A closer look at SAP PAL - If you're not looking for inliers, then you can't see them. You need Autobox glasses to do that!

Posted on Jun 13, Posted by Tom Reilly Category Forecasting

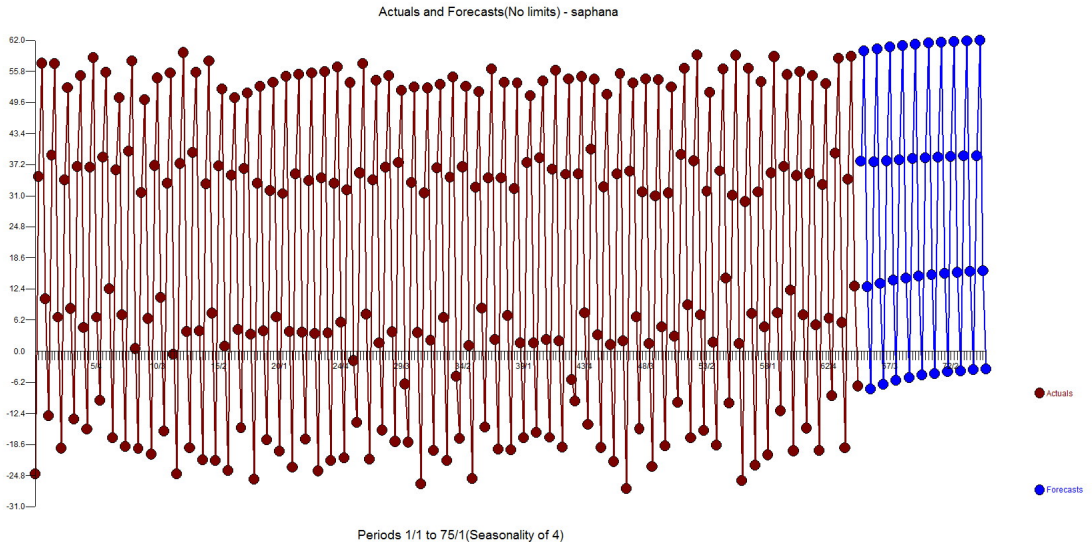
SAP HANA is a Database. PAL is SAP's modeling tool. SAP's naming conventions are a bit confusing. SAP HANA has a very comprehensive user's guide that not only shows an example model of their "Auto Seasonal ARIMA" model on page 349, but also includes the data
which allows us to benchmark against. The bottom line is that the ARIMA model shown is highly overparameterized, ignores outliers, changes in level and a change in the seasonality.

Here is the model built by HANA that we took and estimated in Autobox which matches what SAP shows in their User's guide. We suppressed the identification of outliers, etc. in order to match. None were searched for and identified by SAP. Maybe they don't do this for time series?? We can't tell, but they didn't present any so we can assume they didn't. That's a lot of seasonal factors. Count them....1,2,3,4. That's a red flag of a system that is struggling to model the data. We have never seen two MA4's in a model built by Autobox.

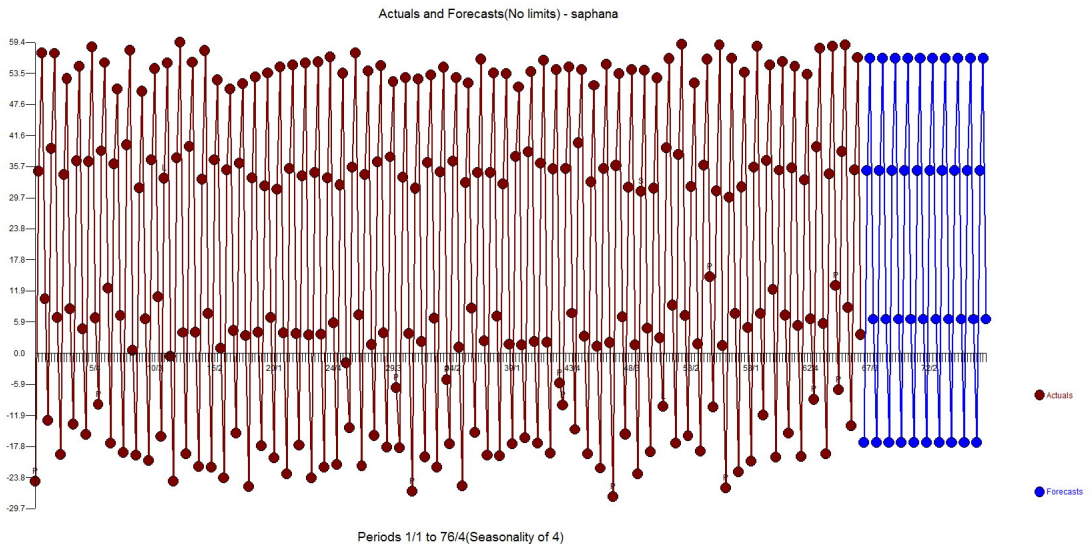
```
[(1-B**4)]Y(T) =  .15000E-01  +     [(1-  .255B** 1)(1-  .749B** 4)]**-1 [
                 (1-  .249B** 4)(1-  .534B** 4)] [A(T)]
```

