

How does SAP stack up against Autobox?- A real daily data comparison using SAP's Predictive Analytics.

Posted on Mar 07, Posted by [Tom Reilly](#) Category [Forecasting](#)

SAP has a [webpage with a tutorial](#) on using their Predictive Analytics 2.3 tool(formerly KXEN Modeler)using daily data. They released this back in December, but didn't see until browsing Twitter. It provides an unusual public record of what comes out of SAP. They didn't publish the model with p-values and all of the output, but this is good enough to compare against. We ran numerous scenarios with different modeling options to understand what the outcome would be using these modeling(ie variable) techniques. Autobox has some default variables it brings in with daily data. We will have to suppress some of those features so that when we use the SAP variables they don't collide with them and make a multicollinear regression.

The [Tutorial](#) is well written and allows you to easily download the 1,724 days of data and model this yourself. While SAP had a .13 MAPE(in sample), they had a challenge at the end for those who get a MAPE less than .12 to contact them. Can you predict what Autobox did? .0724. Guess who is going to contact them? I will also add, that if you can do better contact us as we might have something to learn too. I also suggest that you post how other tools handle this as well as that would be interesting to see as well. Autobox [thrives](#) (1st among automated) on daily data as it did in a daily forecasting competition and is much more difficult to model and something we have dedicated 25 years to perfecting.

After reading the SAP [user's guide](#) let's make the distinction that Autobox uses all of the data to build the model, while SAP (like all other tools) withholds data to "train" on.

Autobox adjusts for outliers. One could argue that by using adjusting for outliers the MAPE will only go down which is true, but it be aware that it allow for a clearer identification of the relationships in the data(ie coefficients / separating signal from noise).

The first approach in the SAP tutorial is running with only historical data and they add in the causals later. Outliers are identified and has a MAPE of .197.

66 Variables

A bunch of very curious variables(66??----PenultimateWednesday) are included that we have never seen before which made us scratch our heads (with delight???). They seem to try and capture the day of the week so we will turn that off some of Autobox's searches to avoid collinearity when we run with these in the first pass. They seem to use a day of year variable which I have never seen before. What book are they getting ideas to use these kind of variables from? Not one that I have ever seen, but perhaps someone can enlighten me? There are two variables that are measuring the number of working days that have occurred in the month and the number left in the month. We did find that some of these variables do have importance in the tests we ran so SAP has some ideas generating useful variables, but many are collinear and this could be called "kitchen sink" modeling. We will do more research into these. There is a holiday variable which also flags working days so the two variables would seem to be collinear. These two end up as the second and third most powerful variables in the SAP model. When we tried these in Autobox, both runs found them significant. Perhaps they measure (implicitly) holidays too? We are not sure, but they help.

AB	AC
MonthWorkingDayInd	ReverseMonthWorkingDayInd
0	0
0	0
0	0
1	20
2	19
3	18
4	17
0	0
0	0
5	16
6	15
7	14
8	13
9	12
0	0
0	0
10	11
11	10
12	9
13	8
14	7
0	0

There are weather variables which are useful and actually represent seasonality so using **both monthly dummies/weekly dummies and the weather variables could be problematic**. The holidays have been all combined into one catch all variable. This assumes that each holiday behaves similarly. It should be noted that a major difference is that SAP does not search for lead or lag relationships around the causals while Autobox can do that. Just try running

[this example](#)

in Autobox and then SAP. We ran with all of these curious variables.

We then reduced these variables and kept only Holiday, gust, rain, tmean, hmean,

dmean, pmean, wmean, fmean, TubeStrike and Olympics and removed the curious other variables.

The question which might arise "how much can you trust the weather predictions?", but here we are looking at only the MAPE of the fit so that is not a topic of concern.

SAP ended up with a .13 MAPE when using there long list of causals. The key here is that no outliers are identified in the analysis. This is a distinction and why Autobox is so different. If you ignore outliers they do still exist and yes they exist in causal problems. By ignoring something that doesn't mean it goes away, but ends up impacting you elsewhere such as the model and you likely aren't even aware of its impact. By not being able to deal with outliers your model with causals will be skewed, but no one talks about this in any school or text book so sorry to ruin this illusion for you. [Alice in Wonderland](#) (search on alice) thought everything was perfect too, until.....

Autobox does stepdown regression, but also does "stepup" where it will search for changes in seasonality(ie day of the week), trend/level/parameters/variance as things sometimes drastically change. If you're not looking for it then you will never find it! s. The MAPE we are presenting can be found in the detail.htm audit report from the Autobox run(hint:near the bottom). We suppressed the search for special days of the month which are useful in ATM data, but not theoretically plausible for this data. Autobox allows for holidays in the Top 15 GDP's, but in general assumes the data is from the US so we will need to suppress that search. We suppressed the search for special days of the month which are useful in ATM daily data as payday's are important, but not theoretically plausible for this data.

