Paper 156

Forecasting: Something Old Something New

David A. Dickey, North Carolina State University

ABSTRACT

ARIMA (AutoRegressive Integrated Moving Average) models for data taken over time were popularized in the 1970s by Box and Jenkins in their famous book. SAS™ software procedures PROC ESM (Exponential Smoothing Models) and PROC UCM (Unobserved Components Models which are a simple subset of statespace models – see PROC SSM) have become available much more recently than PROC ARIMA. Not surprisingly, since ARIMA models are universal approximators for most reasonable time series, the models fit by these newer procedures are very closely related to ARIMA models. In this talk, some of these relationships are shown and several examples of the techniques are given. At the end, the listener will find that there is something quite familiar about these seemingly new innovations in forecasting and will have more insights into how these methods work in practice. The talk is meant to introduce the topics to anyone with some basic knowledge of ARIMA models and the examples should be of interest to anyone planning to analyze data taken over time.

INTRODUCTION

In this paper some relatively recent additions to the SAS/ETS toolkit are introduced. These are PROC ESM for exponential smoothing models and PROC UCM for unobserved components models. This second set of models is a subset of what are called statespace models and SAS provides PROC SSM for dealing with this more general class. Interestingly, the models underlying these procedures are related, sometimes very closely, to the autoregressive integrated moving average (ARIMA) models that have been in popular use in time series analysis for decades, hence the paper's title. This link will be explained with the hope of making these newer procedures understandable to those readers familiar with ARIMA models.

1. NONSTATIONARY ARIMA MODELS.

The notation ARIMA(p,d,q) stands for an AutoRegressive Integrated Moving Average model with p autoregressive lags (lags of response Y after d differences are taken) , and q moving average lags (lags of the white noise error term e). For the purposes this paper, a white noise series e_t is a series of independent identically distributed random variables where t is a time index, t=1,2,3,… A first differenced series is Y_t-Y_{t-1} where Y_t is the response at time t. A first difference is indicated by d=1 in the ARIMA(p,d,q) notation. Anyone familiar with news broadcasts is familiar with differenced series. The up or down numbers for the Dow Jones Industrials Average and the S&P 500 are examples of first differences of the series levels. Differencing the differences, d=2, results in Y_t -2Y_{t-1}+Y_{t-2} as a target series. Differences with d>1 are rare in practical situations but d=1 is fairly common. Seasonal differences of the form Y_t-Y_{t-s}, where s=12 for monthly data, are also common, possibly in combination with ordinary differences. Why do analysts difference series? In the stock market the differences may be of more interest to investors than the levels.

™SAS is the registered trademark of SAS Institute, Cary, NC

From a mathematical standpoint, differencing is used to ensure (1) a constant mean for the resulting series and (2) a covariance between the time t difference and that at time t-j which depends only on the time separation j. These two properties define a condition known as stationarity of the series. Much theory is worked out for stationary series so that getting a series stationary by differencing or otherwise has been the historical approach to analysis.

Here is a graph of U.S. coal futures, western area, from the New York Mercantile Exchange as reported by the U.S. Energy Information Agency and to its right, the series of first differences or changes in the futures price. The series on the left does not appear to have a constant mean, while that to the right seems closer to a constant mean.

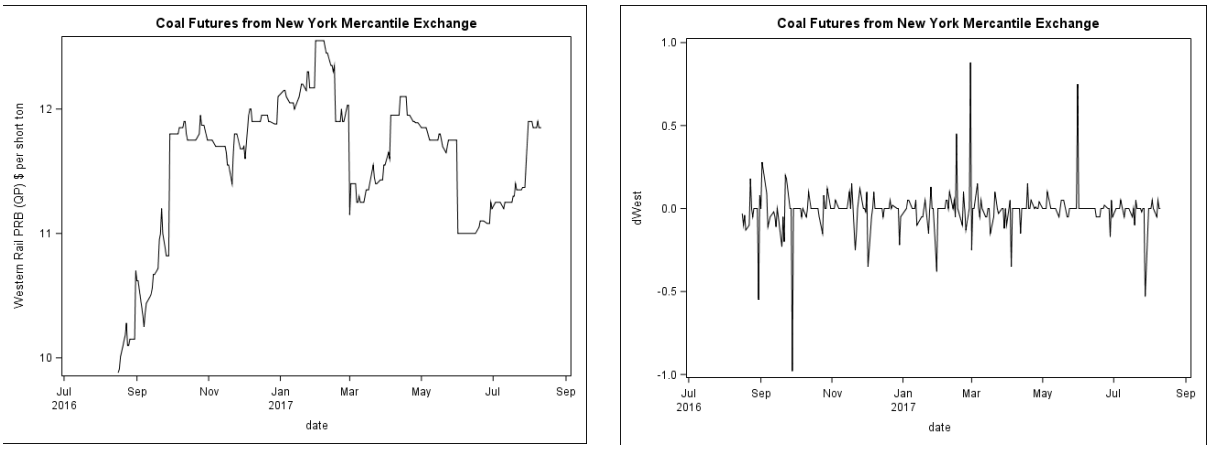

Figure 1. Coal futures prices

The series will be analyzed treating the observations as contiguous even though futures prices are not reported sometimes, weekends for example. A random walk is a series symbolized as ARIMA(0,1,0) and the one step ahead forecast from such a series is just the previous observed value. The model for an ARIMA(0,1,0) is thus

$$Y_t - Y_{t-1} = e_t \text{ or } Y_t = Y_{t-1} + e_t$$

where e_t is a white noise series. Notice that if the left graph in Figure 1 is such a random walk then the right plot is simply e_t versus t. Note also the absence of a mean in the model. There is thus no tendency of the forecasts to return to a historical mean and no reason to use the relative position of the most recent value with respect to the historical mean as a key to investing in a financial product that follows this model. One might question the identical distribution assumption as there are a few unusually large positive or negative values suggesting possibly contaminating outliers or just an unusually heavy tailed distribution.

Based on the model, a random walk forecast predicts the next value from the current one. The forecast one step into the future is thus just the last observed value. Going two steps into the future, there is no observed predecessor value so in the one step ahead forecast, $\hat{Y}_{t+1}$ replaces Y_{t+1}. For j>1, the forecast is a recursion, namely $\hat{Y}_{n+j+1} = \hat{Y}_{n+j}$. Beyond the end of the series, each forecast is just its predecessor, so

the random walk forecast infinitely far into the future is just the last observed value Y_n. The forecast is just a constant. Recursions are the key to the newer procedures detailed here.

A model related to the random walk is the random walk *with drift* β given by

$$Y_t = Y_{t-1} + \beta + e_t .$$

If the observations end at Y_n then it is seen that the forecast one step into the future is

$$\hat{Y}_{n+1} = Y_n + \beta$$

Two steps ahead the forecast is

$$\hat{Y}_{n+2} = \hat{Y}_{n+1} + \beta = Y_n + 2\beta$$

and the j step ahead forecast is, in the same fashion,

$$\hat{Y}_{n+j} = \hat{Y}_{n+j-1} + \beta = Y_n + j\beta$$

The form of the forecast is that of a line with slope β emanating from the last observation. Of course the slope would have to be estimated from the differenced data. Since $Y_t - Y_{t-1} = \beta + e_t$ the best estimate of the slope is the average value of the differences. Since the sum of these differences is just $(Y_2 - Y_1) + (Y_3 - Y_2) + \ldots + (Y_n - Y_{n-1}) = (Y_n - Y_1)$ the estimated slope is $\hat{\beta} = (Y_n - Y_1)/(n-1)$. The forecast is an extension of the line connecting the first observation to the last. See the slanted (green) line in Figure 2.

Models related to random walks are popular, perhaps in part because the random walk model captures the overall pattern of a slowly varying time series. For example, even though the one step ahead forecast is just the preceding value, a plot of the coal futures forecast overlaid on the data series looks at first to do a very good job one step ahead as shown in Figure 2.

In that figure the data are plotted as (red) circles. There is a (blue) line that appears visually to connect the data points, but in fact the blue line is a random walk forecast. Each forecast is just the previous observed value. Despite the apparently excellent forecast through the historical data, each forecast from the random walk is just a short (blue) horizontal line emanating from the previous observation. One reason that the forecast seems so accurate is that the series does not change extremely rapidly in much of the graph and another, more important reason, is that the human eye tends to judge closeness in two dimensions rather than by vertical deviations as the model does. Note that the vertical deviations from the random walk forecast are just the points in the right hand panel of Figure 1. Some are quite large.

The random walk with drift forecast follows the slanted line in Figure 2 after the end of the data which is marked by a vertical reference line. Within the historical data, the one step ahead forecasts add a small drift $\hat{\beta} = 0.0079435$ to the random walk forecast. These differ imperceptibly from the random walk forecasts within the historical data. Despite this historical closeness, beyond the end of the series the linearly increasing forecast and the horizontal forecast become clearly distinct from each other. That drift term changes the forecast into the future dramatically. Finally, if the data were treated (obviously incorrectly) as a mean plus uncorrelated errors, the forecasts would simply form a horizontal line at height equal to the average observed value, indicated by the horizontal reference line stretched across

the width of the graph. Using a mean as the forecast comes nowhere close to reproducing the historical data as does the random walk forecast with or without drift.

