Paper 339-2012

Analyzing the time series of U.S. E-Commerce using Proc Ucm

Anders Milhøj, Department of economics, University of Copenhagen, denmark

ABSTRACT

In this paper Proc Ucm is applied in an analysis of the series of E-Commerce which is published by the U.S. Census Bureau. This rather new procedure decomposes a time series into intuitive components such as levels, trends and seasonality which are easily specified in the SAS® code. The advantages of Proc Ucm as an easy-to-use alternative to a careful econometric analysis will be of focus in the presentation. The results are mainly presented by the graphical output which adds to the attractiveness of the procedure. Also more advanced features of Proc Ucm will be applied such as a discussion of how the seasonality have changed and various ways to include the total retail sales as an independent variable in the model.

INTRODUCTION

In this paper Proc Ucm is applied in an analysis of the series of E-Commerce as published by U.S. Census Bureau. Data is available as a quarterly time series from 1999.4Q, http://www.census.gov/retail/index.html. The series is graphed in Figure 1.

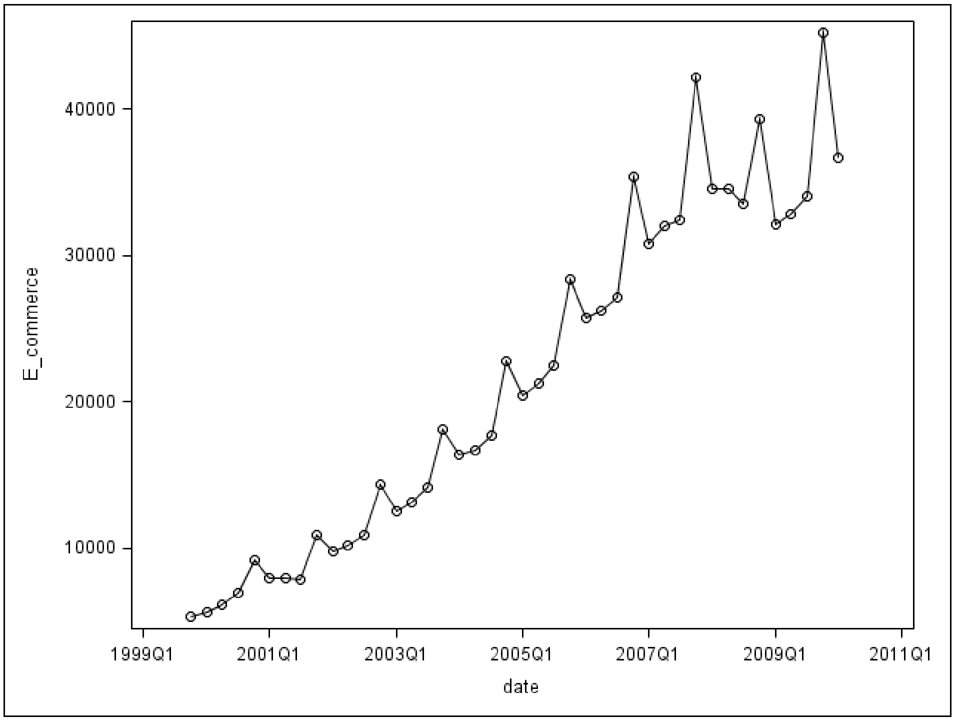


Figure 1 The series of US E-commerce, mill. $.

This presentation of Proc Ucm is intended to demonstrate how easy the procedure is coded and to illustrate the attractive graphical output produced by the procedure. But also more advanced modeling abilities of the procedure will be demonstrated such as the use of independent variables to explain for some of the variation in the series of interest.

In the first application only a simple level-trend model will be fitted in combination with a seasonal component.