With that said, the truth is the forecast from Autobox and SAP HANA is about identical when using all 264 observations. However, you need to consider other origins. If you go back 7 periods and use 257 observations (and assume the same model which I feel safe doing for this example) to where there were back to back outliers the forecast isn't good. The forecast(as

expected) using 256,258,259,260 observations is also bad so yes you want to adjust for outliers and there are consequences if you don't. To reproduce all of this in Autobox, run with Option "4" with this Autobox file, model file and rules .



Here is the forecast using Autobox using all the data(SAP HANA is just about the same)



The example is quarterly data for 264 periods. What SAP HANA doesn't recognize(likely others too?) is that with large samples the standard error is biased to providing false positives to suggest adding variables to the model. When this occurs, you have "injected structure" into the process and created a Type 2 error where you have falsely concluded significance when there was none.

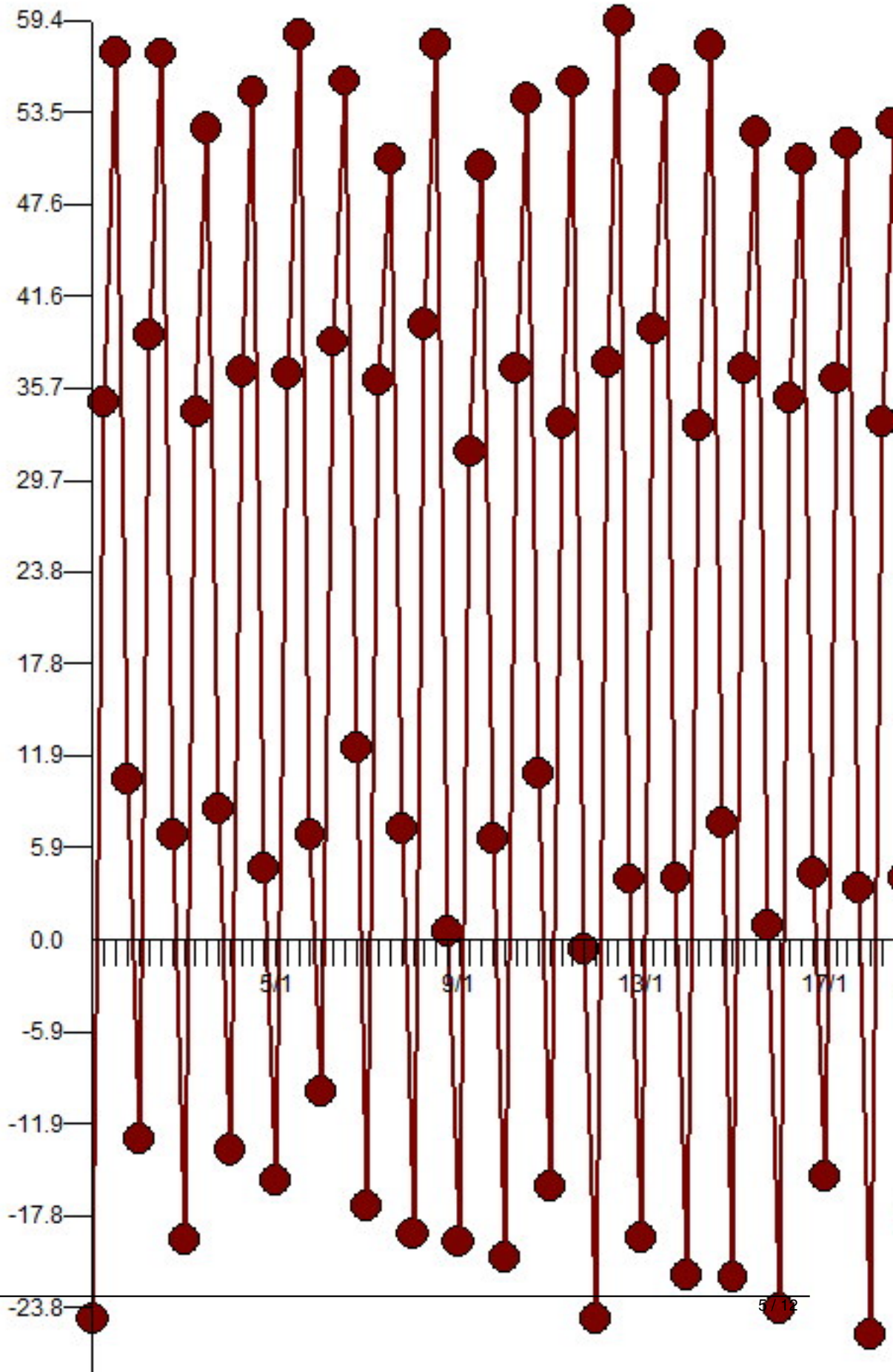
The example is a fun one in a couple of ways. The first observation is an outlier. We had seen other tools forecast CHANGE adversely when you deliberately change the first value to an outlier.

