# Doing Data Science: Case Study 2 - Forecasting Chulwalar Total Etel Exports

### Claire Chu & Joseph Stoffa

### July 19, 2016

## Introduction

Chulwalar is part of the island group Urbano in the northern hemisphere. They are famous for their plants which flower in winter. There are three main plants that Chulwalar exports: Efak is a leafy bush with white flowers, Wuge is a grass like plant with tiny pink flowers and Etel is a flowering tree. Etel comes in two varieties: red flowers and blue flowers. Due to the nature of the products, exports generally are higher towards the end of the year. Chulwalar celebrates its independence on the 1st of December each year. On this day it is custom to give presents to family and friends. Chulwalar also celebrates the March Equinox as a time of rebirth in the northern hemisphere.

The Prime Minister of Chulwalar has asked us to help him in forecasting the exports. In order to do this we have been given as is data and plan data as well as a list of indicators which may affect exports. Our job is to find out the best way to forecast Chulwalar's exports in 2014 based on data collected before this year

## Background

This case study builds on the data analysis and modeling of plant exports from Chulwalar performed by .

## Table of Contents

#### 1.0 [Importing Data](#id-section1)

#### 2.0 [Cleaning Modifying Data](#id-section2)

#### 3.0 [Exploring Data](#id-section3)

#### ---3.1 [Preliminary Analysis](#id-section3.1)

#### ---3.2 [Preliminary Time Series Analysis](#id-section3.2)

#### 4.0 [Correlation of Different External Indicators](#id-section4)

## 1.0 Importing Data

The script *loadData.r* loads the the data for this case study into dataframes *ImportedAsISData*, *ImportedIndicators*, and *ImportedPlanData* from the files *ImportedAsIsDataChulwalar.csv*, *ImportedIndicatorshulwalar.csv*, and *ImportedPlanDataChulwalar.csv* respectively.

## [1] CSV files successfully read. Dataframes loaded:   
## [1] ImportedAsIsData ImportedIndicators ImportedPlanData

## 2.0 Cleaning/Modifying Data

In order to conduct our analysis the data needs to be subsetted into vectors and time series. For this case study we are only interested in the Total Etel exports. Data on other individual plants will be ignored. The script *cleanData.r* takes the data frames for As Is, Planned, and Indicator data and splits them into vectors which are then converted into time series. Below are the resultant time series outputted by the script.

source("Scripts/cleanData.r")

