0386
10	-0.1314	15.0604	0.1299	10	-0.1508
11	0.0644	15.3298	0.1679	11	-0.0999
12	0.0101	15.3366	0.2235	12	-0.0828
13	0.1000	16.0227	0.2479	13	-0.0117
14	-0.1100	16.8767	0.2628	14	-0.1338
15	0.0308	16.9455	0.3221	15	0.0774
16	-0.0264	16.9980	0.3857	16	-0.0028
17	-0.0421	17.1355	0.4452	17	-0.1439
18	0.0710	17.5388	0.4864	18	-0.0588
19	0.0225	17.5806	0.5506	19	-0.0038
20	-0.0358	17.6903	0.6078	20	-0.0958
21	0.0244	17.7435	0.6652	21	-0.0342
22	-0.0586	18.0608	0.7024	22	0.0188
23	0.0615	18.4242	0.7341	23	0.0581
24	-0.0296	18.5115	0.7775	24	-0.0498
25	-0.0057	18.5149	0.8198	25	0.0569

On observing the single Mean ADF values, we see that the data is stationary for both input and output.

4. On observing the output data we see that there is a strong **seasonal** trend at 4,8,12,16,20 lags

Time Series int calls

Difference: (1-B)^2



Mean	-1450590
Std	60866073
N	48
Zero Mean ADF	-11.20887
Single Mean ADF	-11.10484
Trend ADF	-11.03546

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.4628	10.9385	0.0009*	1	-0.4628
2	-0.0402	11.0226	0.0040*	2	-0.3237
3	-0.1956	13.0629	0.0045*	3	-0.5260
4	0.4498	24.1006	<.0001*	4	0.0537
5	-0.1786	25.8809	<.0001*	5	0.0921
6	-0.0952	26.3991	0.0002*	6	-0.0239
7	-0.1104	27.1127	0.0003*	7	-0.1450
8	0.3715	35.3953	<.0001*	8	0.1315
9	-0.2469	39.1453	<.0001*	9	-0.0778
10	0.0074	39.1487	<.0001*	10	-0.0302
11	-0.0752	39.5151	<.0001*	11	-0.0251
12	0.2960	45.3553	<.0001*	12	0.0991
13	-0.1484	46.8649	<.0001*	13	0.1827
14	-0.0548	47.0769	<.0001*	14	0.1410
15	-0.0237	47.1177	<.0001*	15	0.1036
16	0.1636	49.1249	<.0001*	16	0.0079
17	-0.1420	50.6863	<.0001*	17	-0.1041
18	-0.0020	50.6867	<.0001*	18	-0.1103
19	-0.0726	51.1230	<.0001*	19	-0.2092
20	0.2717	57.4525	<.0001*	20	0.0723
21	-0.2558	63.2701	<.0001*	21	-0.0537
22	0.0512	63.5123	<.0001*	22	-0.0351
23	-0.0184	63.5447	<.0001*	23	-0.0133
24	0.1178	64.9322	<.0001*	24	-0.1782

So, we apply a seasonal ARIMA and fit the model.

5. We tried different seasonal ARIMA models starting with seasonal ARIMA (1,2,1)(0,2,0)4 and did model comparison for various models. Even though we got low AIC for seasonal ARIMA (1,2,2)(2,2,2)4, we did not achieve the required result in pre-whitening. So, we had to repeat the process and arrive at the best model. We got the best model at seasonal ARIMA(2,2,2)(2,2,2)4.

Model Comparison

Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights
		Seasonal ARIMA(1, 2, 2)(2, 2, 2)4	32	1.708e+15	1549.3845	1562.8955	0.968	1533.3845	0.111434
		Seasonal ARIMA(1, 2, 1)(2, 2, 1)4	34	2.258e+15	1550.2632	1560.3965	0.967	1538.2632	0.071814
		Seasonal ARIMA(1, 2, 2)(3, 2, 1)4	32	1.615e+15	1550.6333	1564.1443	0.968	1534.6333	0.059683
		Seasonal ARIMA(1, 2, 1)(2, 2, 2)4	33	2.011e+15	1550.7385	1562.5606	0.968	1536.7385	0.056625
		Seasonal ARIMA(1, 2, 1)(1, 2, 3)4	33	2.056e+15	1551.2091	1563.0312	0.967	1537.2091	0.044752
		Seasonal ARIMA(2, 2, 2)(2, 2, 1)4	32	1.981e+15	1551.2498	1564.7609	0.967	1535.2498	0.043850
		Seasonal ARIMA(1, 2, 1)(3, 2, 1)4	33	2.028e+15	1551.3281	1563.1502	0.967	1537.3281	0.042167
		Seasonal ARIMA(2, 2, 2)(2, 2, 2)4	31	1.666e+15	1551.7712	1566.9711	0.968	1533.7712	0.033788
		Seasonal ARIMA(2, 2, 1)(2, 2, 1)4	33	2.259e+15	1551.8285	1563.6506	0.966	1537.8285	0.032833
		Seasonal ARIMA(1, 2, 1)(1, 2, 2)4	34	1.963e+15	1551.8869	1562.0201	0.965	1539.8869	0.031889
		Seasonal ARIMA(1, 2, 3)(2, 2, 2)4	31	1.636e+15	1552.0374	1567.2373	0.968	1534.0374	0.029577
		Seasonal ARIMA(1, 2, 2)(2, 2, 1)4	33	2.343e+15	1552.5075	1564.3296	0.967	1538.5075	0.023381
		Seasonal ARIMA(2, 2, 1)(3, 2, 1)4	32	1.981e+15	1552.5194	1566.0304	0.967	1536.5194	0.023243
		Seasonal ARIMA(1, 2, 1)(2, 2, 3)4	32	2.042e+15	1552.6856	1566.1967	0.968	1536.6856	0.021389
		Seasonal ARIMA(1, 2, 1)(3, 2, 2)4	32	1.944e+15	1552.7735	1566.2846	0.968	1536.7735	0.020469
		Seasonal ARIMA(1, 2, 1)(1, 2, 1)4	35	2.359e+15	1552.9784	1561.4228	0.964	1542.9784	0.018476
		Seasonal ARIMA(1, 2, 3)(2, 2, 3)4	30	1.811e+15	1553.0400	1569.9287	0.969	1533.04	0.017916
		Seasonal ARIMA(2, 2, 2)(2, 2, 3)4	30	1.818e+15	1553.0403	1569.9291	0.969	1533.0403	0.017914
		Seasonal ARIMA(2, 2, 1)(2, 2, 2)4	32	1.863e+15	1553.2864	1566.7975	0.967	1537.2864	0.015839

