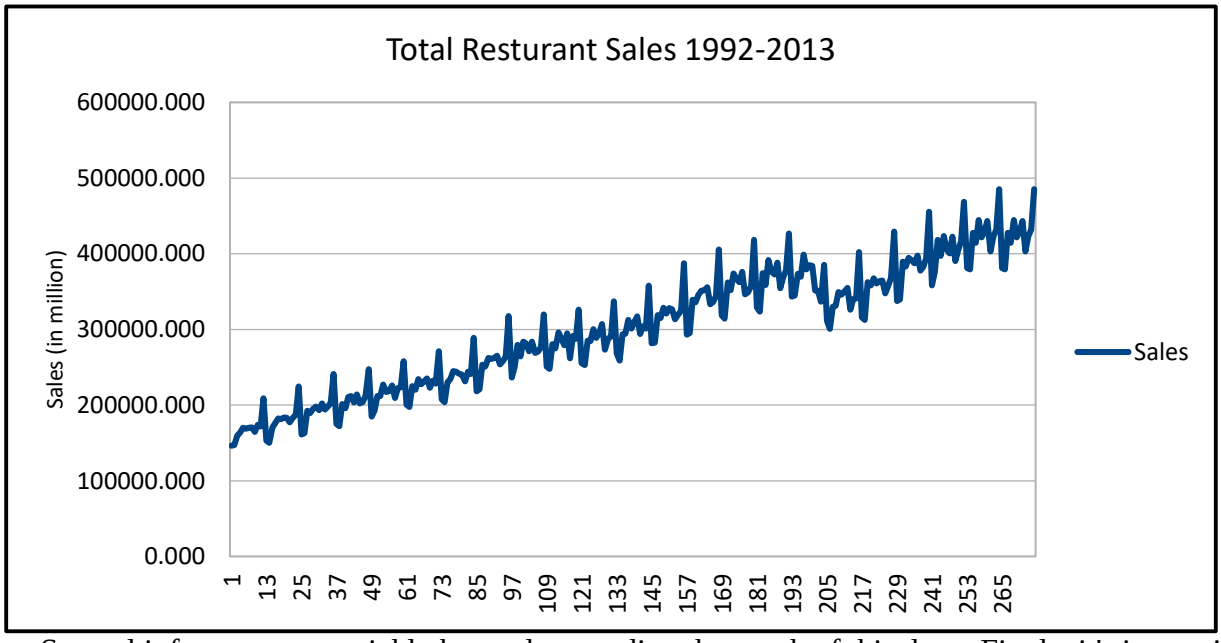
Forecasting Resturant Sales for 2014

By Scott Brown

I've always been a man interested in understanding what numbers within data are trying to tell you. After spending 10 years working at Chick-Fil-A, I often tried to figure out why sales and food predictions were either way too conservative or just totally off. During my last 2 years, I had the opportunity to try my luck with creating projection models but unfortunately, when you have to manage an entire store and kitchen, there just wasn't time for it. I figured I would give it another shot and try to forecast annual restaurant sales in the United States taken directly from the census.