## [1] TotalAsIs:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2008 2313221 1950131 2346635 2039787 1756964 1458302 1679637 1639670  
## 2009 2610573 2371327 2743786 2125308 1850073 1836222 1797311 1851968  
## 2010 2760688 2918333 3227041 1613888 2550157 2317645 1474144 2148521  
## 2011 3112861 2926663 3294784 2577079 2774068 2378227 2222900 2991787  
## 2012 3093088 3679308 3433364 2714899 3011767 2726028 2483834 3055655  
## 2013 4119526 3535744 3560974 3760065 2959933 2787898 2828744 3084113  
## Sep Oct Nov Dec  
## 2008 2882886 2959716 2596494 2656568  
## 2009 3271171 2818888 3310776 3022513  
## 2010 3898571 3348953 3135945 3332886  
## 2011 4151531 3318684 4037076 3429843  
## 2012 4200796 4228724 4618540 3383673  
## 2013 5107775 4562144 4729313 4372181  
## [1]   
## [1] TotalEtelAsIs:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2008 1279668 1053325 1367520 1090725 873568 644479 772658 806741  
## 2009 1583216 1407388 1420801 1141100 919860 858876 910134 843050  
## 2010 1637464 1676161 1549560 813469 1198401 1140024 551268 1012542  
## 2011 1595267 1473528 1469728 1034650 952553 819303 802076 1222812  
## 2012 1519748 1812897 1607280 1008022 1291983 940158 945929 1235146  
## 2013 2109497 1738197 1633944 1745092 1039449 1054201 1003166 1154675  
## Sep Oct Nov Dec  
## 2008 1715265 1795751 1518288 1601324  
## 2009 1981563 1647934 1857836 1615091  
## 2010 2335488 1856264 1678123 1699063  
## 2011 2303271 1591584 1960675 1713991  
## 2012 2330334 2177895 2306324 1618147  
## 2013 3000929 2305605 2284672 2062160  
## [1]   
## [1] YearAsIs:   
## Jan Feb Mar Apr May Jun Jul  
## 2008 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## 2009 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## 2010 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## 2011 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## 2012 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## 2013 26280011 29609916 32726772 37215503 40629676 45408410 26280011  
## Aug Sep Oct Nov Dec  
## 2008 29609916 32726772 37215503 40629676 45408410  
## 2009 29609916 32726772 37215503 40629676 45408410  
## 2010 29609916 32726772 37215503 40629676 45408410  
## 2011 29609916 32726772 37215503 40629676 45408410  
## 2012 29609916 32726772 37215503 40629676 45408410  
## 2013 29609916 32726772 37215503 40629676 45408410  
## [1]   
## [1] TotalAsIs\_2014:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2014 4308161 4155378 3924332 3659121 3898758 3313891 3595106 3502426  
## Sep Oct Nov Dec  
## 2014 5619059 5274287 4841693 4664854  
## [1]   
## [1] TotalPlan:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2008 2243103 2162705 2720911 2011182 1877757 1819924 1682196 1893171  
## 2009 2547980 2247049 2731156 2020158 2098038 1927995 1783692 1907705  
## 2010 2965885 2751170 2906493 2383358 2246893 1992851 2023434 2244997  
## 2011 3113110 2883766 2957893 2601648 2370949 2339881 2105328 2341623  
## 2012 3895396 3588151 3787240 3036434 2907891 2707822 2619486 3784557  
## 2013 3580325 3863212 3606083 3213575 3139128 2998610 2785453 3083654  
## Sep Oct Nov Dec  
## 2008 3325711 2662148 2909966 2574633  
## 2009 3124040 3102251 3154669 2742367  
## 2010 3257717 3536338 3358206 3112906  
## 2011 4086297 3640827 3502334 3280476  
## 2012 4987460 4367319 4205772 4059533  
## 2013 5143757 4149334 4495212 4093664  
## [1]   
## [1] TotalEtelPlan:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2008 1263613 1231125 1489621 1051346 933392 932047 855520 923070  
## 2009 1546801 1378217 1563799 1166229 1057223 983279 913751 980703  
## 2010 1648769 1490577 1538493 1208636 1104777 931127 916160 1096933  
## 2011 1781991 1564272 1455531 1257528 1134418 1018200 843336 974375  
## 2012 2070256 1731099 1663266 1232994 1164076 1018137 932241 1800576  
## 2013 1864733 1837228 1663834 1305603 1172373 1089115 1074687 1217930  
## Sep Oct Nov Dec  
## 2008 2080877 1575579 1561956 1515127  
## 2009 1974166 1886971 1839155 1727567  
## 2010 1832882 2103588 1877929 1862684  
## 2011 2435674 1972649 1873075 1684766  
## 2012 2823873 2224655 2025003 1955509  
## 2013 2916115 2043888 2199880 2133214  
## [1]   
## [1] YearPlan:   
## Jan Feb Mar Apr May Jun Jul  
## 2008 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## 2009 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## 2010 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## 2011 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## 2012 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## 2013 27883407 29387100 32780247 35224132 43947063 44152007 27883407  
## Aug Sep Oct Nov Dec  
## 2008 29387100 32780247 35224132 43947063 44152007  
## 2009 29387100 32780247 35224132 43947063 44152007  
## 2010 29387100 32780247 35224132 43947063 44152007  
## 2011 29387100 32780247 35224132 43947063 44152007  
## 2012 29387100 32780247 35224132 43947063 44152007  
## 2013 29387100 32780247 35224132 43947063 44152007  
## [1]   
## [1] TotalPlan\_2014:   
## Jan Feb Mar Apr May Jun Jul Aug  
## 2014 4474000 4185565 4278119 3985542 3605973 3515173 3269444 3656112  
## Sep Oct Nov Dec  
## 2014 5637391 5157781 5353458 4703185

## 3.0 Exploring Data

### 3.1 Preliminary Analysis

For the preliminary Analysis, the time series have been plotted and the correlation between the As Is and Plan data has been tested in order to ascertain how exact the planning is. These results show a very high planning accuracy and are therefore suitable for modeling. (Wheeler et al.)

source("Scripts/preliminaryAnalysis.r")