6. Our model- seasonal ARIMA(2,2,2)(2,2,2)4

<u>**SARIMA(2,2,2)(2,2,2)4 applied to the output data- International calls.**</u>

Model: Seasonal ARIMA(2, 2, 2)(2, 2, 2)4

Model Summary

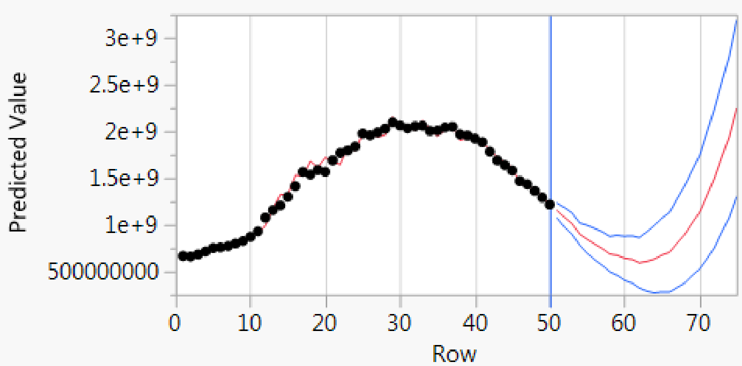
DF	31	Stable	Yes
Sum of Squared Errors	5.1632e+16	Invertible	Yes
Variance Estimate	1.6655e+15		
Standard Deviation	40811082.1		
Akaike's 'A' Information Criterion	1551.77118		
Schwarz's Bayesian Criterion	1566.9711		
RSquare	0.96798017		
RSquare Adj	0.95971698		
MAPE	2.68919294		
MAE	42813313.3		
-2LogLikelihood	1533.77118		

Parameter Estimates

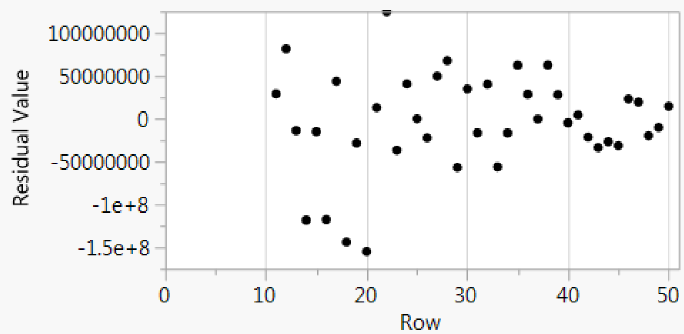
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	0.81731182	0.1803147	4.53	<.0001*
AR1,2	1	2	-0.1426866	0.1510169	-0.94	0.3520
AR2,4	2	4	-1.2	0.1235452	-10.10	<.0001*
AR2,8	2	8	-0.5102552	0.1320927	-3.86	0.0005*
MA1,1	1	1	2.0	0.1165634	16.96	<.0001*
MA1,2	1	2	-0.9988989	0.1166459	-8.56	<.0001*
MA2,4	2	4	4.96227e-5	4.2341e-5	1.17	0.2501
MA2,8	2	8	1.0	0.4691528	2.13	0.0411*
Intercept	1	0	710687.5	100272.1	7.09	<.0001*

Constant Estimate	Mu
637701.531	710687.5

Forecast



Residuals



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.0456	0.0897	0.7646	1	-0.0456
2	0.1635	1.2716	0.5295	2	0.1618
3	-0.0286	1.3088	0.7271	3	-0.0154
4	0.0148	1.3190	0.8581	4	-0.0137
5	0.0804	1.6296	0.8976	5	0.0902
6	-0.0534	1.7707	0.9395	6	-0.0500
7	-0.1349	2.6971	0.9115	7	-0.1728
8	-0.2193	5.2212	0.7337	8	-0.2228
9	-0.0731	5.5106	0.7877	9	-0.0560
10	0.0403	5.6017	0.8475	10	0.0998
11	0.1009	6.1920	0.8603	11	0.1528
12	-0.1458	7.4683	0.8252	12	-0.1391
13	0.1144	8.2829	0.8247	13	0.0948
14	-0.0652	8.5575	0.8583	14	-0.0370
15	0.0761	8.9466	0.8803	15	-0.0677
16	-0.0138	8.9599	0.9151	16	-0.0957
17	-0.0416	9.0862	0.9375	17	-0.0318
18	-0.1026	9.8898	0.9354	18	-0.0647
19	-0.0215	9.9269	0.9547	19	0.0294
20	-0.0789	10.4500	0.9593	20	-0.0902
21	-0.1579	12.6552	0.9202	21	-0.1958
22	-0.0802	13.2553	0.9258	22	-0.0821
23	0.0082	13.2620	0.9461	23	0.0828
24	0.0393	13.4244	0.9586	24	0.0060
25	0.0594	13.8191	0.9648	25	0.0660

<u>SARIMA(2,2,2)(2,2,2)4 applied to the input- mobile data</u>

Model: Seasonal ARIMA(2, 2, 2)(2, 2, 2)4

Model Summary

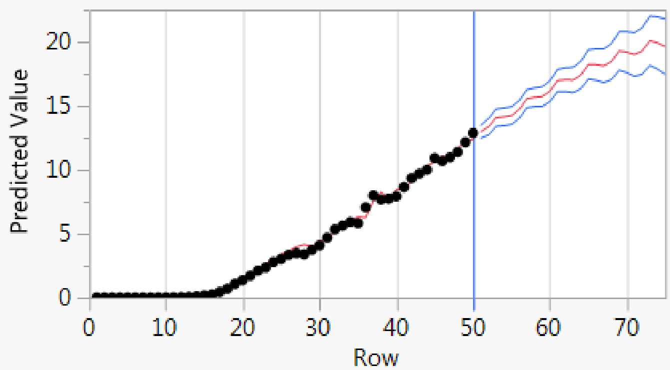
DF	31	Stable	Yes
Sum of Squared Errors	2.01686334	Invertible	Yes
Variance Estimate	0.06506011		
Standard Deviation	0.25506883		
Akaike's 'A' Information Criterion	56.0507722		
Schwarz's Bayesian Criterion	71.2506873		
RSquare	0.99355582		
RSquare Adj	0.99189281		
MAPE	9.6269167		
MAE	0.23695792		
-2LogLikelihood	38.0507722		