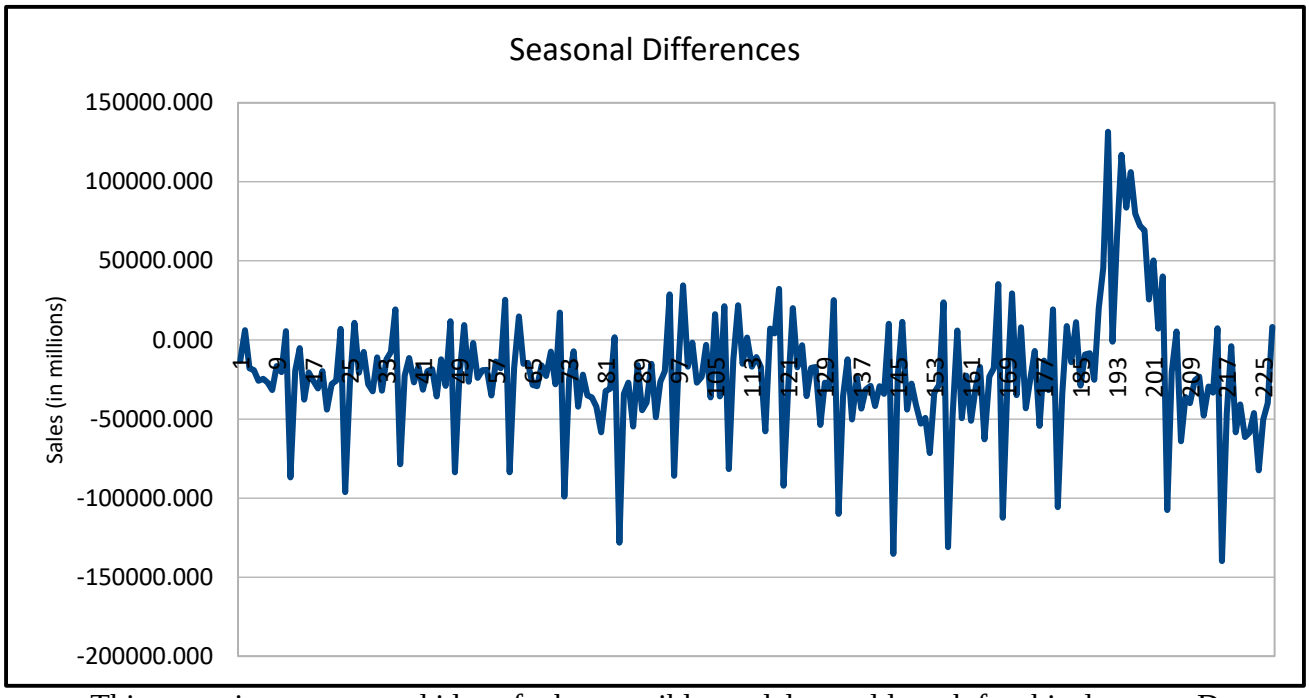
Several inferences can quickly be made regarding the graph of this data. Firstly, it's increasing over time were indicates it has a multiplicative pattern which needs to be taken into consideration for model building. Second, even though the data is multiplicative, it appears to have fairly constant variance throughout that is slightly increasing implying that a log transform would help the data but may not necessarily be required. Another element of interest is the dip around the 200th data point which may be from the effect of some sort of intervention. Lastly, the data appears to have some

seasonality in it as it has consistent peeks around 12 and also around 5 and 8.

Next, I'll determine the residual plot and seasonality for multiplicative decomposition:

This is when I started to have my doubts about this data set. The residual plot doesn't really indicate any kind of trend or correlation and appears fairly random with a bit of fanning despite the data being time based. However, the data has seasonality after running multiplicative decomposition. It dips for both months 1 and 2, jumps a bit for months 5 and 8 and has a spike at month 12. Curious enough, the dip around month 200 is visible in the residual plot as well giving further indication some sort of intervention is in the data. The next step will be to determine the intervention and continue despite the constant variance assumption appearing to not be violated by the randomness of the residuals even though the data is time based.