UNOBSERVED COMPONENT MODELS

In its simplest form a time series, called y_t, is decomposed into a sum of a level term and a remainder term by

$$y_t = \mu_t + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2).$$

The level component μ_t is assumed to be generated by

$$\mu_t = \mu_{t-1} + \eta_t \qquad \eta_t \sim N(0, \sigma_\eta^2)$$

starting up at an initial value μ_0 at time index t = 0 just before the first observation. The level series μ_t could be extended to include a trend by the definition

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \qquad \eta_t \sim N(0, \sigma_\eta^2)$$
$$\beta_t = \beta_{t-1} + \xi_t, \qquad \xi_t \sim N(0, \sigma_\xi^2) .$$

Positive values of β_t correspond to an upward drift in the data series but the actual trend is time dependent as the actual slope β_t by the formula above is allowed to vary. The residual series η_t, ε_t and ξ_t are all assumed to be mutually independent white noise series, meaning that they each consist of identically distributed independent stochastic terms. Their variances give an idea of the stability of the components, as e.g. the value $\sigma_\xi^2 = 0$ gives a model with a constant trend while larger values of σ_ξ^2 allow the trend to fluctuate. Seasonal dummies could be included by

$$y_t = \mu_t + S_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$
$$S_t = - (S_{t-11} + .. + S_{t-1}) + \omega_t \quad \omega_t \sim N(0, \sigma_\omega^2).$$

In addition to these unobserved components independent variables in form of linear regressions could be included in the model. As a further refinement the regression coefficients are allowed to vary by

$$y_t = \mu_t + \gamma_t x_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$
$$\gamma_t = \gamma_{t-1} + \zeta, \qquad \zeta_t \sim N(0, \sigma_\zeta^2).$$

By the Kalman filter and more advanced smoothing algorithms all these components could be estimated from data by Proc Ucm together with the variances of the remainder terms of the components. The essential output consist of plots of the estimated components and numerical values for the variances.

A MODEL WITH A LEVEL, A TREND AND A SEASONAL COMPONENT

The three components are specified by the `level`, the `slope` and the `season` statement .

```
proc ucm data = sasts.E_commerce;
      id date interval = quarter;
      model E_commerce;
      level plot=smooth checkbreak;
      slope plot=smooth;
      season length = 4 plot=(smooth s_annual);
      outlier;
      estimate plot=(panel);
      forecast lead=24 plot=forecasts alpha=0.1;
run;
```

The results are somehow disappointing as the slope component, Figure 2, is rapidly changing during the financial crises in 2007-2010 as the slope changes from being significantly positive to be significantly negative during only two quarters. This explains for the whole effect of the financial crises in the estimated level, Figure 3. This is not what is naturally expected as a trend intuitively has to be persistent and even if it could change over time only small variations are allowed from one quarter to the next quarter. In other words the estimated variance of the slope component is too high. This problem could be mended by specifying a lower value for the slope component variance.

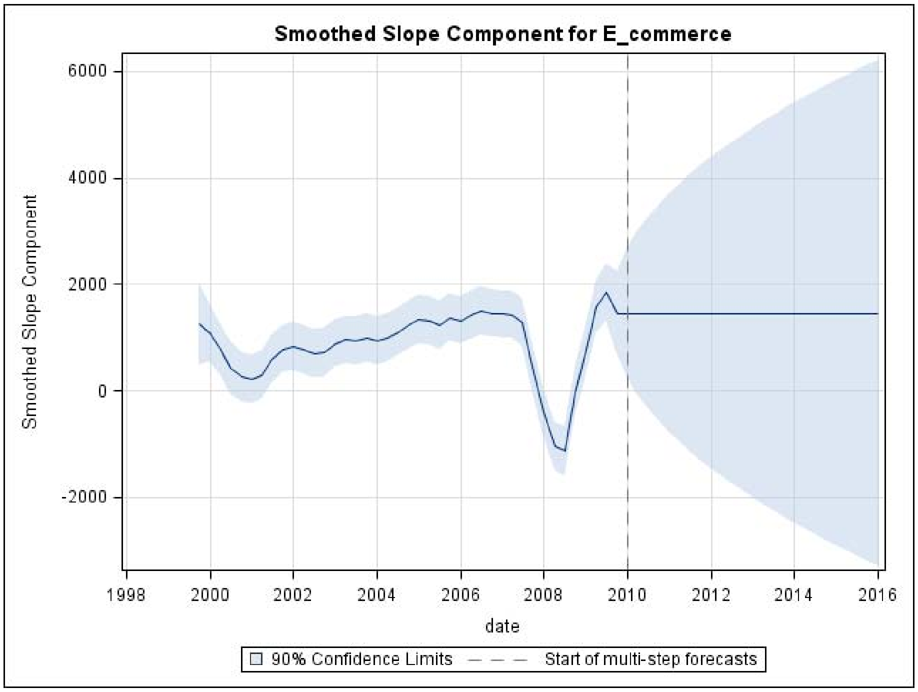


Figure 2 The estimated slope component.