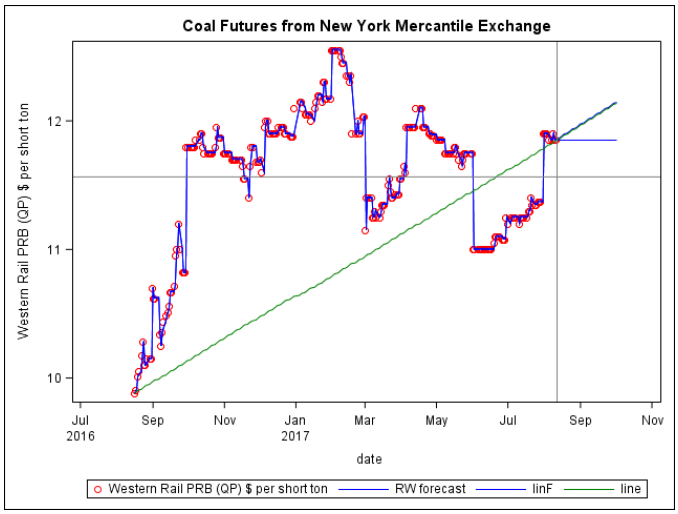

Figure 2. Data with three forecasts – random walk (higher horizontal line segment) simple mean (lower horizontal line) and random walk with drift (slanted line).

Detail: Figure 2 shows a few relatively long almost vertical blue line segments in the historical data. One of these is near the middle of the time period. The point near the top of that line segment is, with a random walk, the forecast of the point at the bottom so the error in the forecast is approximately the length of the line, a fairly large error. It is only because the *horizontal* distances (each being one time unit) from the points to the line segments are small that the forecast appeals to the eye, misleadingly appearing to be excellent. In other words, in a random walk, the one step ahead forecast has the same vertical coordinate as the preceding observation (by definition of a random walk) and is only one time unit to the right. In fact the forecast errors are in the vertical direction. The forecast errors are thus just the differences of the data which appear, as shown in Figure 1, to vary from about -1 to 1.

In summary, the random walk forecast uses the last observation in the data as a forecast into the foreseeable future, ignoring all the data in the past. The alternate assumption of a mean plus uncorrelated errors uses all the data in the series with equal weights no matter how far in the past are the data points, to get a different horizontal line forecast.

2. EXPONENTIAL SMOOTHING FORECASTS

PROC ESM offers several forecasting models depending on whether trends or seasonal patterns exist. Using the coal example to illustrate a simple case, two horizontal forecasts have been described, one that gives 0 weight to the past data and weight 1 to the most recent observation and another, the sample mean, that gives equal weight (1/n) to each of the n data points in a data set. These can be thought of as two extremes on a continuum – bookends so to speak. Why would we use the old data up through about October of 2016? These data seem to represent a temporary past trend that appears to have minimal relevance to the recent data. The random walk forecast downplays past data, in fact it completely ignores all but the most recent observation. Why should we ignore all data points except the last? A compromise might be had by downweighting data from the past with weights that

exponentially decrease as we move further into the past. A prediction formula such as

$$\hat{Y}_{t+1} = \omega \sum_{j=0}^{\infty} (1-\omega)^j Y_{t-j}$$

with 0<ω<1 would do the job since raising a number less than one, like 1-ω, to higher and higher powers, as one goes back into the past, causes early data to receive very small weights and thus to have little influence on the computed simple forecasts. Notice that $\omega \sum_{j=0}^{\infty} (1-\omega)^j = 1$ making the above expression a true weighted average for 0<ω<1. For example, if $\omega = 0.99$ then $\hat{Y}_{t+1} = 0.99 Y_t + 0.99 \sum_{j=1}^{\infty} (0.01)^j Y_{t-j}$ which is close to the random walk forecast Y_t. Computing a forecast in this way is, for hopefully obvious reasons, called exponential smoothing. For this reason, assuming all values of Y prior to the observed series are 0 would not too badly affect the forecast if ω is not too close to 0 and the observed series is long. As shown above, an ω near 1 suggests a very strong discounting of past data and thus the possibility that a random walk forecast, one that that totally ignores all but the most recent data point, might be appropriate. Suppose, in estimating ω, we see the estimates moving toward 1 and being stopped at the software imposed boundary 0.999. This results in a nonconvergence warning. One response based on the above discussion would be to try a random walk forecast.

In contrast, an ω near 0 suggests persistent weights going into the distant past and thus suggests that using the sample mean plus stationary errors might be appropriate. Again, a sequence of estimates approaching 0 and stopped at 0.001 results in nonconvergence and may result when the series is just a mean plus stationary errors.

In Figure 3, the coal futures data are shown along with the two horizontal forecasts just described. The top horizontal line with height equal to the last observation is the random walk forecast and the horizontal reference line extending left to right across the graph is the sample mean. Along with these are plotted five other forecasts using exponential smoothing with exponential decay rate w=(1-ω) = 0.99 (thick cyan line nearest the random walk forecast) 0.985 (thin black line just below the light cyan line), 0.80, 0.40, and 0.10 (nearest to the bottom horizontal line).

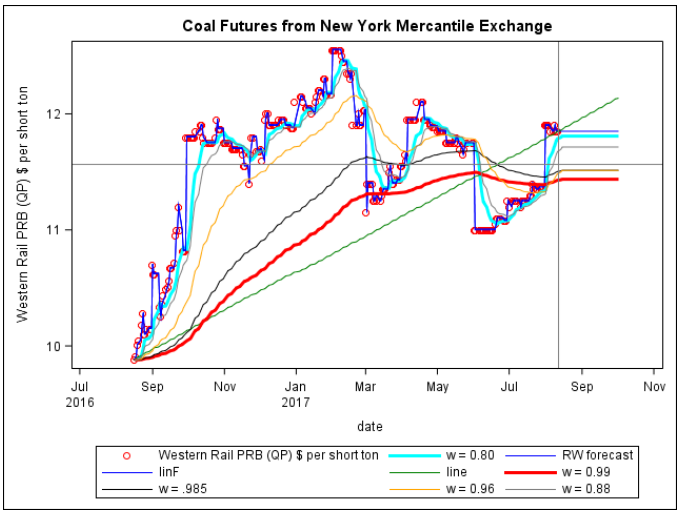

Figure 3. Various coal futures Forecasts.

3. ANOTHER EXAMPLE AND MORE ON WEIGHTS.

Figure 4 shows April snowfall amounts in Denver. When the record showed T (trace amount), 0.05 was substituted, this being the average of 0 and the lowest recorded nonzero amount 0.10. Trace amounts are shown as circles and were more commonly recorded in 1900-1950 than after 1950.

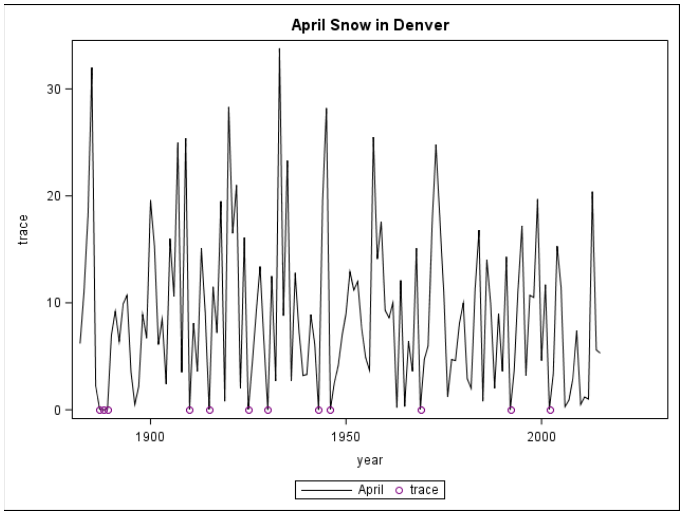

Figure 4. Denver snowfall in April 1882-2015.

In Figure 5 the same plot is shown with two means (ordinary mean 8.92 and weighted average 6.35 with decay rate 1-ω=0.80 being used) and a least squares regression line indicating a decreasing trend in April snow. Should 8.92 or 6.35 or something else be reported as average April snowfall? Is it relevant now to give equal weight to observations carefully made with modern equipment and measurements made only a decade or so after the civil war? Does the linear trend represent what is happening? One might find motivation for downweighting the past based on the fact that modern observations should have less inherent error variability (better accuracy) than those made several decades ago.

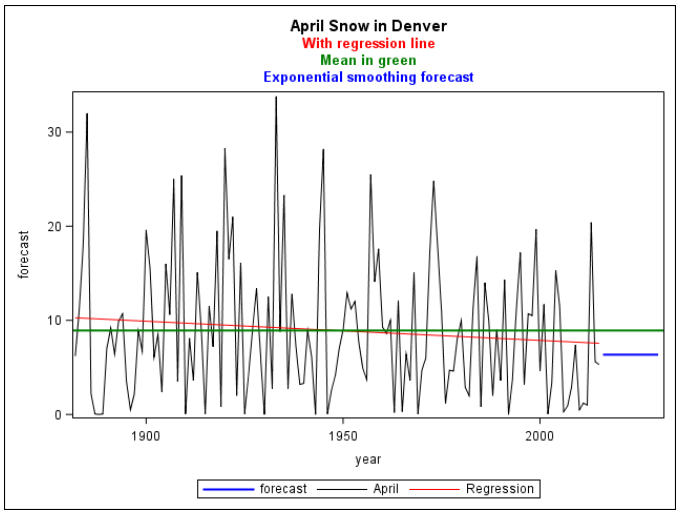

Figure 5. Ordinary and weighted (ω=0.2) means and regression line.

The effect of weighting can be illustrated by adding a right side axis and plot of the weights $\omega(1-\omega)^j$ as in Figures 6 and 7.

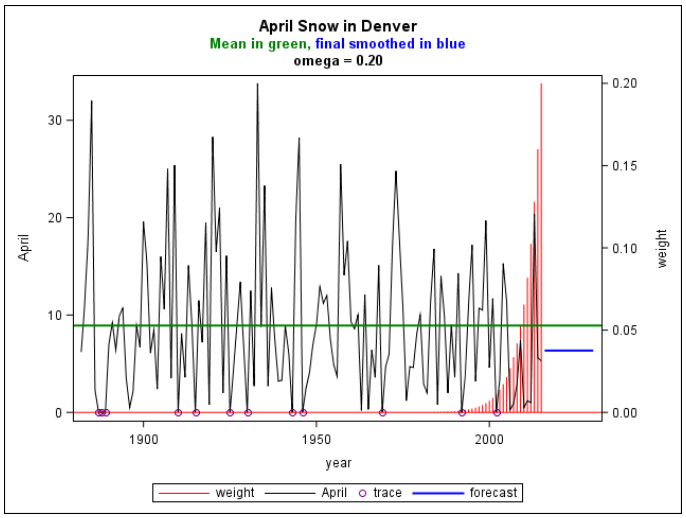

Figure 6. Adding weights $\omega(1-\omega)^j$ to the graph with ω = 0.20 (decay rate $1-\omega$ = 0.80).