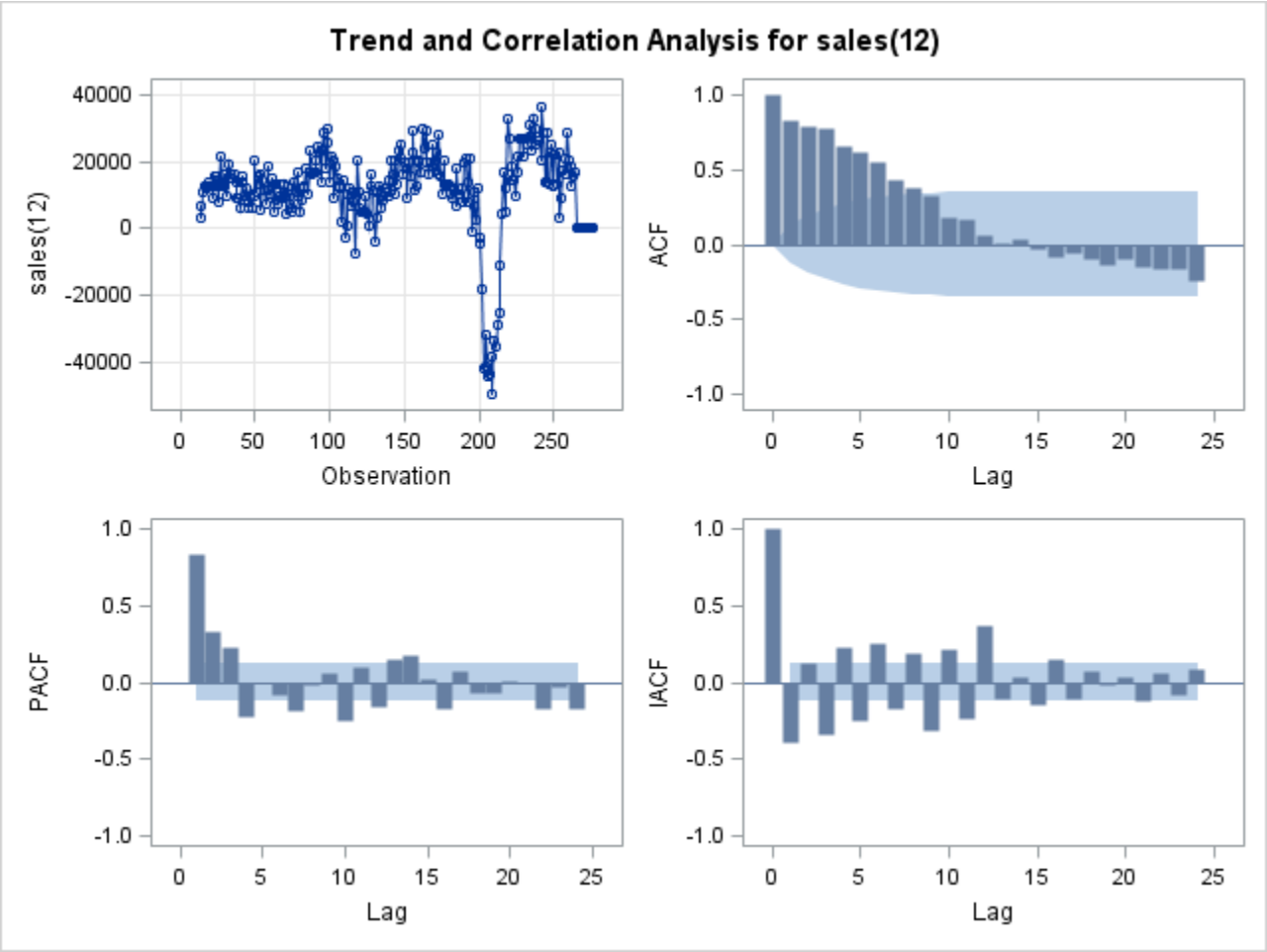
Despite the apparent intervention, it doesn't visibly appear significant in the original data. It does not appear obvious in the first differences but it clearly visible in the graph of the seasonal differences:

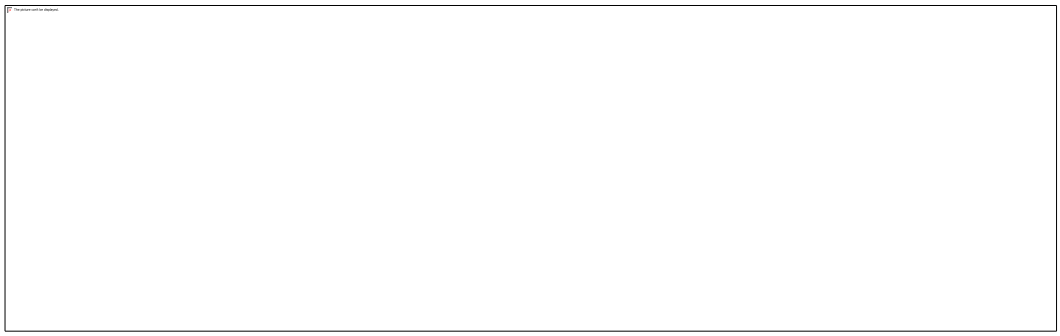


This now gives me a good idea of what possible models would work for this data set. Due to the clear intervention, one model will have the points coded out. Further inspection indicates the dip in data occurs between months 199 and 223. This drop in sales data can be easily explained as being caused by an external event because months 199 and 223 correspond to June of 2007 to June of 2009 which was during the great recession period in which many people (myself included) stopped going out to eat and had to cut back on spending. Despite this, the dip in sales during this time didn't appear significant in the original graph so another model will use all the original data for predictions. While both of these models will be obtained through Arima, I will use an exponential smoothing model, citing Occam's razor, to see if a model with the fewest parameters will still yield a good model.