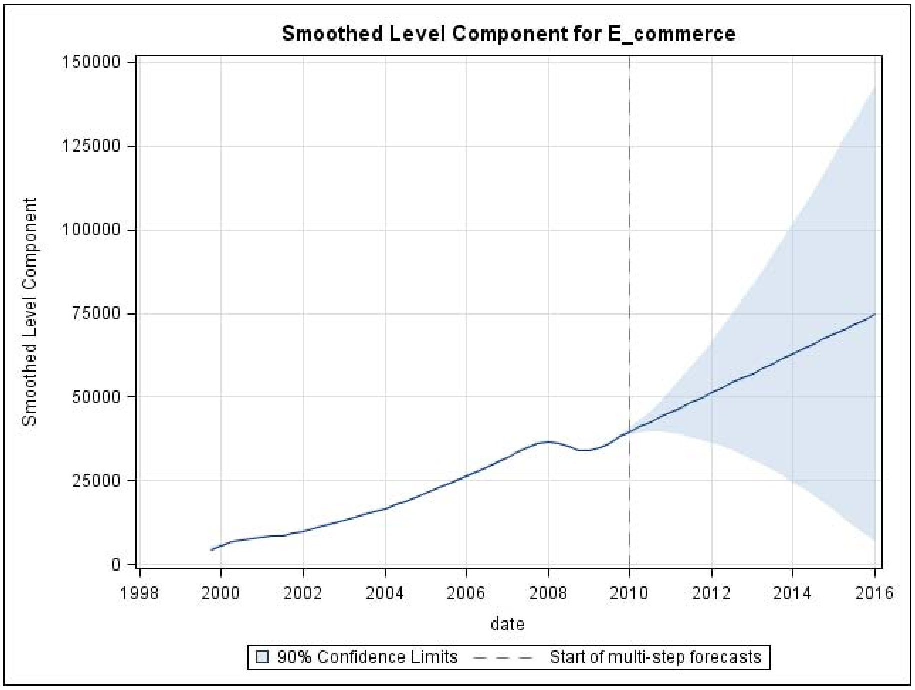


Figure 3 The estimated level component.

```
                          Final Estimates of the Free Parameters

                                                           Approx                 Approx
     Component         Parameter          Estimate       Std Error    t Value    Pr > |t|

     Level             Error Variance   5285.26022        526162.5       0.01      0.9920
     Slope             Error Variance       324902        291936.4       1.11      0.2657
     Season            Error Variance       203445         86369.6       2.36      0.0185


                            Fit Statistics Based on Residuals

                       Mean Squared Error                    2154291
                       Root Mean Squared Error            1467.75038
                       Mean Absolute Percentage Error        4.41064
                       Maximum Percent Error                11.28691
                       R-Square                              0.98121
                       Adjusted R-Square                     0.98010
                       Random Walk R-Square                  0.87429
                       Amemiya's Adjusted R-Square           0.97789
```

The variances of the slope and level components are both insignificant. The fit is almost perfect as $R^2 = 0.98$. This gives a clear indication that the model is over parameterized leading to a too perfect fit for which the model tells nothing more than the observed series alone.