The SAP model shown on page 354 uses an intercept AR1, seasonal differencing, and AR4 and two MA4's. Yes, two MA4's.

Let's take a look at what Autobox does with the example.

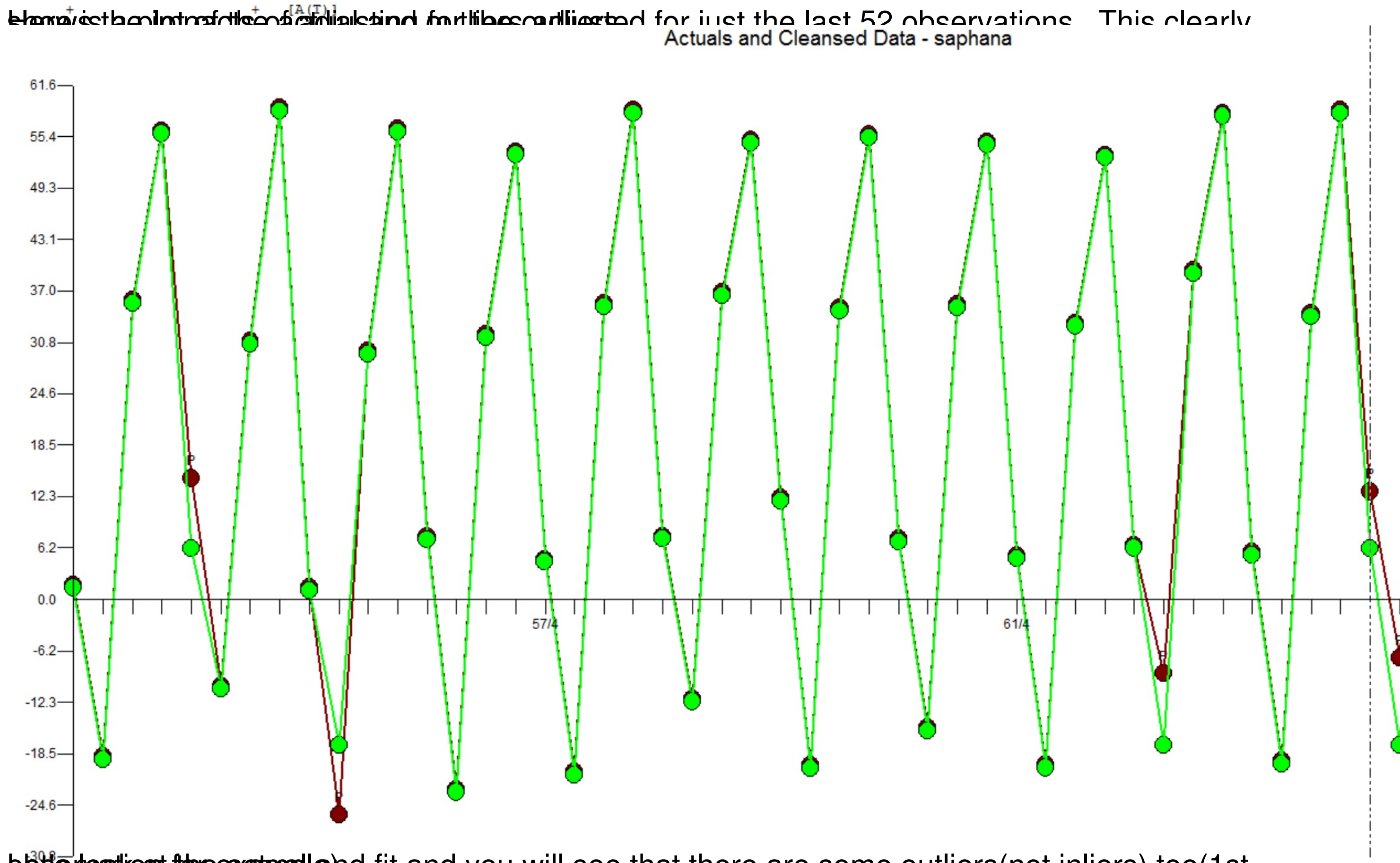
When you look at the plot of the data you, you might notice that the level of the data starts high, goes lower and then goes back to the initial level. this called a "level shift". You can calculate