Parameter Estimates

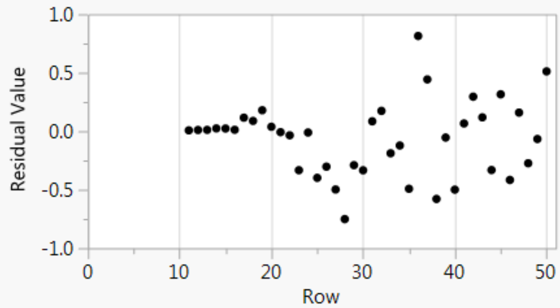
Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
AR1,1	1	1	0.765549	0.1743952	4.39	0.0001*
AR1,2	1	2	-0.228323	0.0791331	-2.89	0.0071*
AR2,4	2	4	-0.205219	0.1047321	-1.96	0.0591
AR2,8	2	8	-0.062072	0.0373120	-1.66	0.1063
MA1,1	1	1	1.966465	0.1494919	13.15	<.0001*
MA1,2	1	2	-0.998820	0.1489615	-6.71	<.0001*
MA2,4	2	4	1.981735	0.3774148	5.25	<.0001*
MA2,8	2	8	-0.981735	0.3556188	-2.76	0.0096*
Intercept	1	0	0.000028	0.0003571	0.08	0.9379

Constant Estimate	Mu
1.64579e-5	2.80627e-5

Forecast



Residuals



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.0623	0.1672	0.6826	1	0.0623
2	0.0490	0.2735	0.8722	2	0.0453
3	0.0075	0.2761	0.9645	3	0.0018
4	-0.1089	0.8294	0.9345	4	-0.1123
5	0.2747	4.4518	0.4864	5	0.2928
6	0.0479	4.5651	0.6007	6	0.0177
7	-0.1346	5.4875	0.6007	7	-0.1869
8	-0.1081	6.1014	0.6359	8	-0.1145
9	-0.1957	8.1764	0.5165	9	-0.1002
10	-0.1265	9.0728	0.5252	10	-0.1873
11	-0.0203	9.0968	0.6130	11	-0.0521
12	-0.1372	10.2265	0.5961	12	-0.0656
13	0.0617	10.4632	0.6557	13	0.1382
14	-0.0042	10.4643	0.7275	14	0.0539
15	-0.0996	11.1307	0.7433	15	-0.0616
16	-0.0545	11.3389	0.7881	16	-0.1160
17	-0.0427	11.4722	0.8310	17	-0.0216
18	0.0778	11.9344	0.8506	18	-0.0189
19	0.0248	11.9835	0.8863	19	-0.1224
20	-0.0464	12.1646	0.9103	20	-0.0686
21	0.0106	12.1745	0.9348	21	0.0841
22	-0.1273	13.6862	0.9122	22	-0.1337
23	0.0002	13.6862	0.9355	23	-0.0555
24	-0.0278	13.7674	0.9517	24	-0.0698
25	-0.0135	13.7878	0.9653	25	0.0152

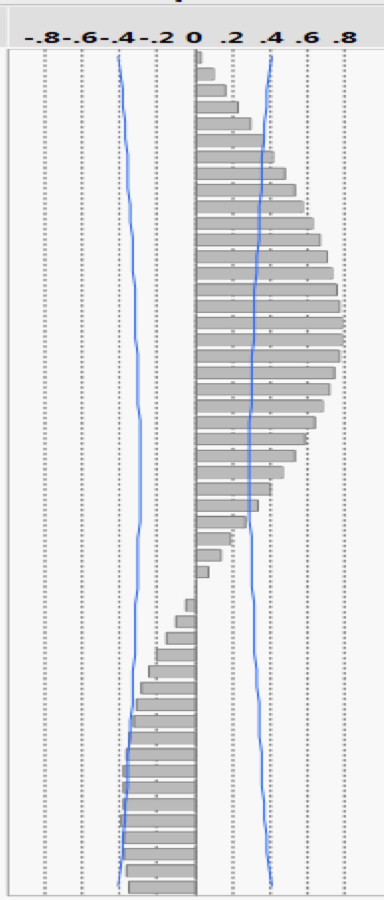
While analyzing the ACF and PACF for the input and output, we find that there is no steep/significant spike in both input and output.

Portmanteau test: On analyzing the **Ljung-Box Q** values for both input and output we find that the Portmanteau test is insignificant. Hence, we come to conclusion that the **model is adequate.**

7. **Cross-correlation-** On doing the cross correlation we get the below plot.

Cross Correlation Plots: Output Series - int calls

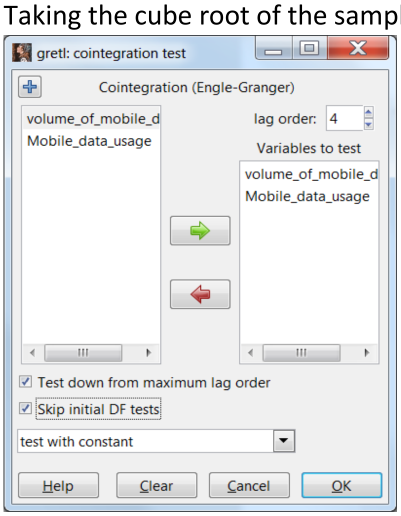
Lag	Corr vs. mobile data
-25	0.0371
-24	0.1015
-23	0.1643
-22	0.2301
-21	0.2996
-20	0.3639
-19	0.4198
-18	0.4747
-17	0.5297
-16	0.5770
-15	0.6215
-14	0.6638
-13	0.7029
-12	0.7313
-11	0.7538
-10	0.7705
-9	0.7810
-8	0.7784
-7	0.7636
-6	0.7421
-5	0.7147
-4	0.6754
-3	0.6319
-2	0.5828
-1	0.5287
0	0.4684
1	0.3996
2	0.3314
3	0.2639
4	0.1972
5	0.1326
6	0.0714
7	0.0109
8	-0.0466
9	-0.1017
10	-0.1537
11	-0.2025
12	-0.2433
13	-0.2774
14	-0.3050
15	-0.3282
16	-0.3482
17	-0.3588
18	-0.3687
19	-0.3748
20	-0.3811
21	-0.3817
22	-0.3778
23	-0.3716
24	-0.3612
25	-0.3430



We see a significant correlation in both the positive and negative side of the lags. To test if our data has **cointegration** we use the **Engle Granger cointegration test using open source software Gretl.**

Engle Granger test
Null hypothesis: H0:Unit root (i.e not cointegrated)
Alternate hypothesis: Ha: No unit root(i.e cointegrated)

Taking the cube root of the sample size we arrive at the lag order of 4.