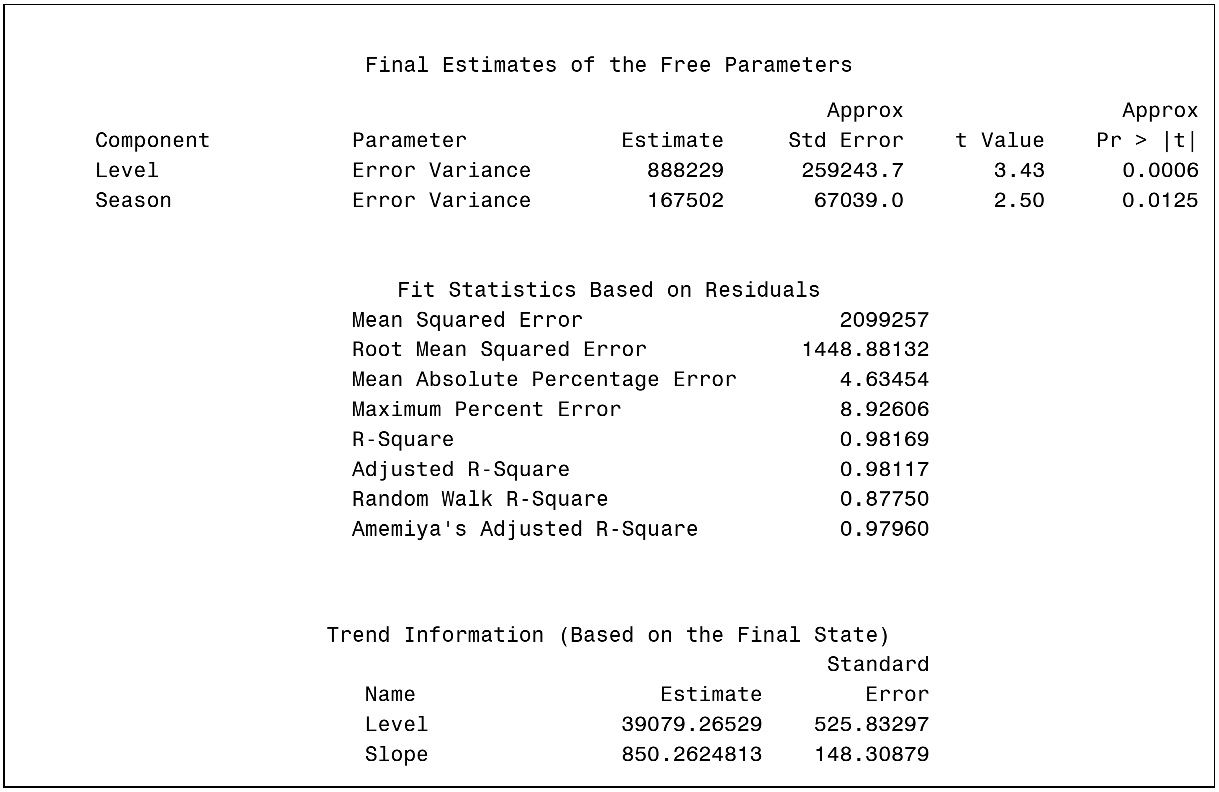
One possible remedy is to fix one of the variances in the numerical estimation algorithm. In the statement below the slope is fixed at a constant value as the innovation variance, `var=0`, for the slope process is fixed by the option `noest` as zero. In this way a model with a fixed linear trend, however at varying intercepts, is included for the volume of E-Commerce.

```
      slope plot=smooth var=0 noest;
```

This could be a reasonable possibility for these data, but it also possible that a specification of a positive variance, but less than the estimated value above would perform even better. A variance at 100000 which is about one third of the previously estimated variance is specified by

```
      slope plot=smooth var=100000 noest;
```