local averages of these 3 groups to verify on your own. of the The biggest culprit is the 4th quarter which seems and then look at the Autobox actual and fit, it becomes easier to see how the data drops down in the middle and then goes back to the previous level.

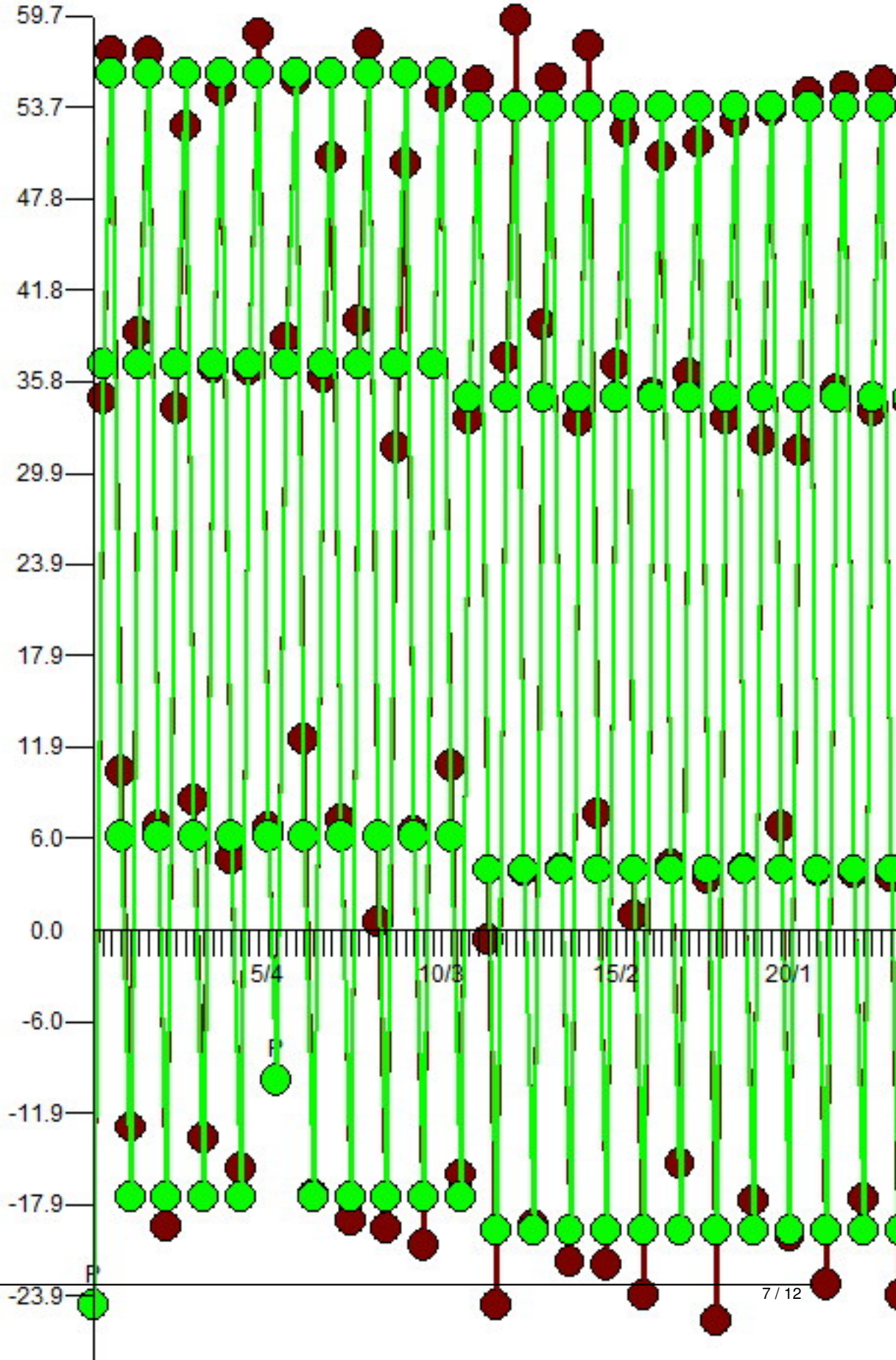

59.4
53.5
47.6
41.6
35.7
29.7
23.8
17.8
11.9
5.9
0.0
-5.9
-11.9
-17.8
-23.8
5/1
9/1
13/1
17/1