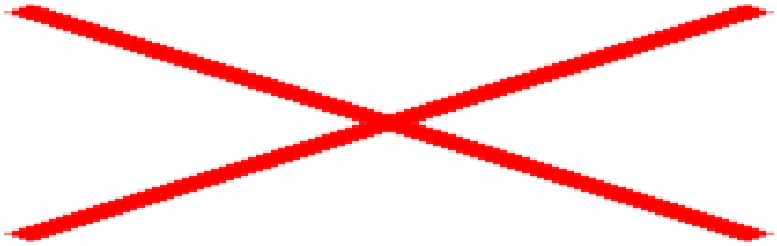
Since this data is seasonal, I must first evaluate 4 versions of the data in proc arima: the original data, first differences, seasonal differences and seasonal first difference. Not surprisingly, the seasonal difference data yielded the most “stationary” of all 4 models. The sac is dying down quicker than the original model and the spac cuts off around 1 with some spikes afterwards. An auto regressive model is confirmed to be the prefered model method since the residuals are shown to be autocorrelated and

not white noise. The seasonal difference model will be used.





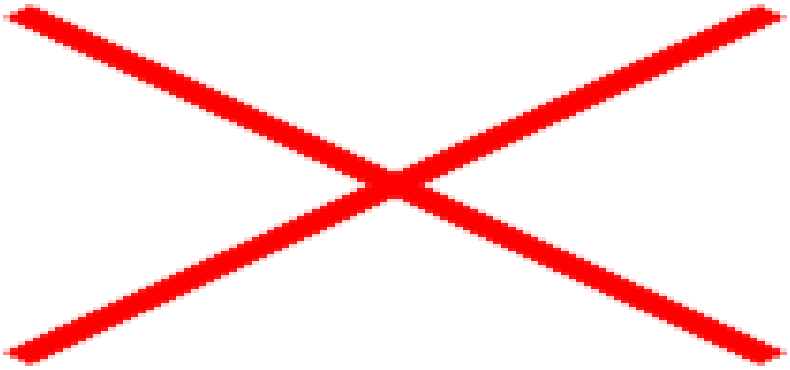
After creating models, the next obstacle was to figure out which model could be the considered the best for predictions. This process was a bit of a headache due to the many different ways to check which model is “best”. The book suggests looking for the model with the smallest standard error which made sense in theory however it seemed that the more arima parameters in the model, the smaller the standard error even if the forecast error wasn't the smallest.