## [1] Correlation between Total As Is and Total Plan:   
## [1] 0.9183402  
## [1] Correlation between Total Etel As Is and Total Etel Plan:   
## [1] 0.9159505  
## [1] Correlation between Year As Is and Year Plan:   
## [1] 0.9627401  
##   
## Call:  
## lm(formula = TotalAsIs ~ TotalPlan, data = TotalAsIs)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -770214 -196776 26017 182579 672705   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.959e+04 1.521e+05 0.589 0.558   
## TotalPlan 9.627e-01 4.959e-02 19.413 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 332600 on 70 degrees of freedom  
## Multiple R-squared: 0.8433, Adjusted R-squared: 0.8411   
## F-statistic: 376.9 on 1 and 70 DF, p-value: < 2.2e-16  
##   
##   
## Call:  
## tslm(formula = TotalAsIs ~ TotalPlan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -770214 -196776 26017 182579 672705   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.959e+04 1.521e+05 0.589 0.558   
## TotalPlan 9.627e-01 4.959e-02 19.413 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 332600 on 70 degrees of freedom  
## Multiple R-squared: 0.8433, Adjusted R-squared: 0.8411   
## F-statistic: 376.9 on 1 and 70 DF, p-value: < 2.2e-16

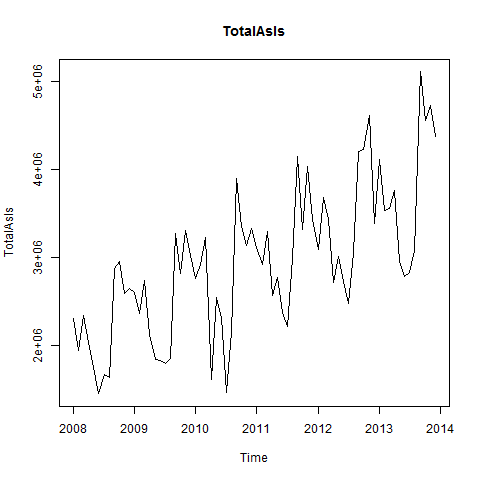


Figure 1

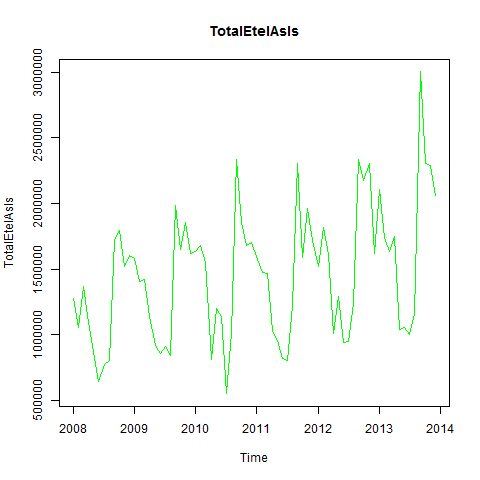


Figure 2

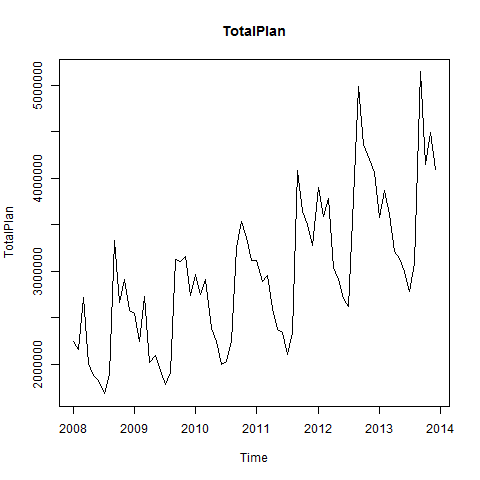


Figure 3

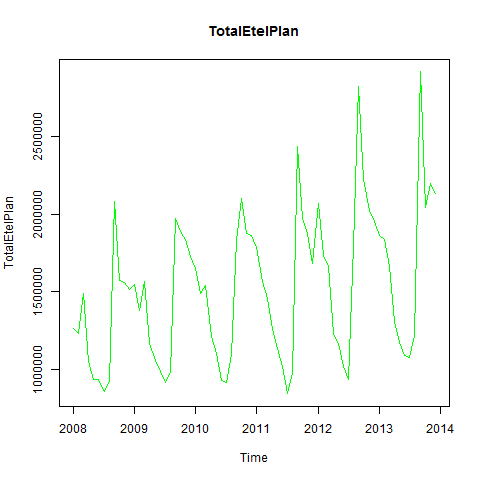


Figure 4

### 3.2 Preliminary Time Series Analysis

The time series can be analyzed using the *stl* function in order to seperate the trend, seasonality, and remainder (remaining) coincidental) components from one another. Thus the individual time series can be shown graphically and tabularly. The trend of the total exports is almost linear. A relatively uniform seasonality can be seen in Figure 5. (Wheeler et al).