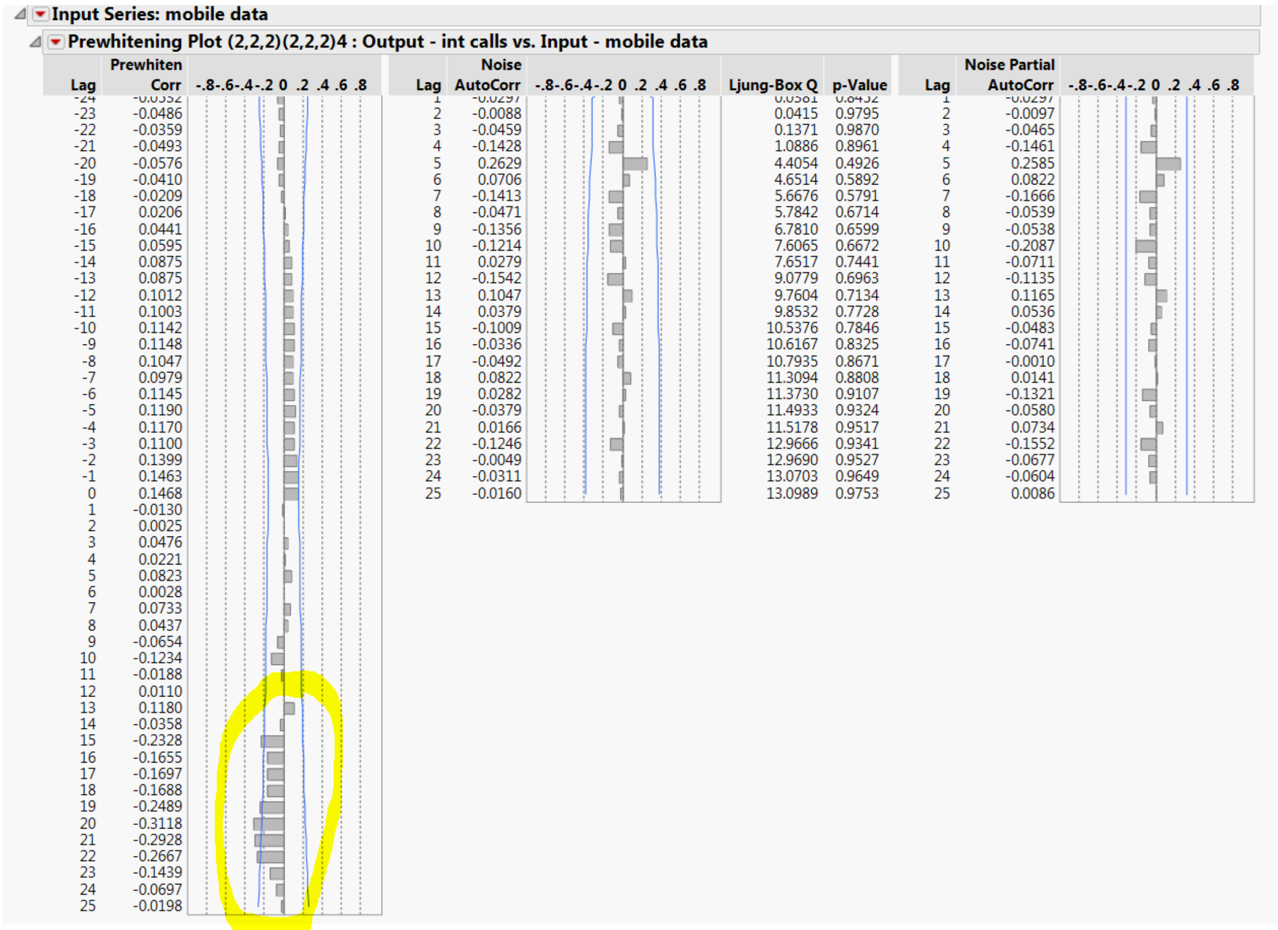


gretl: cointegration test

```
Step 1: cointegrating regression

Cointegrating regression -
OLS, using observations 2004:3-2016:4 (T = 50)
Dependent variable: volume_of_mobile_data

                     coefficient   std. error   t-ratio  p-value
  -------------------------------------------------------------
  const              -1.94404      1.73361      -1.121   0.2677
  Mobile_data_usage   4.02275e-09  1.09524e-09   3.673   0.0006  ***

Mean dependent var   4.126150   S.D. dependent var   4.146641
Sum squared resid    657.6924   S.E. of regression   3.701611
R-squared            0.219390   Adjusted R-squared   0.203128
Log-likelihood      -135.3648   Akaike criterion     274.7296
Schwarz criterion    278.5536   Hannan-Quinn         276.1858
rho                  0.995126   Durbin-Watson        0.018739

Step 2: testing for a unit root in uhat

Augmented Dickey-Fuller test for uhat
testing down from 4 lags, criterion AIC
sample size 45
unit-root null hypothesis: a = 1

  model: (1-L)y = (a-1)*y(-1) + ... + e
  estimated value of (a - 1): 0.00297437
  test statistic: tau_c(2) = 0.133987
  asymptotic p-value 0.991
  1st-order autocorrelation coeff. for e: -0.075
  lagged differences: F(4, 40) = 6.145 [0.0006]

There is evidence for a cointegrating relationship if:
(a) The unit-root hypothesis is not rejected for the individual variables, and
(b) the unit-root hypothesis is rejected for the residuals (uhat) from the
    cointegrating regression.
```

We arrive at the p value of 0.991. So, we cannot reject the null hypothesis. So, the input and output time series are not cointegrated. Hence, we proceed with pre-whitening and then applying the transfer function.

8. **Pre-whitening**- After finding the best model we try to do pre-whitening on the input data.