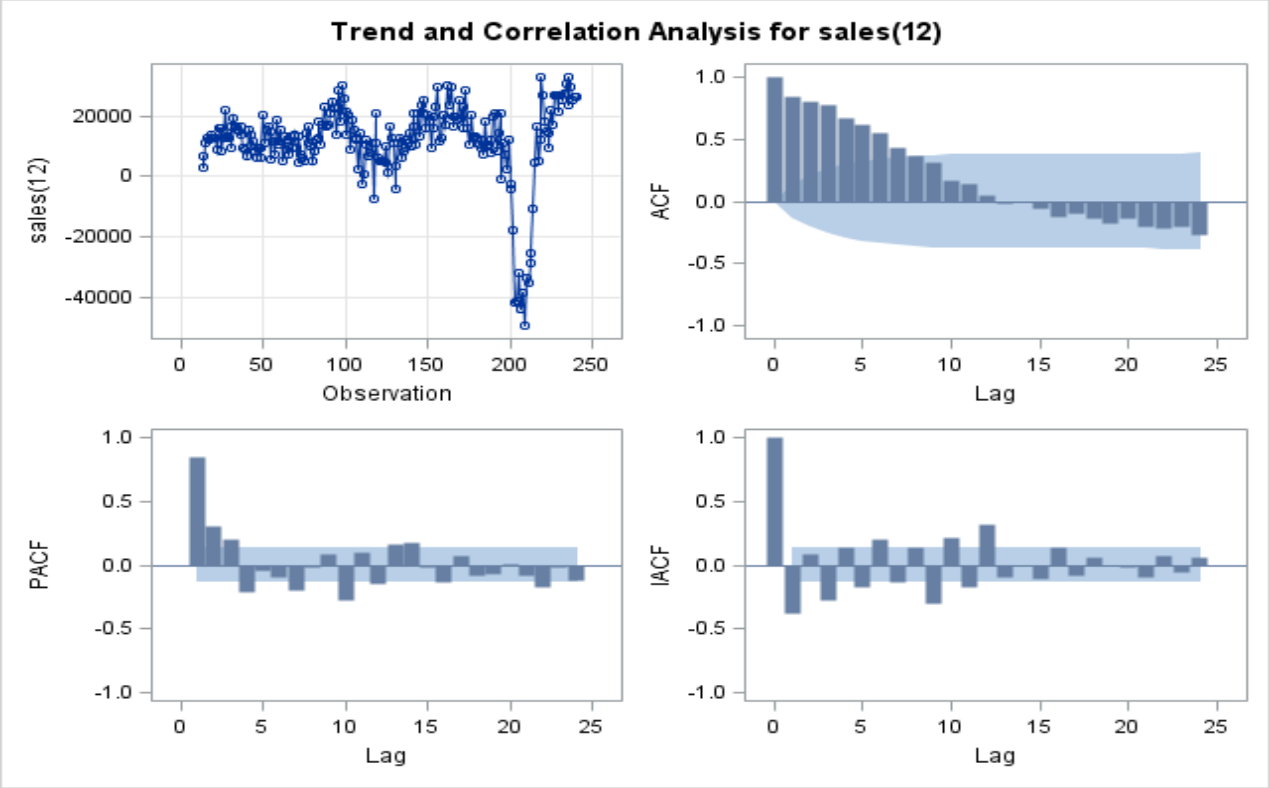
It behaved in a similar way to an R^2 value being inflated by additional variables (instead of using R^2 adjusted). I decided the best method would be to compare the s values using the predicted and actual values.

To add some variety to the model building process, I decided to use data points up to the end of 2010 (month 240) and use this to predict 2011 and compare it with the actual 2011 sales. The model with the smallest s would be picked to predict 2012, 2013 and finally 2014. Next, all data points up to 2011 (month 252) would be used to predict 2012 and again the model with the smallest s would be chosen to predict 2013 and 2014. I decided to go this route because it is described in the Holt's-Winters method that using some but not all of the data will still yield useful results for forecasting.