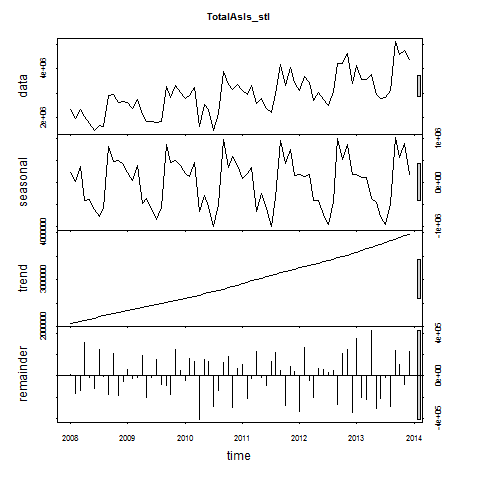


Figure 5

print(TotalAsIs\_stl)

## Call:  
## stl(x = TotalAsIs, s.window = 5)  
##   
## Components  
## seasonal trend remainder  
## Jan 2008 223320.67 2074233 15667.16  
## Feb 2008 17036.99 2096208 -163113.80  
## Mar 2008 361473.74 2118182 -133021.18  
## Apr 2008 -410834.24 2140157 310464.16  
## May 2008 -391831.93 2162114 -13317.80  
## Jun 2008 -608564.13 2184070 -117204.25  
## Jul 2008 -777993.52 2206027 251603.49  
## Aug 2008 -583615.66 2228213 -4927.72  
## Sep 2008 810939.36 2250400 -178453.09  
## Oct 2008 474131.86 2272586 212998.05  
## Nov 2008 488504.52 2294373 -186383.79  
## Dec 2008 395452.58 2316160 -55045.03  
## Jan 2009 217151.38 2337948 55473.99  
## Feb 2009 39716.91 2359168 -27558.10  
## Mar 2009 378507.21 2380389 -15109.96  
## Apr 2009 -467522.18 2401609 191220.87  
## May 2009 -371597.89 2425515 -203844.26  
## Jun 2009 -595724.45 2449421 -17474.54  
## Jul 2009 -827029.12 2473327 151013.28  
## Aug 2009 -567342.69 2495885 -76573.99  
## Sep 2009 843160.68 2518443 -90432.21  
## Oct 2009 447562.71 2541000 -169675.09  
## Nov 2009 497312.47 2562364 251099.75  
## Dec 2009 388265.67 2583727 50520.14  
## Jan 2010 201133.54 2605091 -45536.12  
## Feb 2010 122776.46 2628120 167436.40  
## Mar 2010 442825.47 2651150 133065.83  
## Apr 2010 -652923.75 2674179 -407367.50  
## May 2010 -301149.68 2698691 152615.46  
## Jun 2010 -543850.29 2723203 138292.09  
## Jul 2010 -985987.99 2747715 -287583.18  
## Aug 2010 -487941.31 2774544 -138081.68  
## Sep 2010 972415.73 2801373 124782.46  
## Oct 2010 343206.82 2828202 177544.55  
## Nov 2010 573281.74 2858572 -295909.05  
## Dec 2010 375326.75 2888943 68616.25  
## Jan 2011 84179.43 2919314 109367.89  
## Feb 2011 190940.11 2949475 -213752.60  
## Mar 2011 339598.68 2979637 -24451.98  
## Apr 2011 -661193.66 3009799 228473.57  
## May 2011 -252299.73 3037669 -11300.88  
## Jun 2011 -597799.74 3065538 -89511.39  
## Jul 2011 -1002974.31 3093408 132466.66  
## Aug 2011 -345401.48 3120526 216662.97  
## Sep 2011 951339.44 3147643 52548.18  
## Oct 2011 418464.54 3174761 -274541.80  
## Nov 2011 749466.48 3200972 86637.11  
## Dec 2011 166063.96 3227184 36595.48  
## Jan 2012 173825.10 3253395 -334131.81  
## Feb 2012 131526.89 3279250 268531.13  
## Mar 2012 171949.25 3305105 -43690.50  
## Apr 2012 -412193.90 3330961 -203867.63  
## May 2012 -414897.17 3358540 68124.29  
## Jun 2012 -723606.43 3386119 63515.20  
## Jul 2012 -957183.71 3413699 27319.12  
## Aug 2012 -438041.15 3441507 52189.27  
## Sep 2012 998725.79 3469315 -267244.98  
## Oct 2012 523934.85 3497123 207665.66  
## Nov 2012 847979.72 3527674 242886.44  
## Dec 2012 172550.29 3558224 -347101.49  
## Jan 2013 184195.89 3588775 346555.55  
## Feb 2013 114297.14 3623803 -202355.91  
## Mar 2013 121000.80 3658831 -218857.78  
## Apr 2013 -360531.42 3693859 426737.22  
## May 2013 -462506.26 3728897 -306457.92  
## Jun 2013 -759940.89 3763935 -216096.28  
## Jul 2013 -951772.71 3798973 -18456.45  
## Aug 2013 -468011.67 3834192 -282067.53  
## Sep 2013 1004335.28 3869411 234028.47  
## Oct 2013 554713.70 3904630 102800.01  
## Nov 2013 873598.66 3940742 -85027.53  
## Dec 2013 169104.03 3976853 226223.51