To summarize: We can run this a few different ways, but we can't present all of these results down below as it would be too much information to present here. We included some output and the Autobox file (current.asc-rename that if you want to reproduce the results) so you can see for yourself. What we do know is that including ARIMA increases run time.

MAPE's

- Run using all variables with Autobox default options(suppressing US Holidays, day of month and monthly/weekly dummies). [.0883](#)
- Run using all variables with Autobox default options(suppressing US Holidays, day of month and monthly/weekly dummies). Allow for ARIMA [.0746](#)
- Run using a reduced set of variables(see above) & suppressing US holidays, day of month and monthly/weekly dummies). [.1163](#)

- Run using a reduced set of variables(see above) & suppressing US holidays, day of month and monthly/weekly dummies). Allow for ARIMA [.0732](#)
- Run using only Holiday, Strike/Olympics and rely upon monthly or weekly dummies. [.1352](#)
- Run using only Holiday, Strike/Olympics and rely upon monthly or weekly dummies. Allow for ARIMA [.1145](#)
- Run using a reduced set of variables, but remove the catch all "holiday" variable and create separate 6 main holiday variables that were flagged by SAP as they might each behave differently. (suppressing US Holidays, day of month, and monthly/weekly dummies) [.1132](#)
- Run using a reduced set of variables, but remove the catch all "holiday" variable and create separate 6 main holiday variables that were flagged by SAP as they might each behave differently. (suppressing US Holidays, day of month, and monthly/weekly dummies). Allow ARIMA [.0724](#)

Let's consider the model that was used to develop the lowest MAPE of .0724.

There were 38 outliers identified over the 1,724 observations so the goal is not to have the best fit, but to model and be parsimonious.

So, what did we do to make things right? We started by deleting all kinds of variables. There were [linearly redundant variables](#) such as WorkingDay that is perfectly correlated (inverse here) to Holiday which by definition should never be done when using dummy variables. The variable "Special Event" is redundant with TubeStrike and Olympics as well. Special Event name isn't even a number, but rather text and also is redundant.

W	X	Y
WorkingDay	BeforeHoliday	Holiday
0	1	1
0	1	1
0	0	1
1	0	0
1	0	0
1	0	0
1	1	0
0	1	1
0	0	1
1	0	0
1	0	0

All other software withholds data whereas Autobox uses all of the data to build the model as we

have adaptive technology that can detect change (seasonality/level/trend/parameters/variance plus outliers). We won best [dedicated forecasting tool](#) in J. Scott Armstrong's "Principles of Forecasting". For the record, we politely disagree against a few of the 139 "

[Principles](#)

" as well.