A decay rate of $(1-\omega)$=0.8 seems close enough to 1 that it should incorporate several recent years in the weighted average. However, years prior to 1994 or so have weights visually indistinguishable from 0 in Figure 6. Data inspection shows all these weights are less than 1/500 and in fact their total is less than 0.01. Two more weights are illustrated, using omega = 0.02 and 0.60, in the left and right panels of Figure 7. This begs the question of how to choose the weights.

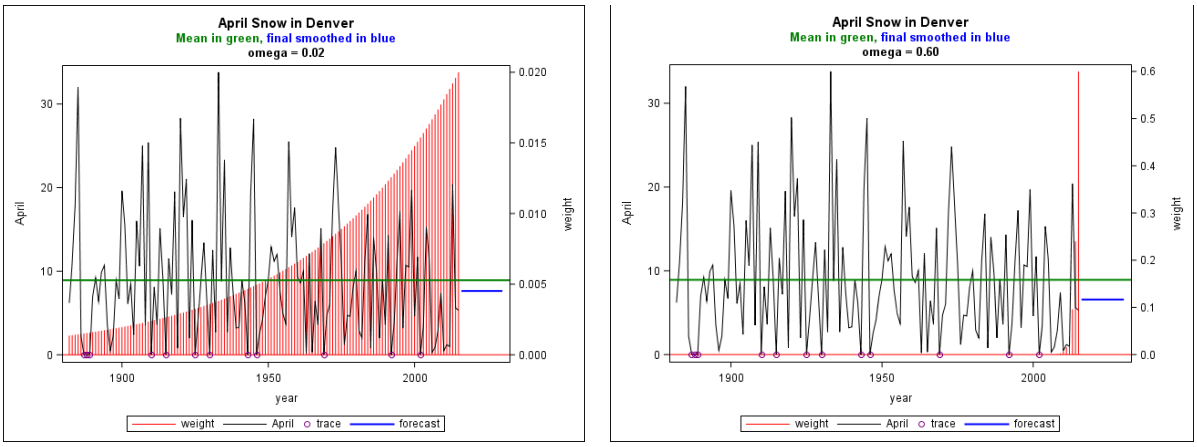

Figure 7. Comparing ω = 0.02 (decay rate 0.98) to ω = 0.60 (rate 0.40).

4. CHOOSING THE WEIGHTS.

Notice that up until now, the exponential smoothing process has been considered as just that – a process that has intuitive appeal and is simple to compute. It has been seen that the process and resulting forecasts can be sensitive to the weights chosen. There has been no discussion of how to estimate and test the weights nor has there been a discussion of forecast standard errors.

Historically, fixed weights for various scenarios were suggested. This evolved into a suggestion to compute error sums of squares for a grid of weights and pick the weight that minimized it. This would be termed estimation using a least squares methodology. If the exponential smoothing process can be viewed as arising from an ARIMA model, which we show it can, then all of the extensive estimation theory for those models can be brought to bear on the weight estimation problem.

As an example, consider forecasting from an ARIMA(0,1,1) or IMA(1,1) given by $Y_{t+1} = Y_t + e_{t+1} - \theta e_t$. We assume 0< θ<1. A sequence of algebraic manipulations allow us to express $\hat{Y}_{t+1}$ as an exponentially weighted average of current and past Y values as follows:

$e_{t+1} = Y_{t+1} - Y_t + \theta e_t$ which holds at all times so $e_t = Y_t - Y_{t-1} + \theta e_{t-1}$ and substituting we have

$$e_{t+1} = Y_{t+1} - Y_t + \theta e_t = e_{t+1} + (Y_{t+1} - Y_t) + \theta\left(Y_t - Y_{t-1} + \theta e_{t-1}\right) = e_{t+1} + (Y_{t+1} - Y_t) + \theta(Y_t - Y_{t-1}) + \theta^2 e_{t-1}$$

Repeated back substitution in this fashion results in

$e_{t+1} = (Y_{t+1} - Y_t) + \theta(Y_t - Y_{t-1}) + \theta^2(Y_{t-1} - Y_{t-2}) + \theta^3(Y_{t-2} - Y_{t-3}) + \ldots$ where the lagged e on the right has disappeared because 0<θ<1 and so θ^k converges to 0 as k increases. One last algebraic rearrangement yields $Y_{t+1} = e_{t+1} + (1-\theta)(Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \ldots)$ and thus the forecast is the same as that of the exponential smoothing method, namely $\hat{Y}_{t+1} = (1-\theta)(Y_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \ldots)$.

Comparing this expression to $\hat{Y}_{t+1} = \omega \sum_{j=0}^{\infty} (1-\omega)^j Y_{t-j}$ we see that they are the same if $\theta = 1-\omega$.

Having seen that an ARIMA(0,1,1) model gives an exponential smoothing model, we have conditional least squares and maximum likelihood available to estimate $\theta=1-\omega$. The likelihood function for the vector D of differenced data with mean 0 is

$$(2\pi)^{-n/2}\left|\Sigma\right|^{-1/2}\exp(-0.5D'\Sigma^{-1}D)$$

where Σ is the MA(1) covariance matrix, a matrix with $1+\theta^2$ in the diagonal, $-\theta$ immediately above and below, and 0 elsewhere.

The left panel of Figure 8 shows data generated as an ARIMA(0,1,1) series with θ = 0.80. The right panel shows the resulting likelihood function for various θ. The maximum likelihood estimate of θ is 0.83. Recall that $\omega=1-\theta$. Note too how much the series on the left changes level, even without a drift term. The maximum likelihood estimate (MLE) of ω, because it is a function of θ, is 1 minus the MLE of θ. The maximum likelihood estimate of ω is thus 1-0.83=0.17.

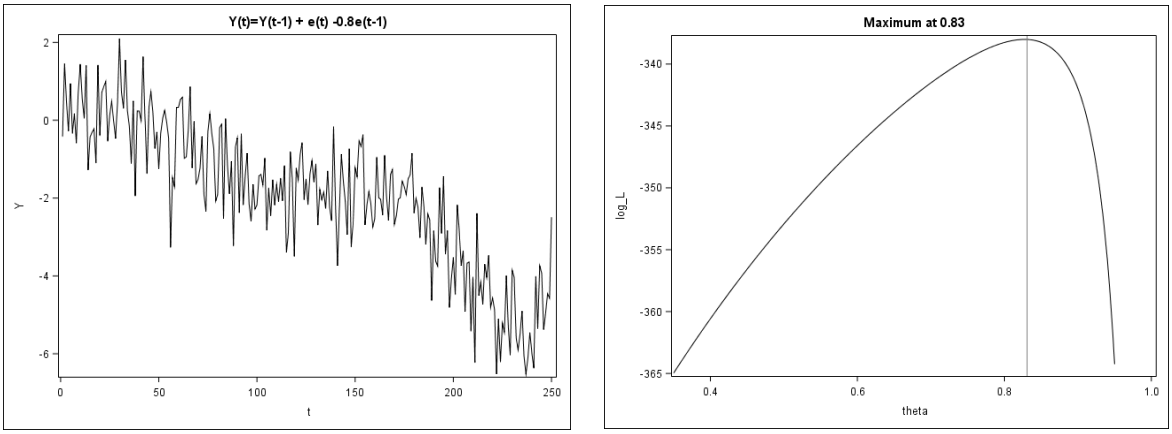

Figure 8. A perfect candidate for exponential smoothing and its likelihood function.

5. EXAMPLE: ATLANTIC TIDES.

The National Oceanic and Atmospheric Administration (NOAA) provides tidal measurements on their web page. Below, in Figure 9, are hourly tides at Wilmington NC from January 21, 2018 through January 26, 2018 with reference lines at each midnight. The typical 2 high tides per day are observed. Predictions are represented with a dashed curve and observations with a solid curve. The NOAA predictions capture the overall seasonality well. Locally there seem to be long stretches of hours in which the forecasts stay over (or under) the actual tides. Can the NOAA predictions be improved locally using exponential smoothing? Given data up to now and the NOAA forecasts of the future, can exponential smoothing improve the forecasts for the next 24 hours or so? Such an improvement is sought in the next analysis.

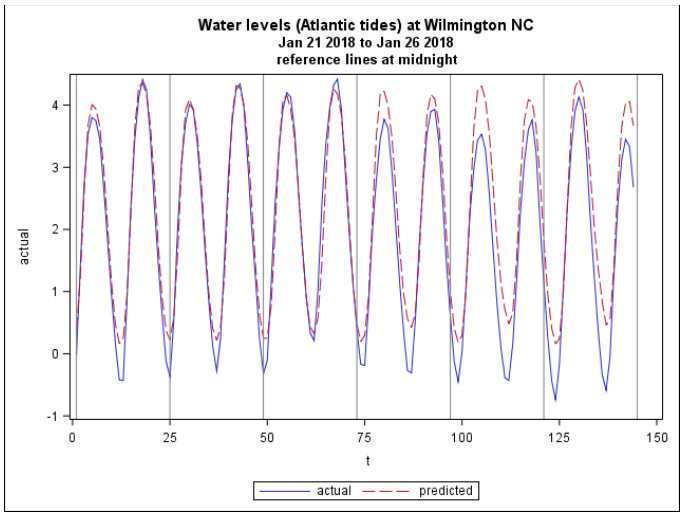

Figure 9. Hourly Atlantic tides January 21-26, 2018.