Input Series: mobile data

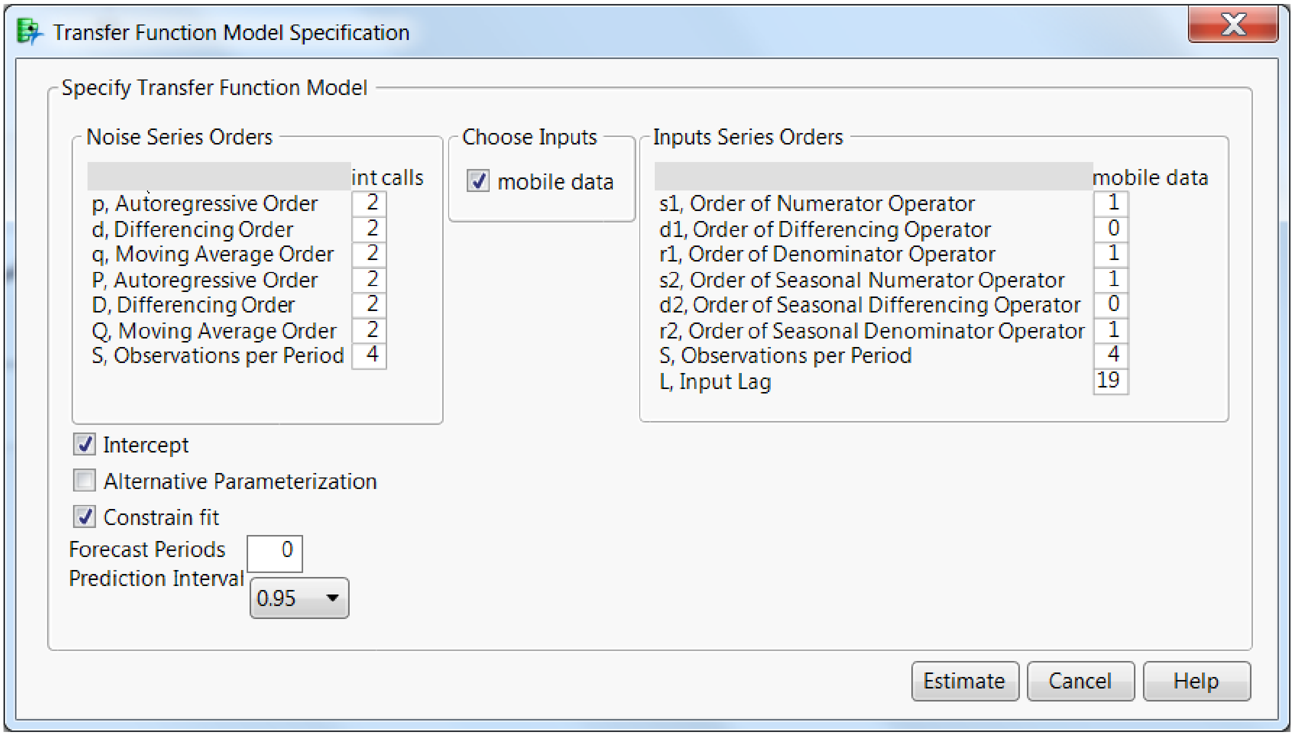
Prewhitening Plot (2,2,2)(2,2,2)4 : Output - int calls vs. Input - mobile data

Lag	Prewhiten Corr
-24	-0.0352
-23	-0.0486
-22	-0.0359
-21	-0.0493
-20	-0.0576
-19	-0.0410
-18	-0.0209
-17	0.0206
-16	0.0441
-15	0.0595
-14	0.0875
-13	0.0875
-12	0.1012
-11	0.1003
-10	0.1142
-9	0.1148
-8	0.1047
-7	0.0979
-6	0.1145
-5	0.1190
-4	0.1170
-3	0.1100
-2	0.1399
-1	0.1463
0	0.1468
1	-0.0130
2	0.0025
3	0.0476
4	0.0221
5	0.0823
6	0.0028
7	0.0733
8	0.0437
9	-0.0654
10	-0.1234
11	-0.0188
12	0.0110
13	0.1180
14	-0.0358
15	-0.2328
16	-0.1655
17	-0.1697
18	-0.1688
19	-0.2489
20	-0.3118
21	-0.2928
22	-0.2667
23	-0.1439
24	-0.0697
25	-0.0198

Lag	Noise AutoCorr	Ljung-Box Q	p-Value	Lag	Noise Partial AutoCorr
1	-0.0297	0.0381	0.8452	1	-0.0297
2	-0.0088	0.0415	0.9795	2	-0.0097
3	-0.0459	0.1371	0.9870	3	-0.0465
4	-0.1428	1.0886	0.8961	4	-0.1461
5	0.2629	4.4054	0.4926	5	0.2585
6	0.0706	4.6514	0.5892	6	0.0822
7	-0.1413	5.6676	0.5791	7	-0.1666
8	-0.0471	5.7842	0.6714	8	-0.0539
9	-0.1356	6.7810	0.6599	9	-0.0538
10	-0.1214	7.6065	0.6672	10	-0.2087
11	0.0279	7.6517	0.7441	11	-0.0711
12	-0.1542	9.0779	0.6963	12	-0.1135
13	0.1047	9.7604	0.7134	13	0.1165
14	0.0379	9.8532	0.7728	14	0.0536
15	-0.1009	10.5376	0.7846	15	-0.0483
16	-0.0336	10.6167	0.8325	16	-0.0741
17	-0.0492	10.7935	0.8671	17	-0.0010
18	0.0822	11.3094	0.8808	18	0.0141
19	0.0282	11.3730	0.9107	19	-0.1321
20	-0.0379	11.4933	0.9324	20	-0.0580
21	0.0166	11.5178	0.9517	21	0.0734
22	-0.1246	12.9666	0.9341	22	-0.1552
23	-0.0049	12.9690	0.9527	23	-0.0677
24	-0.0311	13.0703	0.9649	24	-0.0604
25	-0.0160	13.0989	0.9753	25	0.0086

Patterns in the plot suggest terms in the transfer function model. We also observe that the noise is within the significant limits indicating white noise.
The peak starts 19,peaks at 20 and ends at 22. So, first significant non-zero autocorrelation occurs at 19. So b=19. The values exhibit exponential decay after lag 20. So, s=20-19=1, r is 1 or 2. With these values we plot the transfer function.

9. **Transfer function:**We apply the transfer function with the above b,r and s values.