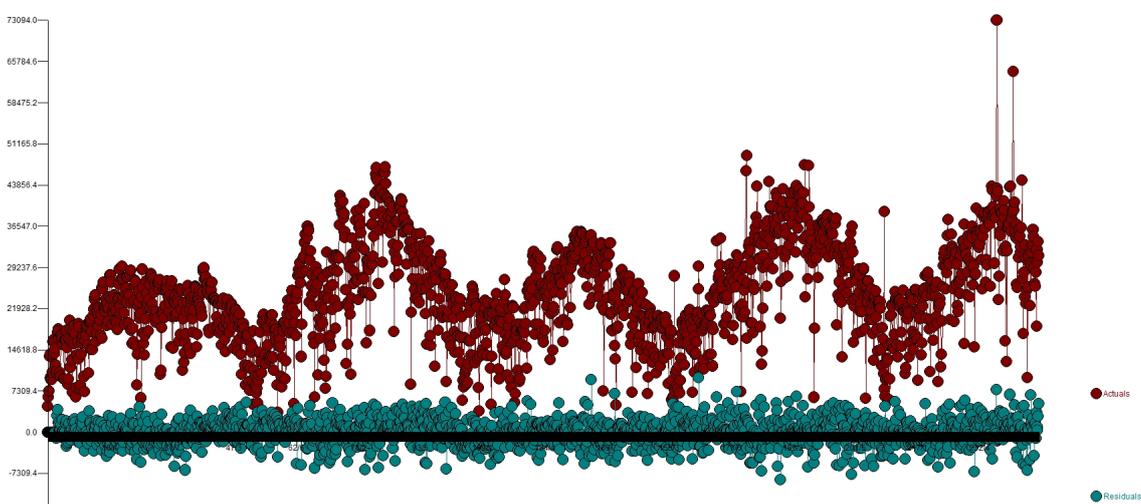
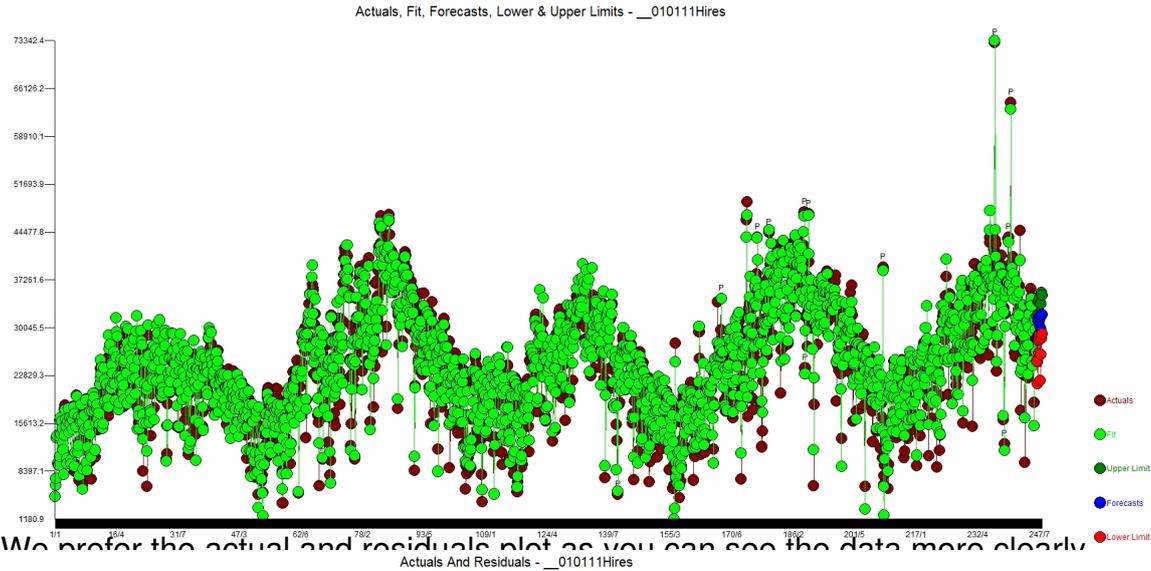
We report the in sample MAPE, in the file "details.htm" seen below...

1710	09/06/15	1.00	29539.000	22997.642
1711	09/07/15	1.00	31917.000	30134.641
1712	09/08/15	1.00	32968.000	32055.560
1713	09/09/15	1.00	33332.000	33675.527
1714	09/10/15	1.00	35956.000	33544.046
1715	09/11/15	1.00	34690.000	34433.560
1716	09/12/15	1.00	29143.000	33340.957
1717	09/13/15	1.00	26342.000	24974.880
1718	09/14/15	1.00	25923.000	24310.646
1719	09/15/15	1.00	28311.000	29442.619
1720	09/16/15	1.00	18764.000	15244.483
1721	09/17/15	1.00	32336.000	31385.990
1722	09/18/15	1.00	30124.000	29795.899
1723	09/19/15	1.00	33855.000	31079.578
1724	09/20/15	1.00	31380.000	26337.365

BASED UPON 1724 ESTIMATES

MODEL MEAN ABS PCT ERROR	=	7.243897
RANDOM WALK MEAN ABS PCT ERROR	=	17.157919
AVERAGE MEAN ABS PCT ERROR	=	28.786044

And this is only a small sample of the data used for the comparison. The full data set is available in the report of this analysis.



Let's review the model

The sign of the coefficients make sense(for the UK which is cold). When it's warmer people will skip the car and use the bike, for example so when Temperature goes up (+ sign) then people rent more bikes. When its gusty people will not and just drive. The tutorial explains the variables names in the back. tmean is average temperature, w is wind, d is dewpoint, h is humidity, p is barometric pressure, d is real feel temperature. All 6 holidays were found to be important with all but one having lead or lag impacts. When you see a B**-2 that means two days before the Christmas volume was low by 5036. Autobox found all 6 days of the week to be important. The SAP Holiday variable was a mixture of Saturday and Sunday and causes some confusion with interpretation of the model. This approach is much cleaner. The first day of the data is a Saturday(1/1/2011) and the variable "FIXED_EFF_N10107" is measuring that impact that Saturday is low by 4114. Sunday is considered average as day 7 is the baseline. See below for more on the day of the week rough verification(ie pivot table/contribution %).