The differences, actual-predicted, are shown in Figure 10 to the left of the vertical line. It seems there is some seasonality in the prediction errors at least over this period of time.

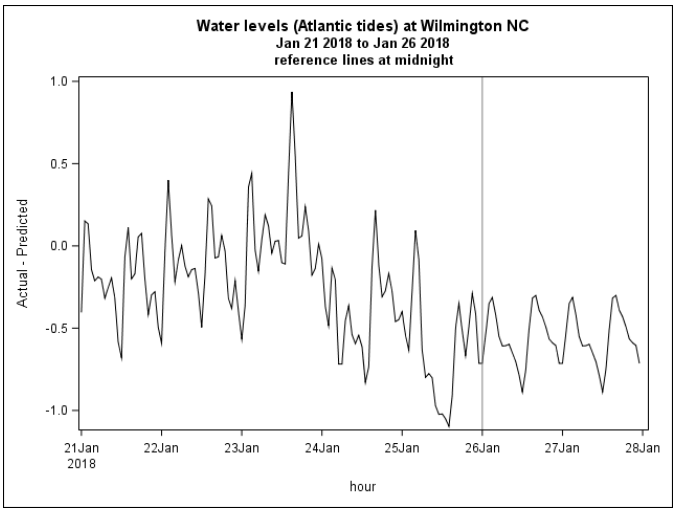

Figure 10. Historical deviations of actual tides from predictions, Wilmington NC (with ESM forecasts).

To the right of the vertical line are the forecasts from a modification known as seasonal exponential smoothing. The data are hourly with period 24 hours. Time is stored as a SAS datetime variable so that PROC ESM "knows" the associated periodicity. The smoothing operation treats all the midnight observations as forming a series to be smoothed, all the 1 a.m. observations as another series etc. up through the 11 p.m. series. Using the same weight for all, the procedure produces 24 smoothed values, one for each hour of the day, used for one day ahead, two days ahead, etc. producing a level forecast with seasonality added. Note that there are two high tides each day. Because these 48 forecasts are all negative, the anticipation is that the NOAA predictions will be too high. In the analysis, the last 24 hours of data were withheld so that an honest assessment of performance will be available.

If there is a trend and seasonality, two methods, Winters' additive and multiplicative methods, are available. The additive method adds seasonal terms to a linear trend forecast where the terms add to 0. The multiplicative method multiplies the trend by seasonal factors that average to 1. Trends with seasonality are often encountered. If the seasonality seems to increase in amplitude as the level increases, the multiplicative method is indicated. If not, use the additive version.

Returning to the tides data, the forecasts of actual – predicted can be added to the future predictions in an attempt to improve the predictions. Figure 11 shows the actual values (red with markers) and two predictions. To the left of the vertical line the predictions are identical to each other and above the actual values. To the right, the blue top curve represents the unadjusted NOAA predictions which are above the actual tides everywhere. The remaining green curve, below the actuals at high tide and above at low tide, are the adjusted predictions which seem closer to the actual tides.

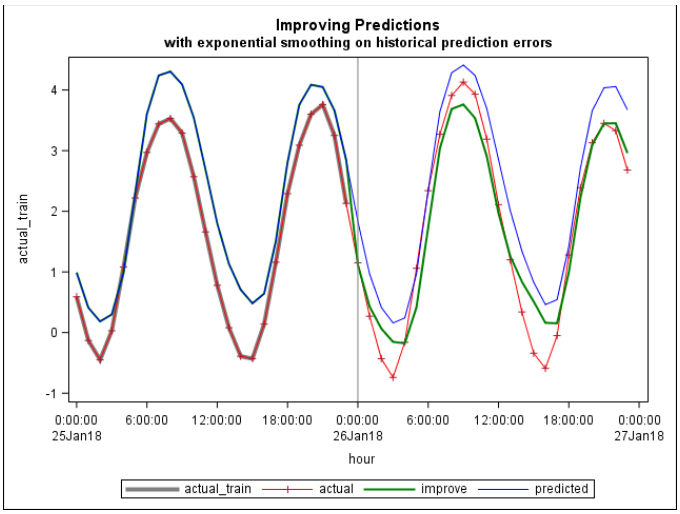

Figure 11. Last day of training data and first day of forecast (validation) data.

To check that the adjustment is really an improvement, the sum of squared prediction errors is computed for the year of withheld data. These are 10.67 and 3.67 for the unadjusted and adjusted predictions. The seasonal exponential smoothing of forecast errors has reduced the prediction error sum of squares to 1/3 of its previous value, a substantial improvement in the short term.

Note also, looking ahead to unobserved components models, that the ARIMA version of the seasonal exponential smoothing model is $Y_t = Y_{t-24} + e_t - \theta e_{t-24}$ and without the error terms, letting $Y_t = Y_{t-24}$ will simply repeat an initial set of 24 numbers periodically into the future. This will be a basic unobserved component as will be a trend produced by a recursion similar to $Y_t = Y_{t-1}+\beta$.

6. SMOOTHING FOR TRENDING DATA.

To finish the discussion of PROC ESM, smoothing of nonseasonal trending data should be discussed. Three methods are available. The first is double exponential smoothing. Single exponential smoothing gives a column of smoothed values, each of which is a forecast of the next observation. We have seen

that the recursion $S_{t+1} = \omega Y_t + (1-\omega)S_{t-1}$ where S is the smoothed series and Y is the observation, gives this column of smoothed values. We can also smooth the smoothed values as $L_{t+1} = \omega S_t + (1-\omega)L_{t-1}$ where L is the double smoothed value. L is used here because the forecast is a linear trend. The equivalent ARIMA model, using the backshift notation B common in ARIMA discussions, is $(1-B)^2Y_t = (1-\theta B)^2 e_t$. Notice that there is no intercept.

Why should the moving average roots be the same? Why should we use the same weights in the first and second smoothing stage? If we allow 2 different weights we have what is known as Holt's method of linear smoothing with the equivalent ARIMA(0,2,2) model being $(1-B)^2Y_t = (1-\theta B)(1-\gamma B)e_t$. Suppose the last 2 Y values in a series are Y_{n-1}= 100 and Y_n= 120 and the forecasts of Y_{n+1} and Y_{n+2} are 125 and 129. The Y value 2 periods ahead involves e values that have not yet occurred so the forecast for period n+3 is just 2(129)-125 = 133, an increase of 4 from 129 and 8 from 125. This is because $(1-B)^2Y_{n+3} = e_{n+3} - (\theta+\gamma)e_{n+2}+\gamma\theta e_{n+1}$. Setting the unknown future e values to 0 this becomes $Y_{n+3}-2Y_{n+2}+Y_{n+1} = 0$ or $Y_{n+3}=2Y_{n+2}+Y_{n+1}$ where these Y values are now replaced by forecasts. Similarly 2(133)-129 is 137, another increase of 4 in the forecast. Continuing in this fashion we get forecasts increasing linearly at rate 4 regardless of whether double smoothing or Holt's method was used to get 125 and 129.

Would a forecast that increases or decreases linearly for all time be appropriate? It might be safer to let the forecast taper off. In terms of ARIMA models, an ARIMA(1,1,2) will show this behavior in the forecasts. In other words, one of the differences (1-B) in Holt's method is replaced with an autoregressive factor $(1-\alpha B)$ with $0<\alpha<1$. As before, the use of a first difference gives a forecast that asymptotes to something that is often not too far from the last observation and certainly removes any tendency to return to the historic mean. The exponential approach to the limit value is at rate α.

The use of exponential smoothing should always involve a "sanity check" obtained by graphing the data and forecasts. To illustrate the dangers of blindly applying exponential smoothing, the three trend models are (clearly inappropriately) applied to the tides data. The code for double exponential smoothing is shown below with obvious changes (method=linear or damptrend) for Holt's method and damped trend smoothing;

```
proc esm data=tides
   lead=&lead out=outD outest = betasD plots=all;
 forecast actual / method=double;
 id hour interval=hour;
run;
proc print data=betasD;
run;
```

Results for these models are below. The common weight for double smoothing, _EST_, has hit the software imposed boundary ω=0.999 as have the two separate weights for Holt's method, rendering the two exponentially smoothed forecasts identical. Forecasts will be like those of an ARIMA (0,2,0) model. The damped trend model appears to have a theoretically equivalent ARIMA(1,1,1) model as one of the implied moving average roots would be 1-0.999 (almost 0).

Forecasting tides
Using double smoothing,lead=24

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	actual	NONE	DOUBLE	WEIGHT	0.999	0.027653	36.1257	9.0334E-74

Forecasting tides
Using Holt's linear smoothing,lead=24

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	actual	NONE	LINEAR	LEVEL	0.999	0.08002	12.4843	1.3209E-24
2	actual	NONE	LINEAR	TREND	0.999	0.12182	8.2005	1.2907E-13

Forecasting tides
Using damped trend,lead=24

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	actual	NONE	DAMPTREND	LEVEL	0.99900	0.14047	7.1121	5.2678E-11
2	actual	NONE	DAMPTREND	TREND	0.99900	0.26862	3.7190	.000287886
3	actual	NONE	DAMPTREND	DAMPING	0.84617	0.05264	16.0761	1.0889E-33

Plots of the double smoothed (Figure 12 left panel) and damped trend forecasts (right panel) follow. Holt's method (not shown) gives exactly the same plot as that of double smoothing for the reasons just explained. If, after seeing all of the active boundary constraints and convergence failure messages, any doubt about the inappropriateness of these methods remains, the graphs should remove it. Vertical lines delimit the last full day of data. The damped trend model is seen from the above table to behave like $(1-B)(1-0.84617B)Y_t=e_t$.