It is interesting to note that the almost linear trend is not seen in the individual segment for Total Etel (Figure 6). The individual trends for all flower exports run in partially opposite directions in the middle of the time scale, which causes the total As Is data trend to be linear. (Wheeler et al).

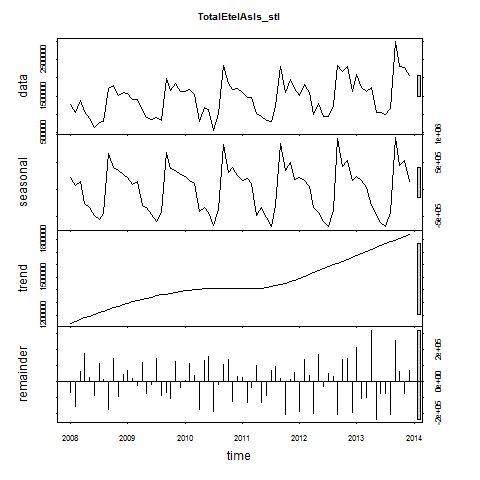


Figure 6

print(TotalEtelAsIs\_stl)

## Call:  
## stl(x = TotalEtelAsIs, s.window = 5)  
##   
## Components  
## seasonal trend remainder  
## Jan 2008 212543.90 1135393 -68268.686  
## Feb 2008 64724.69 1149373 -160772.697  
## Mar 2008 141647.42 1163353 62519.359  
## Apr 2008 -264999.95 1177333 178391.504  
## May 2008 -339776.29 1191010 22334.080  
## Jun 2008 -473456.92 1204687 -86751.057  
## Jul 2008 -560479.22 1218364 114773.472  
## Aug 2008 -439963.23 1231709 14995.207  
## Sep 2008 649360.43 1245054 -179149.730  
## Oct 2008 394715.11 1258400 142636.314  
## Nov 2008 343804.99 1270718 -96234.645  
## Dec 2008 272635.02 1283036 45653.230  
## Jan 2009 216526.66 1295354 71335.502  
## Feb 2009 85366.77 1304840 17181.430  
## Mar 2009 134597.06 1314326 -28121.833  
## Apr 2009 -299576.10 1323812 116864.362  
## May 2009 -339425.04 1333273 -73988.443  
## Jun 2009 -463451.32 1342735 -20407.903  
## Jul 2009 -588086.05 1352197 146023.090  
## Aug 2009 -428334.97 1359570 -88185.300  
## Sep 2009 684000.28 1366944 -69380.862  
## Oct 2009 378107.74 1374317 -104490.629  
## Nov 2009 350361.28 1380061 127414.131  
## Dec 2009 269269.49 1385804 -39982.767  
## Jan 2010 237811.58 1391548 8104.442  
## Feb 2010 164697.76 1396069 115393.780  
## Mar 2010 108887.92 1400591 40081.142  
## Apr 2010 -414588.58 1405112 -177054.835  
## May 2010 -341339.97 1406966 132774.500  
## Jun 2010 -426196.92 1408821 157400.399  
## Jul 2010 -672372.03 1410675 -187034.545  
## Aug 2010 -377863.24 1410092 -19686.420  
## Sep 2010 818966.77 1409509 107012.482  
## Oct 2010 310866.57 1408926 136471.594  
## Nov 2010 396380.37 1408474 -126731.546  