For this specification the parameter estimates seem reasonable and the fit is not much worse. The level plot, Figure 4, more precisely mimics what intuitively could be the underlying tendency of the growth in E-Commerce with a more clear U-shape around 2009. Moreover the confidence interval in the forecasting period is much smaller than in Figure 3 when no trend was incorporated in the model. The forecast plot, Figure 5 presents a reasonable fit in the observation period and also the forecasts some years ahead seem acceptable even if a fixed linear trend in the infinite future is impossible.



```
                    Final Estimates of the Free Parameters

                                                      Approx                   Approx
Component          Parameter          Estimate      Std Error     t Value    Pr > |t|
Level              Error Variance       888229       259243.7        3.43      0.0006
Season             Error Variance       167502        67039.0        2.50      0.0125


                        Fit Statistics Based on Residuals
                   Mean Squared Error                      2099257
                   Root Mean Squared Error              1448.88132
                   Mean Absolute Percentage Error          4.63454
                   Maximum Percent Error                   8.92606
                   R-Square                                0.98169
                   Adjusted R-Square                       0.98117
                   Random Walk R-Square                    0.87750
                   Amemiya's Adjusted R-Square             0.97960



                 Trend Information (Based on the Final State)
                                                      Standard
                    Name                 Estimate        Error
                    Level             39079.26529    525.83297
                    Slope             850.2624813    148.30879
```

The slope is constant and the estimated value, 850, means that the value of E-Commerce increases 850 mill $ steadily each quarter adjusted for seasonal variation. This trend is due to inflation and economic growth but it also reflects the raising importance of E-Commerce during the observation period.