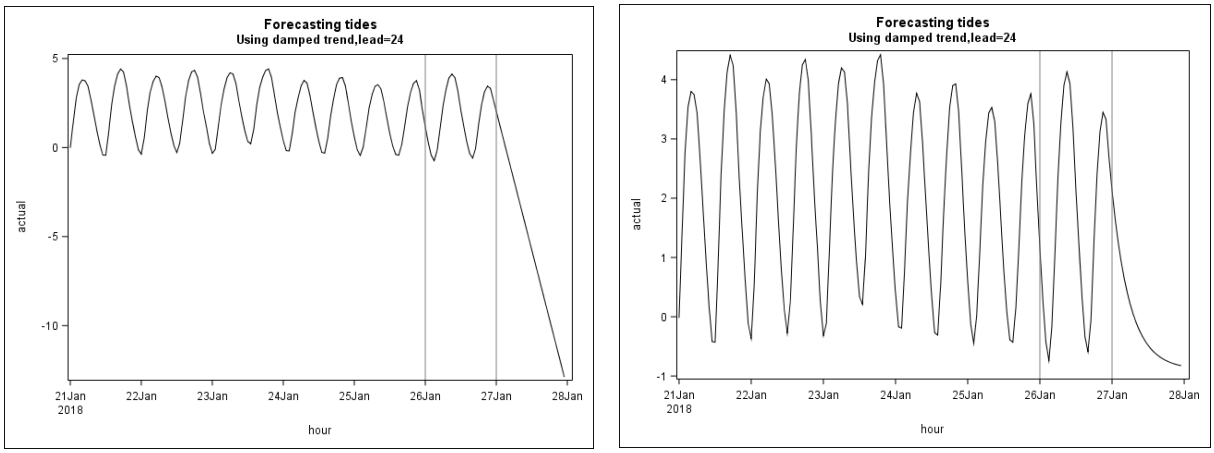

Figure 12. Inappropriate applications of exponential smoothing.

Even a seasonal smoothing model like that used for the prediction errors gives boundary values of the smoothing parameters. Using method=seasonal produces a level parameter estimate 0.999 indicating a level like that of a random walk and thus suggesting a forecast level near the last observation's tide. In contrast the almost 0 seasonal weight suggests seasonal factors that are very regular and thus near the average historical seasonal pattern. These are added to the local level forecast. Here are the results.

Forecasting tides
Using seasonal smoothing,lead=24

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	actual	NONE	SEASONAL	LEVEL	0.999	0.0582	17.1743	0.00000
2	actual	NONE	SEASONAL	SEASON	0.001	47.2954	0.0000	0.99998

As before, the PLOTS=ALL option shows the forecasts which are plotted in Figure 13. Again, the viewer would likely abandon this forecast method based on the graph despite its excellent performance throughout the historical data.

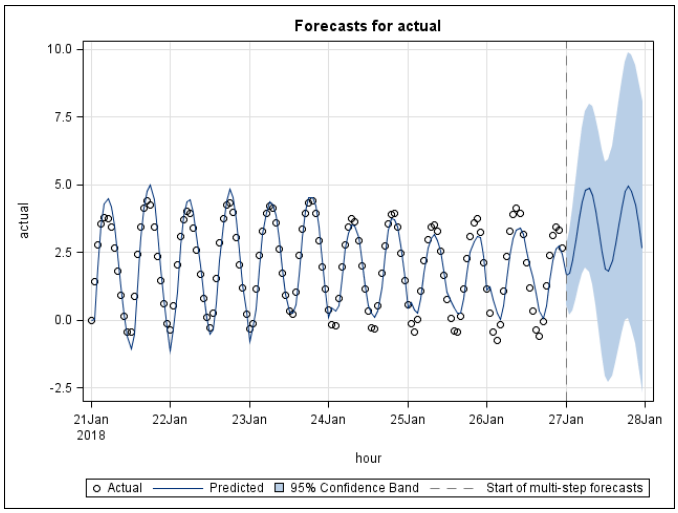

Figure 13. Seasonal exponential smoothing forecasts are unappealing.

As seen before, the rapidly expanding error bands typical of ARIMA models with unit roots are seen. Once again the lesson that a good fit within the historical data does not necessarily imply a good forecast is illustrated. More information, mathematical details, and graphics are available in the (new) third edition of SAS for Forecasting Time Series by Brocklebank, Dickey, and Choi (2018) for these methods, unobserved components, and statespace models.

7. AN APPROPRIATE USE OF LINEAR TREND.

An example that sheds a favorable light on the trending exponential smoothing methods is afforded by a series of yearly U.S. corn yields in bushels per acre (BPA). Prior to the early 1940s, yields were fairly constant as shown in Figure 14. With the invention of effective herbicides and genetic improvement programs, yields increased in a surprisingly linear fashion. For forecasting, the older constant yield data could be completely excluded or could be downweighted by exponential smoothing. To capture the current linear trend one of the recently discussed trending forecast models might be appropriate.

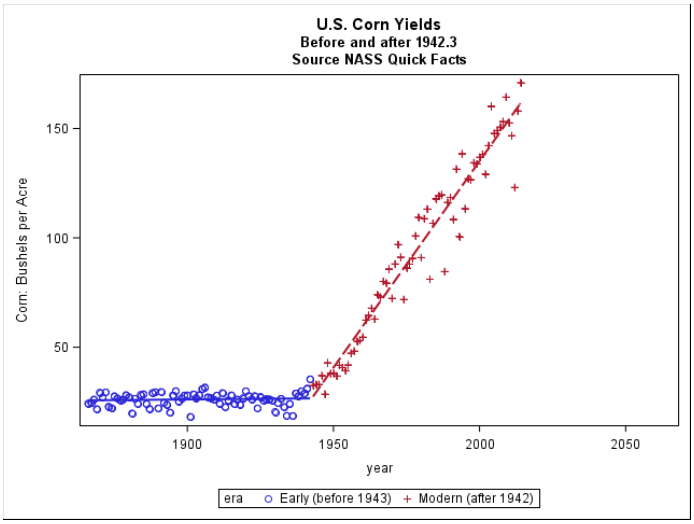

Figure 14. U.S. Corn yields in bushels per acre.

Code for Holt's method ;

```
PROC ESM out=outcorn outest=cornbetas lead=20 plot=all;
 forecast BPA / method=linear;
 id date interval=year;
run;
```

Both parameters are nice - significantly different from 0 and not at all close to 1.

U.S. Corn Yields
Using Holt's (linear) method, 20 years ahead

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	BPA	NONE	LINEAR	LEVEL	0.16878	0.033389	5.05486	.000001262
2	BPA	NONE	LINEAR	TREND	0.14042	0.043678	3.21493	.001604255

The ESM forecast plot in the left panel of Figure 15 shows a forecast that appears little affected by the early era data and in fact is most clearly aligned with data from about 2000 on.

Changing to the damped trend approach (method=damptrend) produced a forecast almost indistinguishable from Holt's linear method and an estimated damping parameter 0.99230 indicating very little change in the slope for several years of forecasts.

U.S. Corn Yields
Using damped trend smoothing, 20 years ahead

Obs	_NAME_	_TRANSFORM_	_MODEL_	_PARM_	_EST_	_STDERR_	_TVALUE_	_PVALUE_
1	BPA	NONE	DAMPTREND	LEVEL	0.16249	0.035483	4.5793	.000009913
2	BPA	NONE	DAMPTREND	TREND	0.15569	0.057960	2.6861	.008066752
3	BPA	NONE	DAMPTREND	DAMPING	0.99230	0.011570	85.7617	8.649E-127

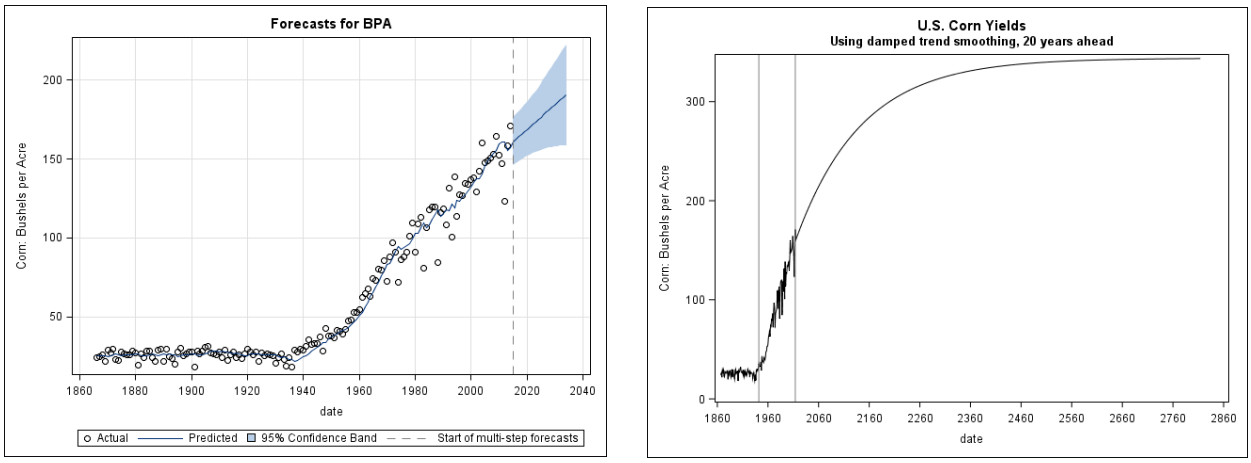

Figure 15. Predictions of U.S. corn yields (BPA) – Holt's method, left, and damped trend, right.

We did not hit the boundary but came very close, indicating that there is almost no curvature in the forecast for several years. It is consistent with the visual impression that corn yields have increased at a surprisingly linear rate in the US. To assure the reader that damped trend was really used, a request for a ridiculous 80 year forecast was issued producing the right panel of Figure 15. The plot is a custom SGPLOT because addition of forecast error bands, as ESM does, gives such a wide forecast error band that the data occupy only a tiny proportion of the vertical axis. A ridiculous yield asymptote is also observed, likely far beyond the amount of corn that can physically occupy an acre of land.