Transfer function model 1: r=1

Time Series int calls

Transfer Function Model (1)

Model Summary

Sum of Squared Errors	7.2809e+15
Variance Estimate	3.6405e+15
Standard Deviation	60336172.8
Akaike's 'A' Information Criterion	631.369014
Schwarz's Bayesian Criterion	642.185256
RSquare	0.98558823
RSquare Adj	0.97838235
MAPE	2.49020664
MAE	45517233.6
-2LogLikelihood	603.369024

Failed: Cannot Decrease Objective Function Hessian is not positive definite.

Parameter Estimates

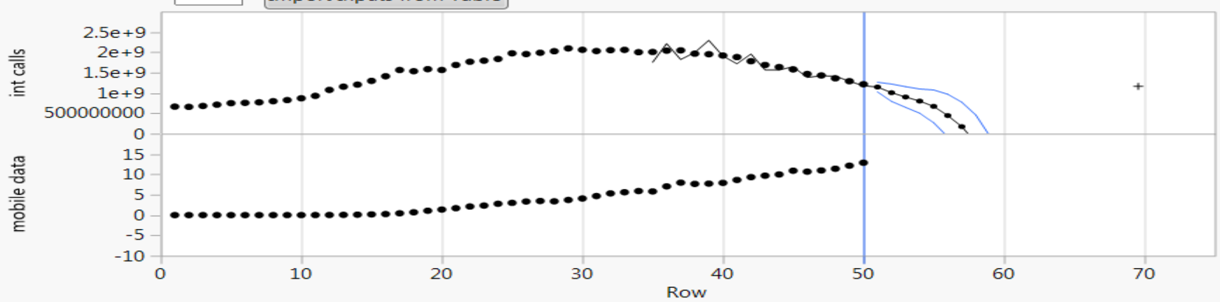
Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
mobile data	Num0,0	0	0	4.5	39.15042	0.11	0.9193
mobile data	Num1,1	1	1	-0.1046485	2.65823	-0.04	0.9722
mobile data	Num2,4	2	4	-0.3615232	9.36975	-0.04	0.9727
mobile data	Den1,1	1	1	1.9	.	.	.
mobile data	Den2,4	2	4	0.57932805	0.00051	1135.8	<.0001*
int calls	AR1,1	1	1	-0.5019074	5.3874e-5	-9316	<.0001*
int calls	AR1,2	1	2	-0.0747277	0.00044	-171.6	<.0001*
int calls	AR2,4	2	4	-1.2	.	.	.
int calls	AR2,8	2	8	-0.8744993	0.00023	-3829	<.0001*
int calls	MA1,1	1	1	7.30963e-6	1.9088e-6	3.83	0.0619
int calls	MA1,2	1	2	1.0	.	.	.
int calls	MA2,4	2	4	-0.2810013	.	.	.
int calls	MA2,8	2	8	-0.173595	0.00082	-211.3	<.0001*
	Intercept	0	0	609542.5	.	.	.

$$\left(\left(1-2\bullet B\right)+B^{2}\right)\bullet\left(\left(1-2\bullet B^{4}\right)+B^{8}\right)\bullet \text{int calls}_{t}=609542.5461+\left(\frac{\left(\left(4.4819+0.1046\bullet B\right)\bullet\left(1+0.3615\bullet B^{4}\right)\right)}{\left(\left(1-1.8817\bullet B\right)\bullet\left(1-0.5793\bullet B^{4}\right)\right)}\right)\bullet \text{mobile data}_{t-19}+\left(\frac{\left(\left(\left(1+0\bullet B\right)-B^{2}\right)\bullet\left(\left(1+0.281\bullet B^{4}\right)+0.1736\bullet B^{8}\right)\right)}{\left(\left(\left(1+0.5019\bullet B\right)+0.0747\bullet B^{2}\right)\bullet\left(\left(1+1.2392\bullet B^{4}\right)+0.8745\bullet B^{8}\right)\right)}\right)\bullet e_{t}$$

▷ **Residuals**

◿ **Interactive Forecasting**

Prediction Interval 0.95 Import Inputs from Table



◿ **Iteration History**

Iter	Iteration History	Step	Delta-Criterion	Obj-Criterion
149	603.39026269	BFGS	.	0.00841875
150	603.39023528	BFGS	.	0.01245607
151	603.39022961	BFGS	.	0.01113282
152	603.3902124	BFGS	.	0.00848455
153	603.39019923	BFGS	.	0.00459073
154	603.39018829	BFGS	.	0.00121133
155	603.39018593	BFGS	.	0.00103775
156	603.39016268	BFGS	.	0.00476796
157	603.39010162	BFGS	.	0.01180114
158	603.38887624	BFGS	.	0.04746038
159	603.38826583	BFGS	.	0.02531639
160	603.38818795	BFGS	.	0.02022978
161	603.38813298	BFGS	.	0.04423416
162	603.38750692	BFGS	.	0.02521044
163	603.38724895	BFGS	.	0.03477343
164	603.38653858	BFGS	.	0.01979129
165	603.38305283	BFGS	.	0.05454874
166	603.38192712	BFGS	.	0.07397987
167	603.38145461	BFGS	.	0.03686729
168	603.37879015	BFGS	.	0.04645923
169	603.37795861	BFGS	.	0.07506168

Transfer function model 2: r=2

ime Series int calls

Transfer Function Model (2)

◿ **Model Summary**

DF	0
Sum of Squared Errors	1.4688e+15
Variance Estimate	.
Standard Deviation	.
Akaike's 'A' Information Criterion	628.345239
Schwarz's Bayesian Criterion	640.706658
RSquare	0.99709275
RSquare Adj	0.99527572
MAPE	7.52620212
MAE	133517489
-2LogLikelihood	596.345245

Failed: Maximum Iterations Exceeded Hessian is not positive definite.