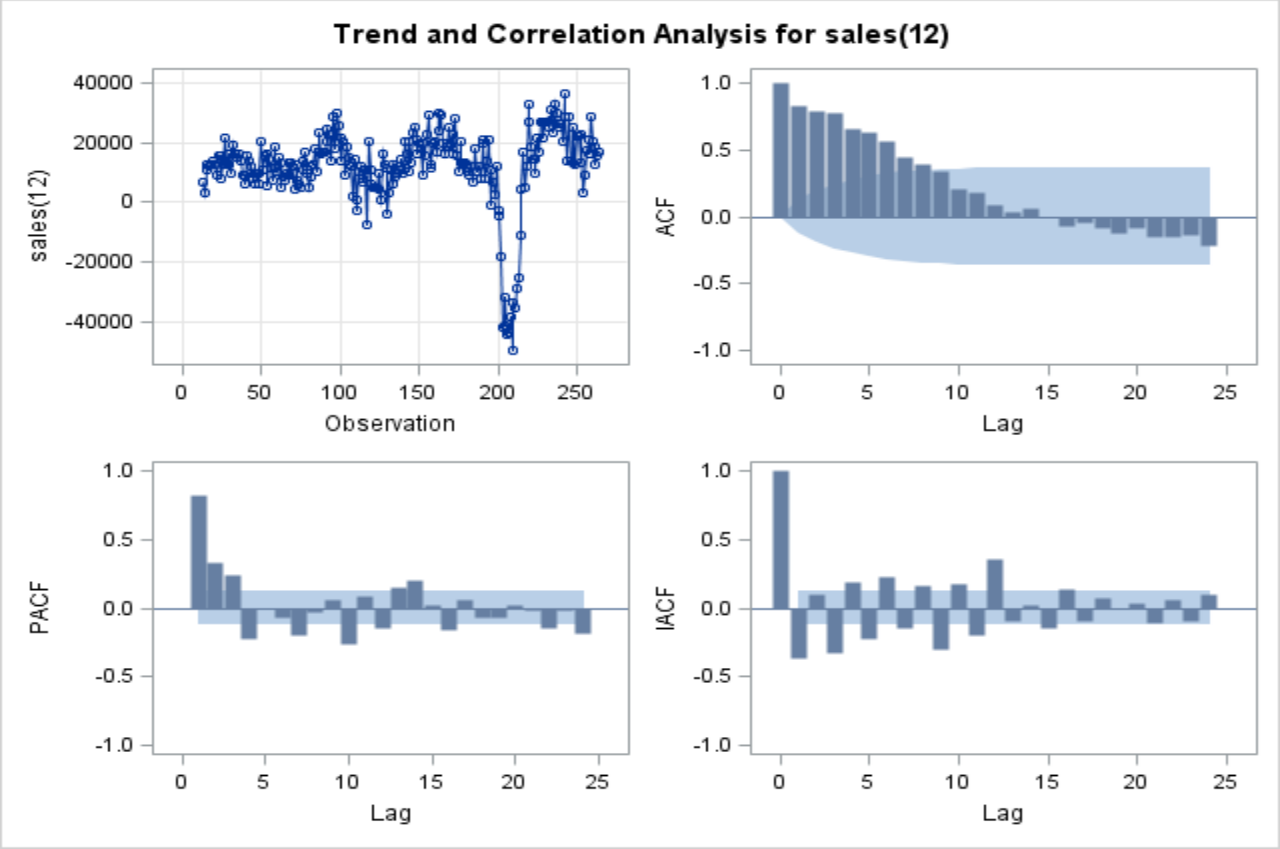
After finding the best models for all 3 methods, the following results are obtained:



Not surprisingly, an Arima model produced the best results for all 3 years. However, Holts-Winters produced the 2nd best of all the 2013 models which shows that using the “simple” exponential smoothing approach is indeed useful. Finding the Arima models was completed through trial and error after looking at each Sac/Spac plot:



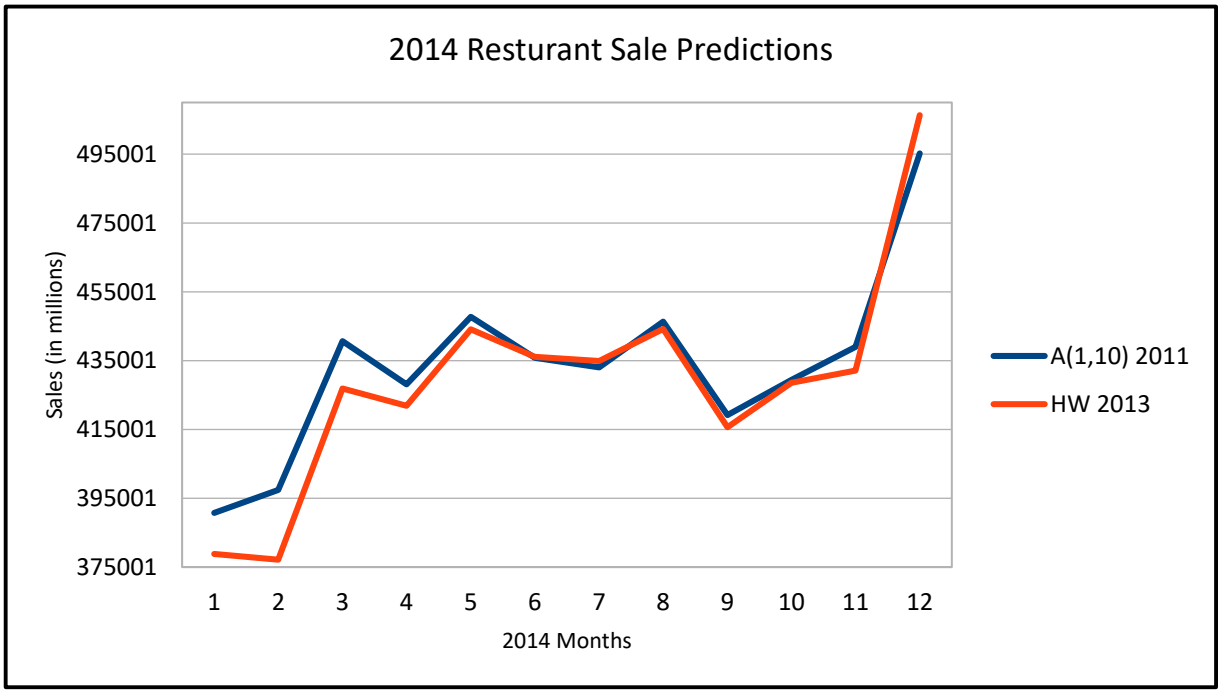
For predicting 2011 using the seasonal differences, the Sac is dying down quicker than the original Sac and the Spac cuts off at 1 with spikes at 2, 3, 4 ,7 and 10 (although these are significant they are much less than the 1st spike). After simulation, an arima model of p=(1,2,3,4) with no moving average produced the smallest s of 1533.472. Setting the auto regressive part of the model to 1,2,3 and 4 makes sense since these are where the first major spikes are in the Spac (a technique the book suggests to try for seasonal differenced data). However, this model should produce more adequate predictions with a moving average of order 12 added in to it due to the data being seasonally differenced. Perhaps adding another year of data will yield a more enlightening model.



For 2012 and 2013, both produced similar Sac and Spac's to the 2011 model however the best model for predicting 2012 is Arima p(1,10) q(12) and predicting 2013 is Arima p(1,10)(5) q(12). This is a bit odd since the arima model is taking only the 1st and 10th autoregressive spikes (order 1 and 10) into consideration for predictions while again seting the moving average to order 12. Perhaps cutting out the variance of all the other spikes led to a better model. This is similar for the 2013 model except a quadratic type model is found to yield the smallest s. I believe this maybe from the seasonality in the data since months 1,2, and 12 are the major months with seasonality (and thus a 10 month point gap between 2 and 12). The multipliying factor of 5 in the 2013 model could be from the spikes in months 5 and 8 which is arround the middle of the 10 month gap.

When comparing the best models for each year:

To my surprise, the 2011 model produced the best overall predictions. It also yielded the smallest s value of all models that I made (1247.469). It definitely seemed like a good idea to start at 2010 and work my way forwards since I found the best model was not from all data up to 2013. Now that I have found the best Arima model, I will match it with the 2013 Holts-Winters predictions:



Overall, the 2011 model seems like it would be adequate for predictions but it still has its own issues. Despite it producing the smallest prediction error for the data, it's residuals are still autocorrelated and can't be considered white noise. Also, the 10^{th} order specified in the auto regressive part of the model is found to be not significant at $\alpha=.05$:

Despite the model infirmity, I feel it's adequate for forecasting. It has been shown to be by far the best in terms of forecasting error and the autocorrelation could be explained due to the relatively slow pace in which the Sac is dying down. Ultimately, more work may need to be done for it but it's a good start for making 2014 predictions. I have speculated that creating 2 seperate models based on the intervention points (one model up to month 198 and another model from month 199 on) and taking a sort of "model average" between the two would be a good continuing point for the data set. If anything, it should help with the inflated varaince that was present in the 2012 and 2013 predictions using the original adjusted data. Afterall, less variance will help lead to a better model.

References Included:

"Monthly & Annual Retail Trade."- *Time Series Data*. N.p., 31 Jan. 2013. Web. 24 Apr. 2014.

Bowerman, Bruce L., Richard T. O'Connell, Anne B. Koehler, and Bruce L. Bowerman. *Forecasting, Time Series, and Regression: An Applied Approach*. Belmont, CA: Thomson Brooks/Cole, 2005. Print.

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ: Prentice Hall, 1994. Print.