## Dec 2010 260838.84 1408023 30201.636  
## Jan 2011 162908.25 1407571 24787.879  
## Feb 2011 200345.42 1408473 -135290.178  
## Mar 2011 98058.73 1409375 -37705.374  
## Apr 2011 -475991.41 1410277 100364.893  
## May 2011 -326861.16 1414388 -134973.648  
## Jun 2011 -508609.13 1418499 -90586.968  
## Jul 2011 -690240.90 1422610 69706.513  
## Aug 2011 -303950.41 1431899 94863.512  
## Sep 2011 841755.54 1441187 20328.047  
## Oct 2011 348453.95 1450476 -207345.878  
## Nov 2011 482953.99 1463581 14139.587  
## Dec 2011 177522.68 1476687 59781.396  
## Jan 2012 219295.41 1489792 -189339.836  
## Feb 2012 167343.82 1504763 140790.260  
## Mar 2012 47565.92 1519733 39980.664  
## Apr 2012 -327162.02 1534704 -199519.887  
## May 2012 -429104.97 1550797 170291.408  
## Jun 2012 -592640.70 1566889 -34090.529  
## Jul 2012 -686742.76 1582982 49689.871  
## Aug 2012 -391971.46 1597629 29488.941  
## Sep 2012 926501.74 1612275 -208442.888  
## Oct 2012 411978.67 1626922 138994.561  
## Nov 2012 520321.23 1641216 144786.748  
## Dec 2012 157545.15 1655510 -194908.427  
## Jan 2013 228279.54 1669805 211412.939  
## Feb 2013 159093.45 1685663 -106559.227  
## Mar 2013 32529.79 1701521 -100106.827  
## Apr 2013 -295207.15 1717379 322919.852  
## May 2013 -454478.79 1733056 -239128.325  
## Jun 2013 -615888.66 1748733 -78643.269  
## Jul 2013 -688356.21 1764410 -72887.536  
## Aug 2013 -417329.54 1779986 -207981.755  
## Sep 2013 944212.59 1795563 261153.561  
## Oct 2013 431257.90 1811139 63207.703  
## Nov 2013 530407.87 1827035 -72771.039  
## Dec 2013 148706.84 1842931 70522.228

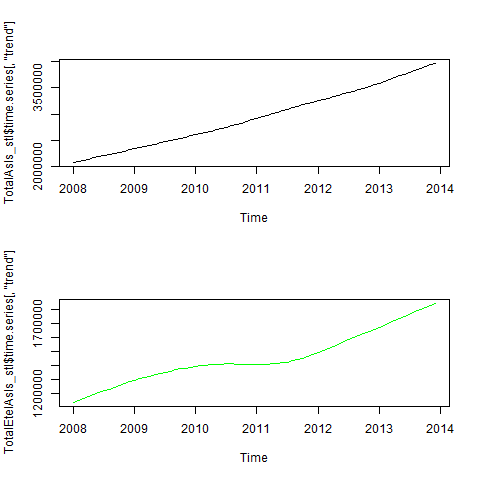


Figure 7

The modification of the seasonality component can also be changed into a montly view. It only makes sense to do this if the seasonality component as the trend looks almost identical and the remainder is then randomly spread. (Wheeler et al.)

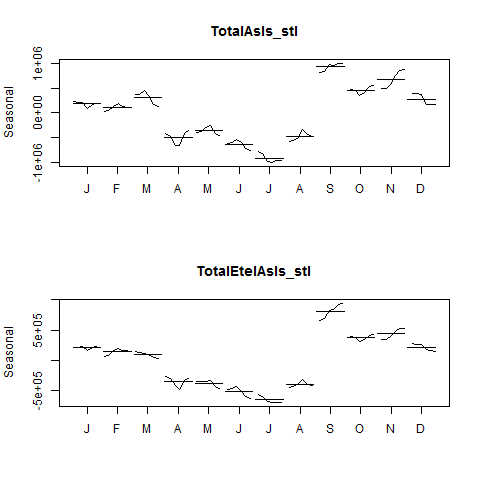


Figure 8

## 4.0 Correlation of Different External Indicators