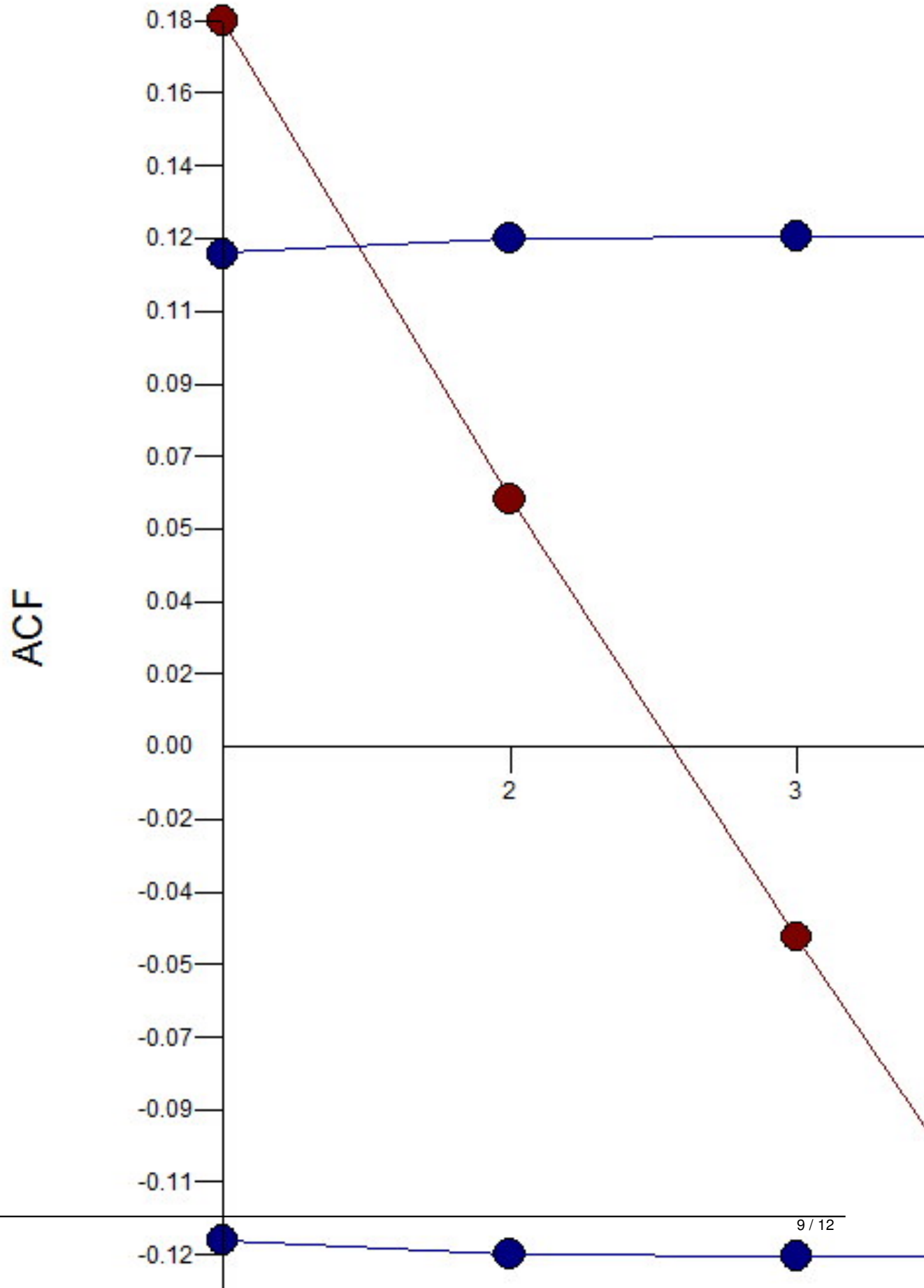
```
  AUTOMATIC FORECASTING SYSTEMS
         HATBORO PA 19040
            215-675-0652
      VERSION: 06/13/2016 08:48


MODELLING OUTPUT SERIES:saphana


Y(T) =   6.1016                          saphana
         +[X1(T)][(- 23.5289)]          FIXED_EFF_N10104
         +[X2(T)][(+ 30.8381)]          FIXED_EFF_N10204
         +[X3(T)][(+ 49.8556)]          FIXED_EFF_N10304
         +[X4(T)][(+ 2.5268)]           :LEVEL SHIFT      51/  1   201
         +[X5(T)][(- 2.1701)]           :LEVEL SHIFT      11/  2    42
         +[X6(T)][(- 10.5115)]          :PULSE            29/  4   116
         +[X7(T)][(+ 10.0705)]          :PULSE            65/  1   257
         +[X8(T)][(- 9.6615)]           :PULSE            42/  4   168
         +[X9(T)][(+ 9.6023)]           :PULSE            43/  1   169
         +[X10(T)[(- 9.0415)]           :PULSE            33/  4   132
         +[X11(T)[(- 8.7095)]           :PULSE            56/  1   221
         +[X12(T)[(+ 8.0667)]           :PULSE            54/  4   216
         +[X13(T)[(+ 8.1955)]           :PULSE            63/  1   249
         +[X14(T)[(- 7.8727)]           :PULSE            47/  1   185
         +[X15(T)[(- 7.0977)]           :PULSE             1/  1     1
         +[X16(T)[(- 6.8727)]           :PULSE            31/  1   121
         +[X17(T)[(+ 7.6123)]           :PULSE             6/  1    21
         +[X18(T)[(- 2.4304)]           :SEASONAL PULSE   49/  2   194
         +[X19(T)[(+ 6.8505)]           :PULSE            51/  1   201
         +[X20(T)[(+ 6.4517)]           :PULSE            64/  4   256
                       +      [A(T)]
```

Here is the plot of the actuals and the cleansed for just the last 52 observations. This clearly