8. INTRODUCING UNOBSERVED COMPONENTS MODELS (UCM).

Exponential smoothing models involve recursive calculations and one or more unit roots in their ARIMA representations. Typically testing for unit roots is not done. In practice a hypothesized second unit root, as Holt's method and double smoothing imply, is usually rejected by hypothesis testing. Because the random walk with drift model is not uncommon in practice, it is likely that the inability of exponential smoothing models to include an intercept in their ARIMA equivalents, implies that the only way to account for (remove) a constant drift term is to difference the series a second time. Thinking of something like a random walk with drift as an alternative can sometimes be helpful when the parameter estimates hit the boundary. The recursive calculations already shown can be used to produce deterministic components in the UCM technology when they include no random innovation term e_t.

The recursion $Y_t=Y_{t-1}+e_t$ describes a random walk. The recursion $Y_t=Y_{t-1}$ with initial value 10 describes a horizontal line – a constant mean 10 for example. Each term is the same as its predecessor. The recursion $Y_t=2Y_{t-1}-Y_{t-2}+e_t$ describes an ARIMA(0,2,0) model whereas $Y_t=2Y_{t-1}-Y_{t-2}$ describes a sequence whose second difference is 0 rather than e_t. With starting values $Y_{-1}=8$ and $Y_0=10$, we find $Y_1=2(10)-8=12$, $Y_2 = 2(12)-10=14$, and in general $Y_t = 10+2t$, a deterministic straight line. The seasonal random walk with period s and initial values Y_{-s+1} … Y_0 becomes just that sequence of s numbers repeated over and over again – a deterministic periodic function – when no error term is present. It is an exactly periodic sequence, the kind of thing handled by seasonal dummy variables in PROC REG or PROC ARIMA.

From trigonometry, we find that $\sin(\omega(t+1))= \sin(\omega t)\cos(\omega)+\cos(\omega t)\sin(\omega)$ while $\cos(\omega(t+1))= \cos(\omega t)\cos(\omega)-\sin(\omega t)\sin(\omega)$. A general sinusoidal wave with amplitude α and phase shift δ can be represented as $\alpha\sin(\omega t+\delta)= \alpha\ (\sin(\omega t)\cos(\delta) + \cos(\omega t)\sin(\delta)$, a linear combination $A\sin(\omega t)+B\cos(\omega t)$ of $\sin(\omega t)$ and $\cos(\omega t)$. Using the trigonometric identities above, these can be moved forward one time unit in a vector recursion:

$$\begin{pmatrix} \sin(\omega(t+1)) \\ \cos(\omega(t+1)) \end{pmatrix} = \begin{pmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega) & \cos(\omega) \end{pmatrix} \begin{pmatrix} \sin(\omega t) \\ \cos(\omega t) \end{pmatrix}$$

The components of the sinusoidal wave at time t+1 are related through this recursion to the components at time t. This gives a way to model seasonality using sinusoids, an alternative to using seasonal lags. With no innovation term e_t, the seasonal lag recursion becomes $Y_t=Y_{t-s}$ where s is the seasonal period. It thus produces a repeating sequence as does the dummy variable approach in regression and is thus called the dummy variable approach.

To summarize, seasonal unobserved components of a dummy variable nature or a sinusoidal nature can be expressed as recursions. As with the trend components a (vector of) normal error terms can be added to allow local perturbations. Expressing the response as a linear combination of such unobserved components, with or without random normal innovations and/or a random normal error term, comprises the basic idea of the unobserved component methodology. With the added random innovations the components correspond to models with one or more unit roots. Even though multiple nonseasonal unit roots are rare in practice, the argument for something like a trend component that involves two unit roots, is that it seems to provide reasonable looking forecasts in practice. Note that this is much like the arguments for the linear exponential smoothing methods such as that illustrated for the corn yield data.

To illustrate these ideas, we start with some generated examples. The constant unobserved component, say C_t, satisfies the recursion $C_t=C_{t-1}+e_t$ when innovation variation is present. Any previously shown random walk illustrates its behavior. Without the random innovation term e_t, C is just a constant for all time. Turning to a trend component, the goal is to express the usual linear trend $\alpha+\beta t$ as a recursion. Following the previously shown logic, the recursion $Y_t=2Y_{t-1}-Y_t$ with initial values $Y_{-1}=\alpha-\beta$ and $Y_0=\alpha$ will do the job.

The STATESPACE and UCM procedures use a model representation known as the statespace representation. At a cursory and simplified level, the methodology involves a "state vector" $\mathbf{Z}_t$ whose elements are combined linearly to relate the state vector to the (vector of) observations $\mathbf{Y}_t$. This is done using an observation equation involving $\mathbf{AZ}_t$. There is also a "transition equation" that relates $\mathbf{Z}_{t+1}$ to $\mathbf{Z}_t$. This is where the recursions thus far studied enter the picture. For a linear trend model we have $Y_t=\mu_t+e_t$ where the time t mean μ_t is $\alpha+\beta t$ and the parameters α and β are constant. Because a linear trend increases by its slope β with each unit increase in time, we see that $\mu_{t+1} = \mu_t + \beta_t$ where $\beta_{t+1} = \beta_t$. Placing μ_t and β_t into the state vector with no innovation variance gives this vector transition equation:

$$\begin{pmatrix}\mu_{t+1}\\ \beta_{t+1}\end{pmatrix}=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix}+\begin{pmatrix}0\\ 0\end{pmatrix}$$

Initial values β_0 and μ_0 set the slope and intercept. The vector of zeros in the above expression could, alternatively, contain innovation errors. This will allow the slope and intercept to vary over time. It adds useful modelling features. We will refer to this as making the slope and intercept local in nature. The observation equation in this example is

$$Y_t = AZ_t + e_t = \begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}\mu_t\\ \beta_t\end{pmatrix} + e_t = \alpha + \beta t + e_t .$$

Is there some reward for making what was initially a simple problem into something so complicated? The answer is yes – we can make the slope or intercept or both local in nature by adding innovation errors to the transition equation rather than setting them to 0. Here is some code that illustrates the situation;

```
%let sigmu=0;
%let sigbeta=0;
%let sigY=3;

Data linear;
*(1) Initialize;
 mu=10; beta=2;

*(2) Transition and observation;
 Title "Statepace data, linear";
 Title2 "Standard deviations: Y &sigY, mu &sigmu, and beta &sigbeta";
do t=1 to 50;
  beta=beta+&sigbeta*normal(93875);
  mu = mu + beta + &sigmu*normal(38715);
  Y = mu + &sigy*normal(2837577);
  output;
end;

*(3) Forecast;
 do t=51 to 65;
  beta=beta;
  mu = mu + beta;
  Y = .;
  output;
 end;

*(4) Graphs;
 ods listing gpath = "%sysfunc(pathname(work))";

proc sgplot;
 scatter X=t Y=Y;
 series X=t Y=mu;
 refline 50 / axis=X;
run;
```

The macro variables allow the incorporation of random normal innovations in the state vector and an error term. In Figure 16, the error variance is 9. The upper left panel has both innovation variances 0.

The upper right has a random intercept innovation (sigmu=2, level becomes local), the bottom left has a random slope innovation (sigbeta=1) and the bottom right has both.

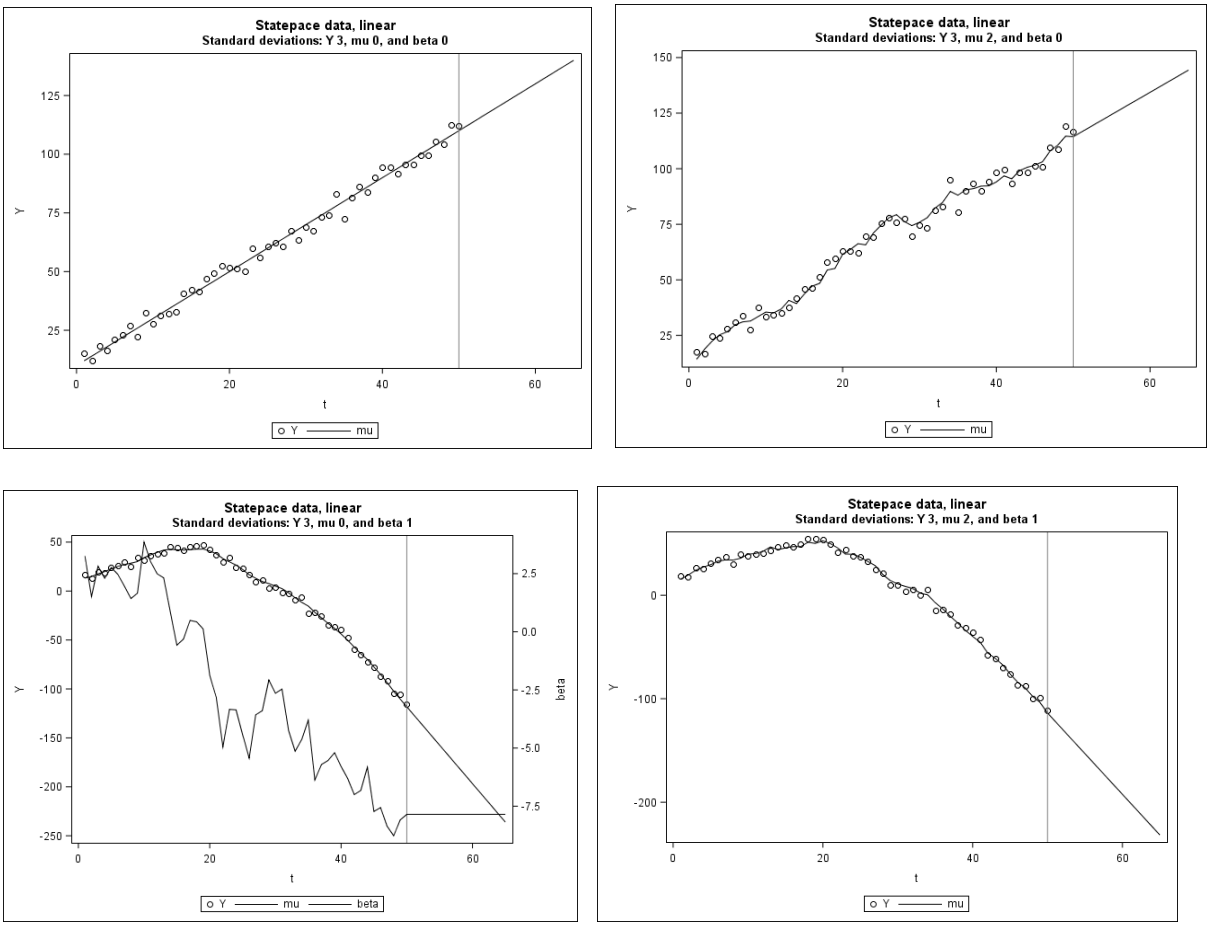

Figure 16: Effects of innovation terms in linear UCM models.