◿ **Parameter Estimates**

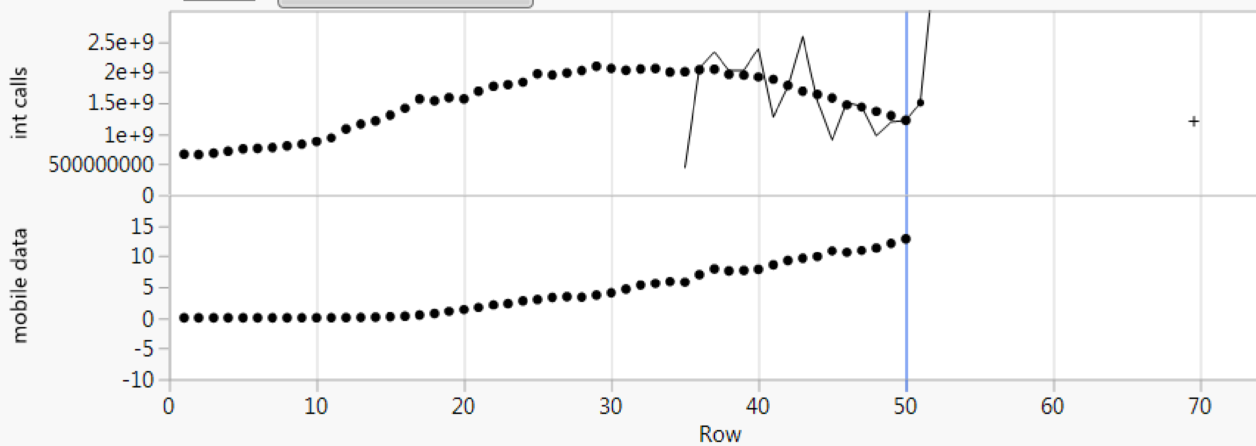
Variable	Term	Factor	Lag	Estimate	Std Error	t Ratio	Prob>|t|
mobile data	Num0,0	0	0	0.3723895	0.027431	13.58	-2.000*
mobile data	Num1,1	1	1	1.2	0.084469	13.72	-2.000*
mobile data	Num2,4	2	4	19.1	2.327461	8.22	-2.000*
mobile data	Den1,1	1	1	4.8	2.797179	1.72	-2.000*
mobile data	Den1,2	1	2	4.8	6.169937	0.77	-2.000*
mobile data	Den2,4	2	4	-0.930807	.	.	.
mobile data	Den2,8	2	8	0.31944365	.	.	.
int calls	AR1,1	1	1	-0.3609078	0.000210	-1715	-2.000*
int calls	AR1,2	1	2	-0.2279949	0.000491	-464.2	-2.000*
int calls	AR2,4	2	4	-1.4	0.000206	-6754	-2.000*
int calls	AR2,8	2	8	-0.9116142	0.000142	-6435	-2.000*
int calls	MA1,1	1	1	7.50931e-9	.	.	.
int calls	MA1,2	1	2	1.0	.	.	.
int calls	MA2,4	2	4	0.01237895	0.000076	162.61	-2.000*
int calls	MA2,8	2	8	-1.0	.	.	.
	Intercept	0	0	609542.7	0.000016	4e+10	-2.000*

$$\left(\left(1-2\bullet B\right)+B^{2}\right)\bullet\left(\left(1-2\bullet B^{4}\right)+B^{8}\right)\bullet \text{int calls}_{t}=609542.689+\left(\frac{\left(\left(0.3724-1.1592\bullet B\right)\bullet\left(1-19.1365\bullet B^{4}\right)\right)}{\left(\left(\left(1-4.8022\bullet B\right)-4.7737\bullet B^{2}\right)\bullet\left(\left(1+0.9308\bullet B^{4}\right)-0.3194\bullet B^{8}\right)\right)}\right)\bullet \text{mobile data}_{t-19}+\left(\frac{\left(\left(\left(1+0\bullet B\right)-B^{2}\right)\bullet\left(\left(1-0.0124\bullet B^{4}\right)+B^{8}\right)\right)}{\left(\left(\left(1+0.3609\bullet B\right)+0.228\bullet B^{2}\right)\bullet\left(\left(1+1.3927\bullet B^{4}\right)+0.9116\bullet B^{8}\right)\right)}\right)\bullet e_{t}$$

Residuals

Interactive Forecasting

Prediction Interval 0.95 Import Inputs from Table



Iteration History

Iter	Iteration History	Step	Delta-Criterion	Obj-Criterion
239	596.35097483	BFGS	.	58.5568976
240	596.34943743	BFGS	.	73.9100994
241	596.34793627	BFGS	.	38.7286612
242	596.34765365	BFGS	.	98.2680111
243	596.34637328	BFGS	.	108.33457
244	596.34597535	BFGS	.	62.9728208
245	596.34573139	BFGS	.	251.992696
246	596.34562064	BFGS	.	72.4438209
247	596.34528782	BFGS	.	53.0637109
248	596.34523873	BFGS	.	0.00584799
249	596.34523871	BFGS	.	1.16620959

On comparing both models:

Time Series Basic Diagnostics

Model Comparison

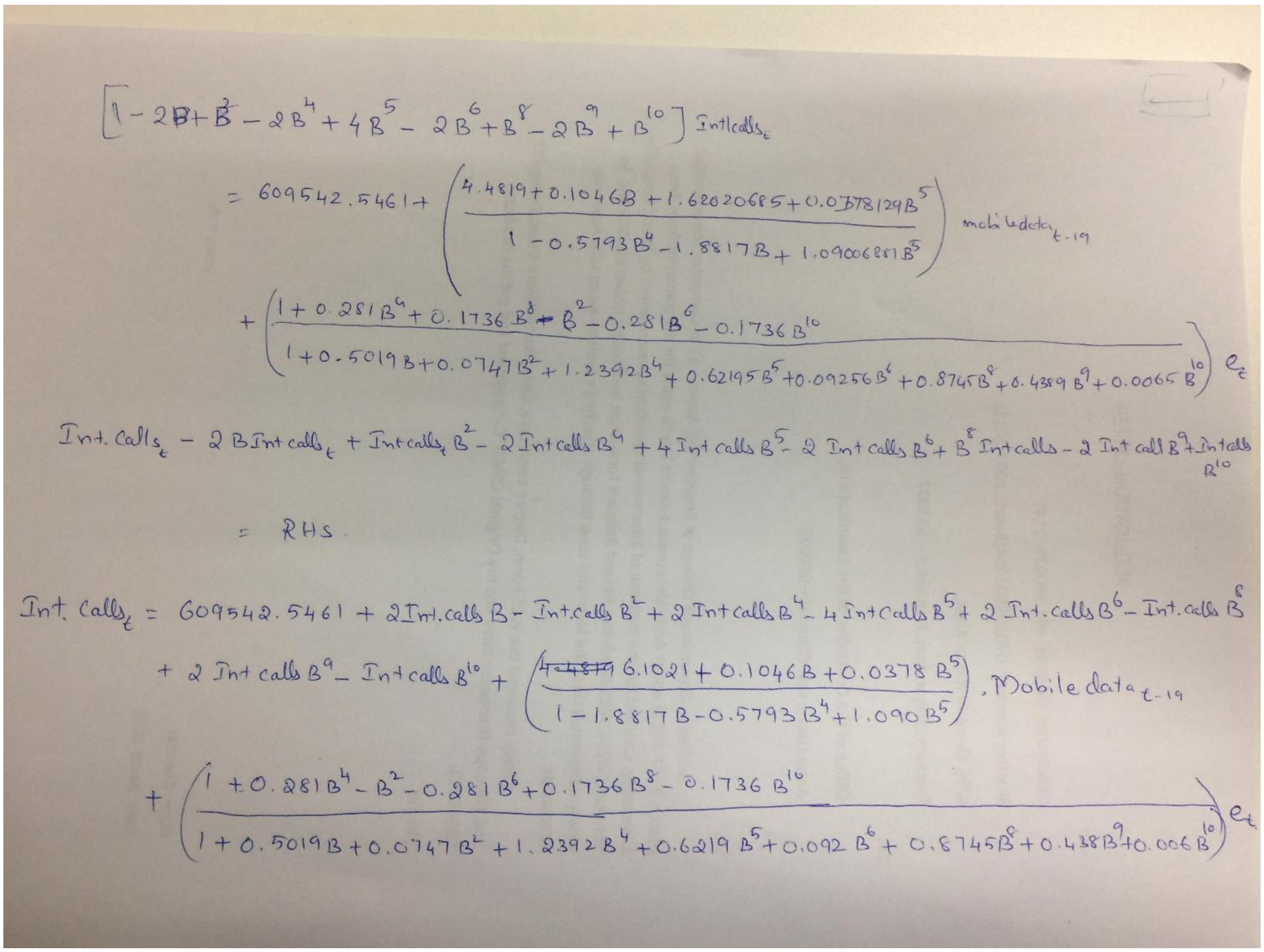
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	MAE
☑	☐	Transfer Function Model (2)	0	.	628.34524	640.70666	0.997	596.34524	0.819341		7.526202	133517489
☑	☐	Transfer Function Model (1)	2	3.64e+15	631.36901	642.18526	0.986	603.36901	0.180659		2.490207	45517234

Even though the transfer function 2 has lower AIC, it has higher MAPE. So, we choose **transfer function 1** with AIC of 631.36 and MAPE of 2.49 as the best model.

10. Expanded Transfer Function Equation

$$
\begin{aligned}
\text{IntCalls}_t &= 1 + 609542.546 \; + \; 2\,\text{IntCalls}_t\text{B} \; - \; \text{IntCalls}_t B^2 \; + \; 2\,\text{IntCalls}_t B^4 - 4\,\text{IntCalls}_t B^5 \; + \; 2\,\text{IntCalls}_t B^6 \\
&- \; \text{IntCalls}_t B^8 \; + \; 2\,\text{IntCalls}_t B^9 \; - \; \text{IntCalls}_t B^{10} \; + \left(\frac{6.1021 + 0.1046\,B + 0.0378\,B^5}{1 - 1.8817\,B - 0.5793\,B^4 + 1.090\,B^5}\right) MobileData \\
&+ \left(\frac{1 - B^2 + \; 0.281\,B^4 - 0.281\,B^6 + 0.1736\,B^8 - 0.1736\,B^{10}}{1 + 0.5019\,B + 0.0747\,B^2 + 1.2392\,B^4 + 0.6219\,B^5 + 0.092\,B^6 + 0.8745\,B^8 + 0.438\,B^9 + 0.006\,B^{10}}\right) e_t
\end{aligned}
$$

Below the derivation steps for solving this equation.



$$[1 - 2B + B^2 - 2B^4 + 4B^5 - 2B^6 + B^8 - 2B^9 + B^{10}]\ \text{Intcalls}_t$$

$$= 609542.5461 + \left(\frac{4.4819 + 0.1046B + 1.62020685 + 0.0378129B^5}{1 - 0.5793B^4 - 1.8817B + 1.09006881B^5}\right) \text{mobiledata}_{t-19}$$

$$+ \left(\frac{1 + 0.281B^4 + 0.1736B^8 - B^2 - 0.281B^6 - 0.1736B^{10}}{1 + 0.5019B + 0.0747B^2 + 1.2392B^4 + 0.62195B^5 + 0.09256B^6 + 0.8745B^8 + 0.4389B^9 + 0.0065B^{10}}\right) e_t$$

$$\text{Int.Calls}_t - 2B\,\text{Intcalls}_t + \text{Intcalls}_t B^2 - 2\,\text{Intcalls}\,B^4 + 4\,\text{Intcalls}\,B^5 - 2\,\text{Intcalls}\,B^6 + B^8\,\text{Intcalls} - 2\,\text{Intcall}\,B^9 + \text{Intcalls}\,B^{10}$$

$$= \text{RHS}.$$

$$\text{Int.Calls}_t = 609542.5461 + 2\,\text{Int.calls}\,B - \text{Int.calls}\,B^2 + 2\,\text{Intcalls}\,B^4 - 4\,\text{IntCalls}\,B^5 + 2\,\text{Int.calls}\,B^6 - \text{Int.calls}\,B^8$$

$$+ 2\,\text{Intcalls}\,B^9 - \text{Intcalls}\,B^{10} + \left(\frac{\cancel{4.4819}\ 6.1021 + 0.1046B + 0.0378B^5}{1 - 1.8817B - 0.5793B^4 + 1.090B^5}\right) \cdot \text{Mobiledata}_{t-19}$$

$$+ \left(\frac{1 + 0.281B^4 - B^2 - 0.281B^6 + 0.1736B^8 - 0.1736B^{10}}{1 + 0.5019B + 0.0747B^2 + 1.2392B^4 + 0.6219B^5 + 0.092B^6 + 0.8745B^8 + 0.438B^9 + 0.006B^{10}}\right) e_t$$

Conclusion: Our final transfer function has an AIC of 631.36 and MAPE of 2.49. Hence, we arrive at the conclusion that mobile data indeed affects the international calls as shown above in the equation.