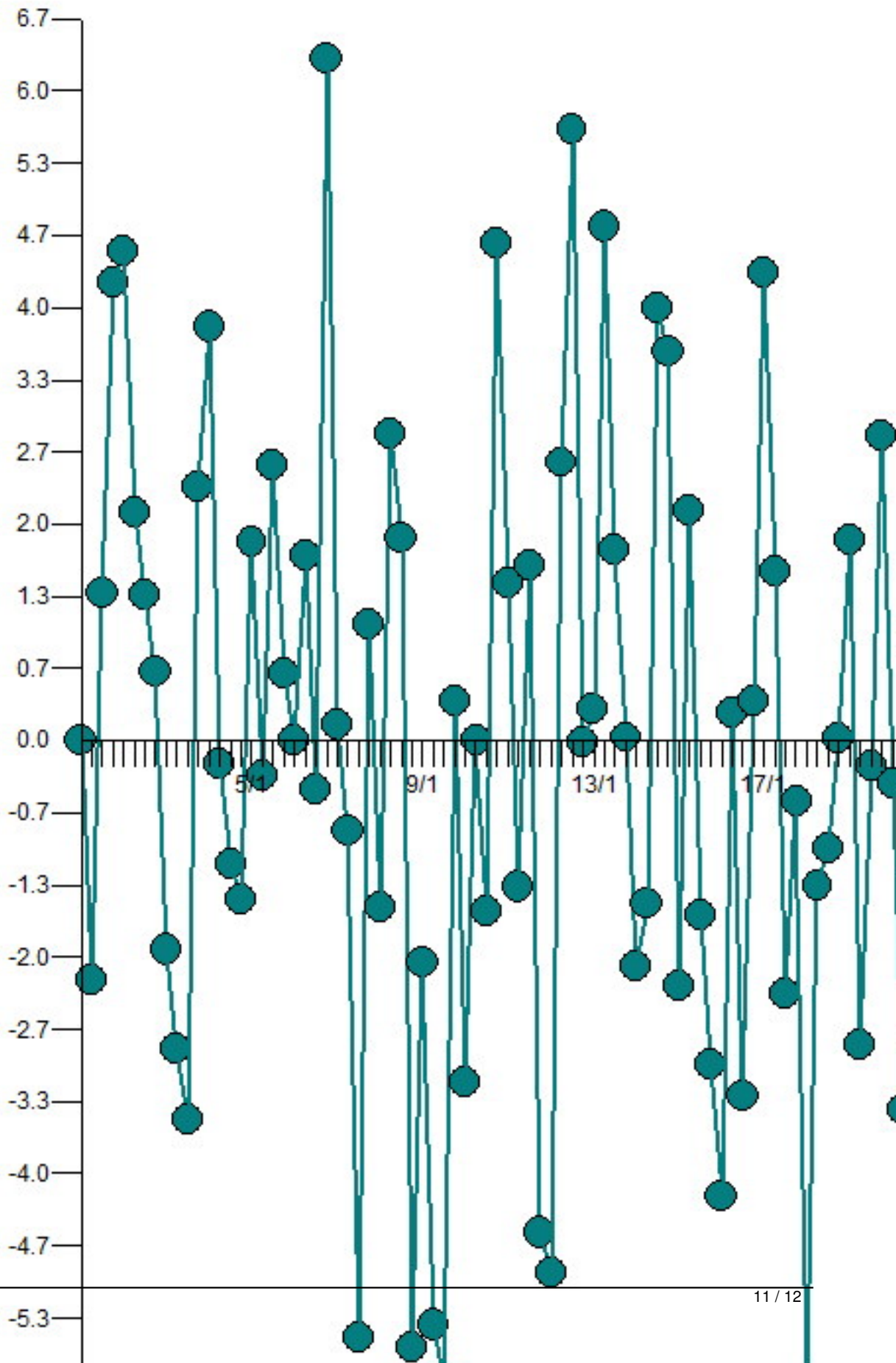


Let's look at the actual and fit and you will see that there are some outliers(not inliers) too(1st observation for example).


59.7
53.7
47.8
41.8
35.8
29.9
23.9
17.9
11.9
6.0
0.0
-6.0
-11.9
-17.9
-23.9
5/4
10/3
15/2
20/1
P
P


ACF
0.18
0.16
0.14
0.12
0.11
0.09
0.07
0.05
0.04
0.02
0.00
-0.02
-0.04
-0.05
-0.07
-0.09
-0.11
-0.12
2
3

Here are the Autobox random, free of pattern and trend residuals(ie N.I.I.D.):


6.7
6.0
5.3
4.7
4.0
3.3
2.7
2.0
1.3
0.7
0.0
-0.7
-1.3
-2.0
-2.7
-3.3
-4.0
-4.7
-5.3
5/1
9/1
13/1
17/1

Tags: Tagged in: ARIMA forecasting inliers outliers SAP HANA