In the bottom left panel, a right hand axis and added plot track the progress of β_t which, from the previous discussion, is a random walk. The final slope around -7.5 dominates the forecast but comparing Y axis labels in the bottom row panels, an innovation effect on μ_t can also be seen. A similar program shows the sinusoidal component with (right panel) and without (left panel) innovation variances in Figure 17.

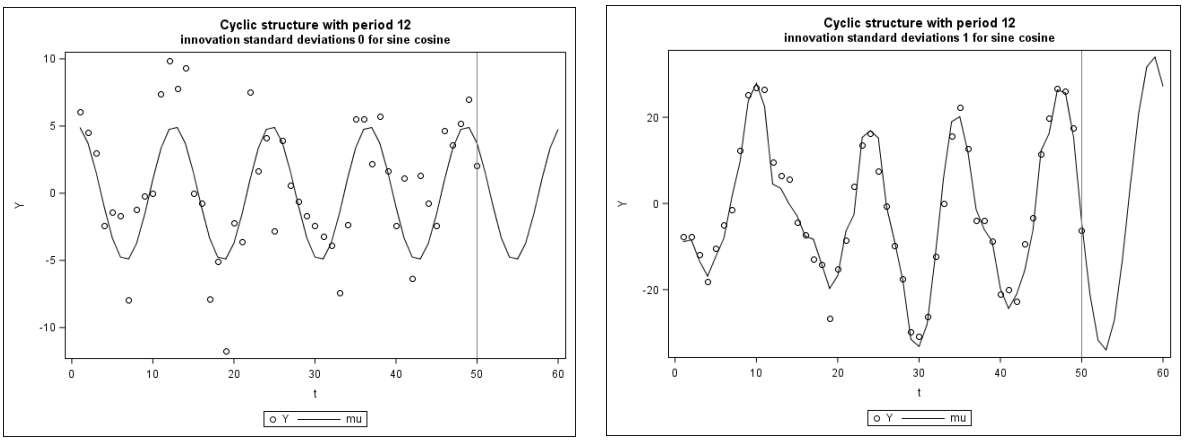

Figure 17. Cyclic component: effect of innovation variances.

The effect of adding a bivariate standard normal innovation to the state vector is a dramatic increase in the amplitude as well as an irregular periodicity in the local level (μ_t).

9. REAL DATA EXAMPLES.

For a real data example, consider the Dow Jones Industrial Average in Figure 18.

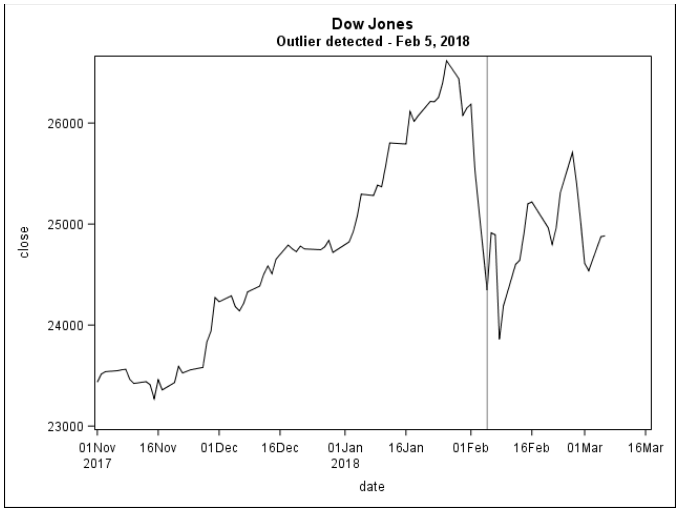

Figure 18. Dow Jones Industrial Average November 1, 2017 through March 6, 2018.

An initial PROC UCM analysis is produced as follows. The 5 day weekday type of date is appropriate here. The seasonal dummy variable approach of period 5 is used;

```
proc ucm data=dow plot=all;
  id date interval=weekday;
  model close;
  level;
  slope;
  season type=dummy length=5;
```

```
  irregular;
run;
```

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	0.82940	3.05842	0.27	0.7862
Level	Error Variance	69907	11126.3	6.28	<.0001
Slope	Error Variance	0.01416	.	.	.
Season	Error Variance	0.02507	.	.	.

The variances for the seasonal and slope components are so small that there are no test statistics available. Removing one variable at a time results in a model with 0 variance for slope, seasonal, and irregular as one might expect from the full model tests, ultimately leading to code without an irregular statement and innovation variances set to 0 except for the level component;

```
proc ucm data=dow plots=all;
  id date interval=weekday;
  model close;
  level;
  slope variance=0 noest;
  season type=dummy length=5 variance=0 noest;
run;
```

This means that the slope and seasonal are deterministic. They are constant over time, global not local. The output has two interesting parts. There is a table indicating a significant innovation variance for the level component. This, in fact, is the only random error term in the model as no irregular statement was included.

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Level	Error Variance	69909	11123.4	6.28	<.0001

The seasonal and slope terms are global because they have no innovation terms. This just means that they are constant, not necessarily 0. The second interesting table shows that they are not just constant, but in fact, there is no evidence that they are nonzero.

Significance Analysis of Components
(Based on the Final State)

Component	DF	Chi-Square	Pr > ChiSq
Level	1	382800	<.0001
Slope	1	0.33	0.5633
Season	4	0.17	0.9966

Setting the slope and intercept terms to 0 by leaving out the associated statements results in a new such table.

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Level	Error Variance	66171	10210.3	6.48	<.0001

What does this say about the Dow Jones Average Y_t? The level μ_t is the only thing left in the state vector and satisfies $\mu_t = \mu_{t-1} + e_t$ with e having an estimated variance 66171. This is a random walk, ARIMA(0,1,0) and is the only component in Y which is consistent with the idea that the Dow Jones Industrial Average is a simple random walk, a commonly held view. It is no problem that there is no irregular component. The innovation variance for level provides the error term in the model. If, in addition, the irregular component were needed, the ARIMA(0,1,0) model would become ARIMA(0,1,1). There is an automatic outlier detector in the procedure. It motivates the reference line in Figure 18.

Outlier Summary

Obs	date	Break Type	Estimate	Standard Error	Chi-Square	DF	Pr > ChiSq
69	05FEB2018	Additive Outlier	-871.11500	181.89361	22.94	1	<.0001

As a second example, consider the monthly U.S. employment series from the Bureau of Labor Statistics (BLS series CEU0000000001) in Figure 19. The series is not seasonally adjusted. Markers denote changes in presidential administrations.

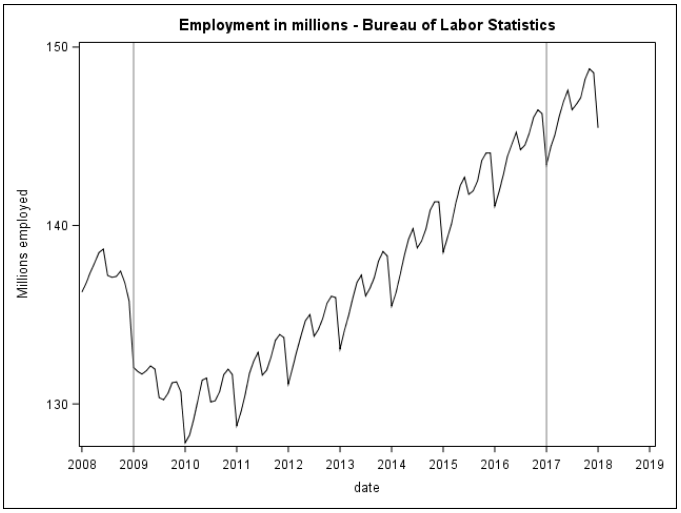

Figure 19. Seasonally unadjusted employment in millions.

Using PROC UCM, level, trend (slope), and seasonal components are requested as well as an irregular term, the term often referred to as an error term;

```
proc ucm data=employment plot=all;
  id date interval=month;
  model employed;
  level;
  slope;
```

```
      /*; season length=12 type=trig keepharmonics= 1;
      season length=12 type=trig keepharmonics= 2;
      season length=12 type=trig keepharmonics= 3;
      title3 "Trigonometric approach"; * */;
  season type=dummy length=12;
  title3 "Dummy approach";
  irregular;
  forecast back=12 lead=48;
run;
```

Notice here that there are several random innovations or error terms. Recall the random level term μ_t from the previous example. There was no need of an additional error term because $Y_t=\mu_t$ seemed to suffice as a model. It is possible even with real data to have 0 irregular variance. In the code above, all components are allowed to have random innovations (the default).