Note the "level shift" variables added to the model. This meant that the volume changed up or down for a period and Autobox identified and ADAPTED to it. We call this " [step up regression](#) " (nothing there right? Yes, we own that world!) as we are identifying on the fly deterministic variables and adding them to the model. The runs with the SAP variables fit 2012 much better. The first time trend began at period 1 with volume steadily increasing 10.5 units each day. This gets tampered down with the second time trend beginning at 177 making the net effect +4.3 increase per day. 38 outliers were identified which is the key to whole analysis. They are sorted by their entry into the model and really their importance.

```

Y(T) = 13336.          ___010111Hires
+ [X1(T)] [(+ 1285.4  )]      tmean
+ [X2(T)] [(- 318.48  )]      tmean
+ [X3(T)] [(- 60.0051)]      cust
+ [X4(T)] [(- 476.15  )]      rain
+ [X5(T)] [(- 978.49  )]      tmean
+ [X6(T)] [(+ 246.32  )]      tmean
+ [X7(T)] [(+ 15642.  )]      TubeScrIke
+ [X8(T)] [(+ 8580.4  )]      Olympics
+ [X9(T)] [(+ 5036.3  )]      B** -2- 8815.0  B** -1+ 4724.4
- 8497.8  B** 1- 10529.  B** 2- 7037.4  B** 3
- 5540.4  B** 4]
+ [X10(T)] [(- 2612.2  )]      B** -1- 7678.3  - 4424.7  B** 1
- 4410.7  B** 2- 5222.0  B** 3]
+ [X11(T)] [(- 4529.2  )]      B** -2+ 3502.5  B** -1] SUMMERBANKDAY
+ [X12(T)] [(- 3456.5  )]      EARLYBANKDAY
+ [X13(T)] [(- 7095.3  )]      - 4898.8  B** 1] SPRINGBANKDAY
+ [X14(T)] [(- 6505.3  )]      B** -2- 8441.1  B** -1- 8475.4
- 4977.7  B** 3]
+ [X15(T)] [(- 4114.4  )]      NEWYEARS
+ [X16(T)] [(- 5461.4  )]      FIXED EFF_M10107
+ [X17(T)] [(- 767.83  )]      FIXED EFF_M10207
+ [X18(T)] [(+ 839.67  )]      FIXED EFF_M10307
+ [X19(T)] [(+ 1020.6  )]      FIXED EFF_M10407
+ [X20(T)] [(+ 712.75  )]      FIXED EFF_M10607
+ [X21(T)] [(+ 7599.8  )]      :LEVEL SHIFT 03/09/12 62/ 7 434
+ [X22(T)] [(- 4241.3  )]      :LEVEL SHIFT 03/01/13 115/ 7 791
+ [X23(T)] [(+ 5615.1  )]      :LEVEL SHIFT 09/02/14 174/ 7 1238
+ [X24(T)] [(- 2149.4  )]      :LEVEL SHIFT 12/13/14 207/ 1 1443
+ [X25(T)] [(+ 16509.  )]      :PULSE 08/09/14 189/ 1 1317
+ [X26(T)] [(+ 15901.  )]      :PULSE 08/14/14 180/ 1 1324
+ [X27(T)] [(- 14010.  )]      :PULSE 02/04/14 162/ 4 1131
+ [X28(T)] [(+ 12280.  )]      :PULSE 09/20/14 196/ 1 1359
+ [X29(T)] [(+ 14008.  )]      :PULSE 07/09/15 836/ 6 1651
+ [X30(T)] [(- 8416.4  )]      :PULSE 08/02/11 31/ 4 214
+ [X31(T)] [(+ 7782.9  )]      :PULSE 10/04/12 93/ 1 645
+ [X32(T)] [(- 7967.0  )]      :PULSE 09/01/11 31/ 3 213
+ [X33(T)] [(+ 12228.  )]      :PULSE 08/17/14 190/ 2 1325
    
```

Note the Seasonal pulse where the first day becomes much higher starting at period 1639 and forward with an average 3956.8 higher volume. Thats quite a lot and if you do some simple plotting of the data it will be very apparent. Day 1 and Day 2 were always low, but over time Day 1 has become more average, Note the AR1 and AR7 parameters.

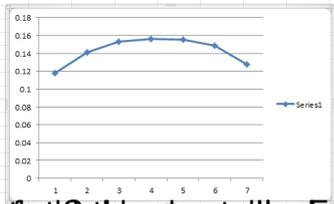
How does SAP stack up against Autobox?- A real daily data comparison using SAP's Predictive Analytics

+ [K34 (T) (- 8073.2)]	: PULSE	09/15/13	142/	2	989
+ [K35 (T) (- 7057.6)]	: PULSE	06/27/11	26/	3	178
+ [K36 (T) (- 5225.9)]	: PULSE	07/15/13	133/	3	927
+ [K37 (T) (- 4144.6)]	: PULSE	08/03/11	31/	5	215
+ [K38 (T) (- 3956.8)]	: SEASONAL PULSE	06/27/15	235/	1	1639
+ [K39 (T) (- 4389.5)]	: PULSE	07/17/13	133/	5	929
+ [K40 (T) (- 10249.)]	: PULSE	08/06/15	240/	6	1679
+ [K41 (T) (- 7742.0)]	: PULSE	03/24/12	65/	1	449
+ [K42 (T) (- 5856.3)]	: PULSE	08/23/11	34/	6	235
+ [K43 (T) (- 5501.8)]	: PULSE	10/03/12	92/	5	642
+ [K44 (T) (- 5501.5)]	: PULSE	06/04/11	23/	1	155
+ [K45 (T) (- 9190.8)]	: PULSE	06/08/14	180/	2	1255
+ [K46 (T) (- 7590.2)]	: PULSE	08/10/14	189/	2	1318
+ [K47 (T) (- 4469.6)]	: PULSE	03/25/12	65/	2	450
+ [K48 (T) (- 13544.)]	: PULSE	12/25/14	208/	6	1655
+ [K49 (T) (- 4573.3)]	: PULSE	08/02/15	240/	2	1675
+ [K50 (T) (- 4916.9)]	: PULSE	11/01/12	96/	6	671
+ [K51 (T) (- 7850.6)]	: PULSE	05/18/14	177/	2	1234
+ [K52 (T) (- 4212.0)]	: PULSE	09/22/12	91/	1	651
+ [K53 (T) (- 4664.8)]	: PULSE	07/22/12	82/	2	569
+ [K54 (T) (- 8840.1)]	: PULSE	10/05/14	197/	2	1374
+ [K55 (T) (- 10065.)]	: PULSE	06/04/12	75/	9	521
+ [K56 (T) (- 5051.0)]	: PULSE	09/26/11	94/	7	238
+ [K57 (T) (- 9435.7)]	: PULSE	07/26/15	239/	2	1668
+ [K58 (T) (- 4761.6)]	: PULSE	03/16/14	166/	2	1171
+ [K59 (T) (- 5522.3)]	: PULSE	09/03/14	188/	2	1311
+ [K60 (T) (- 11119.)]	: PULSE	04/15/12	68/	2	471
+ [K61 (T) (- 4212.9)]	: PULSE	02/05/11	6/	1	36
+ [K62 (T) (- 5190.2)]	: PULSE	04/28/11	17/	6	118
+ [K63 (T) (- 5785.0)]	: PULSE	04/23/11	17/	1	113

Let's consider the day of the week data by building a pivot table.

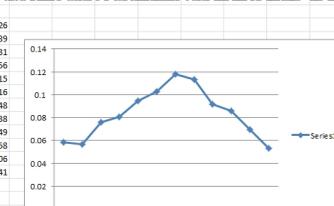
1	vol	day
2	4555	7
3	6250	1
4	7282	2
5	13430	3
6	13757	4
7	9595	5
8	9294	6
9	9338	7
10	10558	1
11	16058	2
12	16412	3
13	13894	4
14	15911	5
15	14938	6

Row Labels	Sum of vol	Sum of day	Count of day
1	4979985	4979985	0.118174
2	5943216	5943216	0.141117
3	6456356	6456356	0.153301
4	6574921	6574921	0.156116
5	6540346	6540346	0.155295
6	6260932	6260932	0.148661
7	5362809	5362809	0.127336
(blank)			
Grand Total	42115565	42115565	



utilized the milkOvith milk and the 5m, the breakage and the temperature at higher

Row Labels	Sum of volume	Sum of milkOvith	Sum of milk
1	2137447	2137447	0.05826
2	2070632	2070632	0.056439
3	2782068	2782068	0.075831
4	2951758	2951758	0.080456
5	3467564	3467564	0.094515
6	3764777	3764777	0.102616
7	4330927	4330927	0.118048
8	4165464	4165464	0.113538
9	3369736	3369736	0.091849
10	3147820	3147820	0.0858
11	2533677	2533677	0.069606
12	1945968	1945968	0.053041
(blank)			
Grand Total	36687838	36687838	



Tags Tagged in: centerets automa grell rats outlier forecasting daily Monthly weekly