Two seasonal approaches have previously been described. The trigonometric approach is commented out here, but is listed to show the syntax. It keeps 3 harmonics, each with a different innovation variance. Had all 3 harmonics been listed in one statement, it would imply a common innovation variance.

The active seasonal dummy variable approach is the one being used. The seasonal component, then, is $S_t = S_{t-12}+e_{St}$. If there is no innovation variance the seasonal component is an exactly repeating sequence given by $S_t = S_{t-12}$. It is deterministic. The procedure thus gives a way to distinguish slowly varying seasonal random walk behavior from very regular, exactly periodic behavior. The first step is to see which components are local, that is, which of them have significant innovation variance. Our strategy, as before, will be to see which variances can be set to 0 and then which of the resulting deterministic components can be omitted.

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	0.00006670	.	.	.
Level	Error Variance	0.01044	0.0029837	3.50	0.0005
Slope	Error Variance	0.00594	0.0020874	2.84	0.0045
Season	Error Variance	0.00021424	0.0001851	1.16	0.2470

One of the variance components is miniscule and has no standard error or test. Omitting this still leaves the seasonal innovation variance insignificant. This is not too surprising given the very regular seasonal pattern seen in Figure 19. The level and slope are local, having significant innovation variance;

```
proc ucm data=employment plots=all;
  id date interval=month;
  model employed;
  level;
  slope;
  season type=dummy length=12 variance=0 noest;
  title3 "Dummy approach"; * */;
  forecast lead=24 plot=decomp outfor=for;
run;
```

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Level	Error Variance	0.01409	0.0028114	5.01	<.0001
Slope	Error Variance	0.00544	0.0019879	2.73	0.0063

Again, there is no surprise. The slope started negative and became positive. The level is also obviously changing over time. The plots=all option produces many plots. In Figure 20, the left panel shows the local level as a solid blue line. The thick vertical red reference line is at the last observed data point, January 2018.

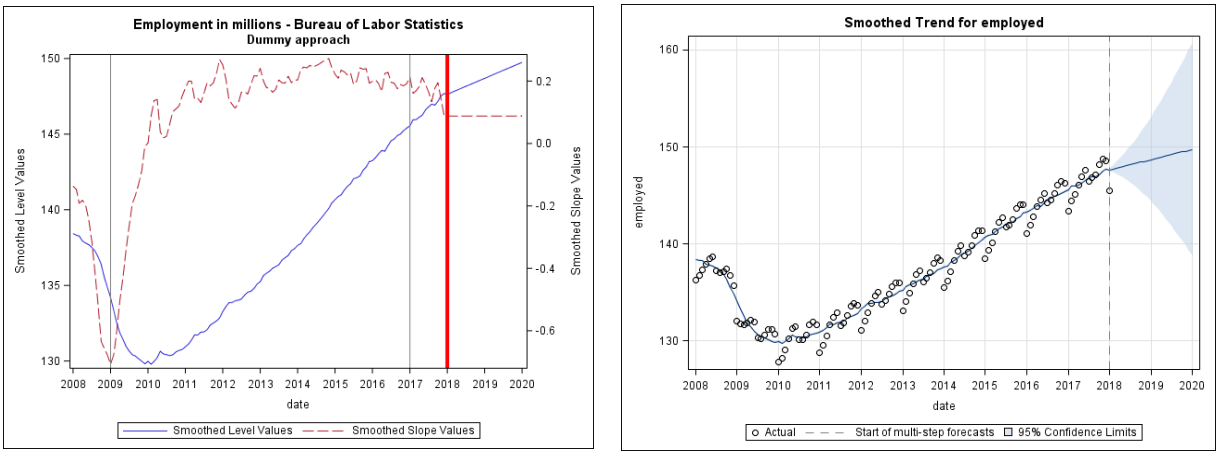

Figure 20. Local level and slope with forecasts (left) and with data and forecast intervals (right).

The blue line is a deseasonalized series. In the last few months the local slope, the red broken line, has dropped a bit, especially in the last month, January 2018 where the right axis shows the slope to be approximately 0.09. To the right of the thick vertical red line, the forecast of the slope is the final, January 2018, smoothed slope so the dashed red line becomes horizontal in the future. This constant slope results in a linear forecast of employment, in blue, for the next 2 years. The right panel is a similar plot produced by the UCM procedure. The extremely fast widening of the error bands is typical of the double unit root nature of models with a slope and level component. The similarity of the yearly deviations of the actual data around the local level shows the aforementioned regularity of the seasonal data component and lack of an irregular component. A table of final (January 2018) values adds some details.

Trend Information (Based on the Final State)

Name	Estimate	Standard Error
Level	147.633598	0.0546661
Slope	0.088055639	0.1115646

The deseasonalized forecast emanates from 147.63 (left axis of Figure 19) and increases at rate 0.088 (right axis, less than 1 standard error above 0) per month for a rise of 24(0.088) = 2.11 over the 2 year

forecast. The blue line forecast rises from 147.63 to 149.74. A table of test statistics for these two January 2018 state vector components along with the seasonal components shows insignificance of the slope, as expected, and significance of the local level and the global (no innovation), regular seasonal components.

Significance Analysis of Components
(Based on the Final State)

Component	DF	Chi-Square	Pr > ChiSq
Level	1	7293473	<.0001
Slope	1	0.62	0.4299
Season	11	7430.08	<.0001

Note that $(0.08805/0.11156)^2 = 0.62$ relates the two tables to each other. Perhaps a level forecast with seasonality added, obtained by omitting the slope statement, would be acceptable. Finally, a forecast plot that includes the seasonality and a panel with the smoothed seasonal are shown in Figure 20. The rapidly expanding forecast intervals typical of these unit root models are evident. The excellent fit in the historical data belies the large uncertainty in the forecasts as often happens with these models.

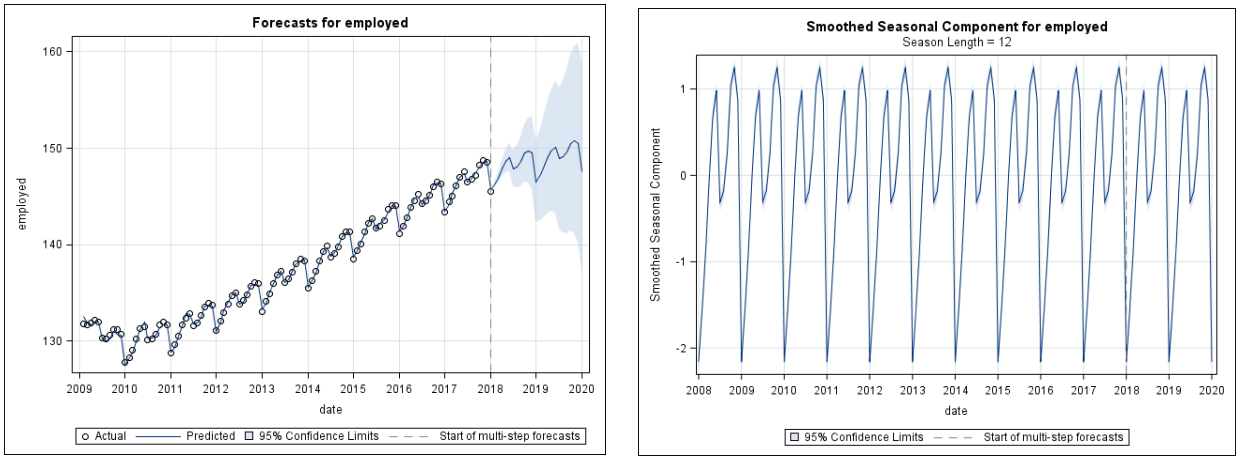

Figure 20. Employment forecast with error bands, left, and seasonal component, right.

As a last example, we return to April snow in Denver. Specifying level, slope, and irregular terms gives

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	56.74810	7.01443	8.09	<.0001
Level	Error Variance	5.798252E-7	0.0004164	0.00	0.9989
Slope	Error Variance	0.00002466	0.0001352	0.18	0.8552

Imposing 0 variance on the slope and level terms sequentially in the way previously shown suggests treating both as deterministic. With that restriction, the next table suggests that the slope can be

omitted. Based on the estimates from the final state vector, the level is not 0 (of course) but the slope is not significantly different from 0, consistent with the idea that the global trend line is just horizontal, no trend in April snowfall. The irregular term for the last observation is not significantly different from 0, an uninteresting result.

Significance Analysis of Components
(Based on the Final State)

Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	3.05	0.0810
Level	1	34.04	<.0001
Slope	1	1.46	0.2268

Omitting the slope statement and restricting the level to be constant (0 variance) the model is just a mean plus irregular. The error variance around this line is about 57. The irregular term for the last observation is near 0 but clearly the error variance is greater than zero.

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	56.89656	7.00348	8.12	<.0001

A possible follow up would be to run an ARMA(p,q) model with just a mean and check for autocorrelation. The UCM models assume none, but it would be prudent to check. While we have carefully checked for more complex UCM structure, we have arrived at a simple mean model with estimated mean being just the simple average 8.92 as would be given by PROC MEANS.

Trend Information (Based on the Final State)

Name	Estimate	Standard Error
Level	8.915671642	0.6527429

10. UCM SUMMARY AND LESSONS LEARNED.

In summary, the UCM procedure has some similarities with ARIMA modelling. Many of the underlying models have unit roots and thus forecast error bands can spread dramatically as forecasts move into the future, despite good historical performance. Forecasts often appear very reasonable and the procedure allows the user to distinguish deterministic from random, slowly changing effects. Nice decompositions of series, accompanied by informative graphs are produced as are output data sets for custom graphics and tables.

REFERENCE.

Brocklebank, J., D. A. Dickey, and B. Choi (2018) SAS for Forecasting Time Series. 3rd edition SAS Institute, Cary NC.

CONTACT INFORMATION.

dickey@stat.ncsu.edu