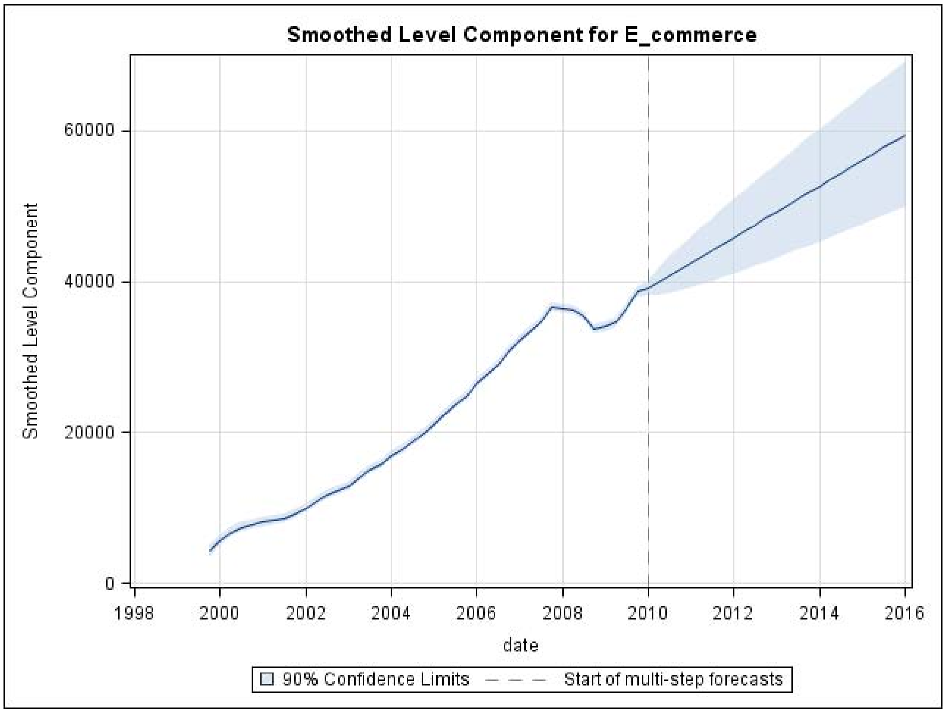


Figure 4 Smoothed level component

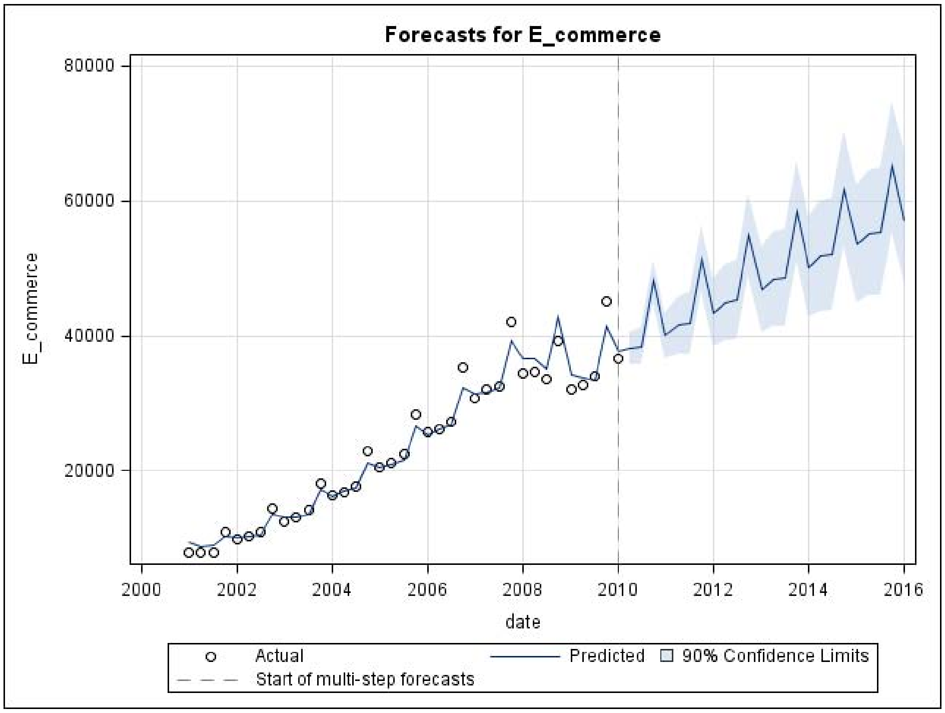


Figure 5 Forecast plot.

The option `plot=(smooth s_annual)` in the `season` statement specifies that the seasonal component is printed and that the dummy variables for each quarter are plotted against the year in order to see how the seasonal structure evolves over the years. For both plots the smoothed version is specified - if the filtered versions are wanted the option should be changed to `plot=(filter f_annual)`. These plots are given in Figure 6 and 7 as each plot only gives plots for two quarters. In the data set the first observation is for the E-Commerce in fourth quarter of 1999 and as this is the first estimated dummy it is denoted called seasonal factor one even if it is for the fourth quarter and it is plotted as the first diagram of the two plots in Figure 6. The dummy variable for fourth quarter have increased rapidly from 2000 to 2010 which may be seen as a sign that more and more Christmas gifts are bought over the internet by common American families, while in the early years E-Commerce presumable mainly was used for other kind of goods and perhaps not by typical American families. The seasonal dummies by definition sum to nearly zero so the fact that the fourth quarter dummy increases of course has the consequence that the other three dummies decreases.

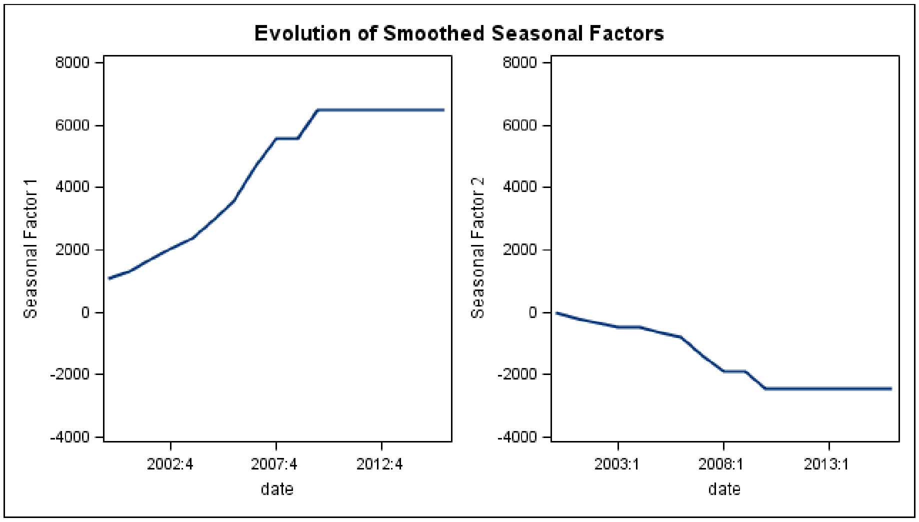


Figure 6 Seasonal factors for fourth and first two quarters.

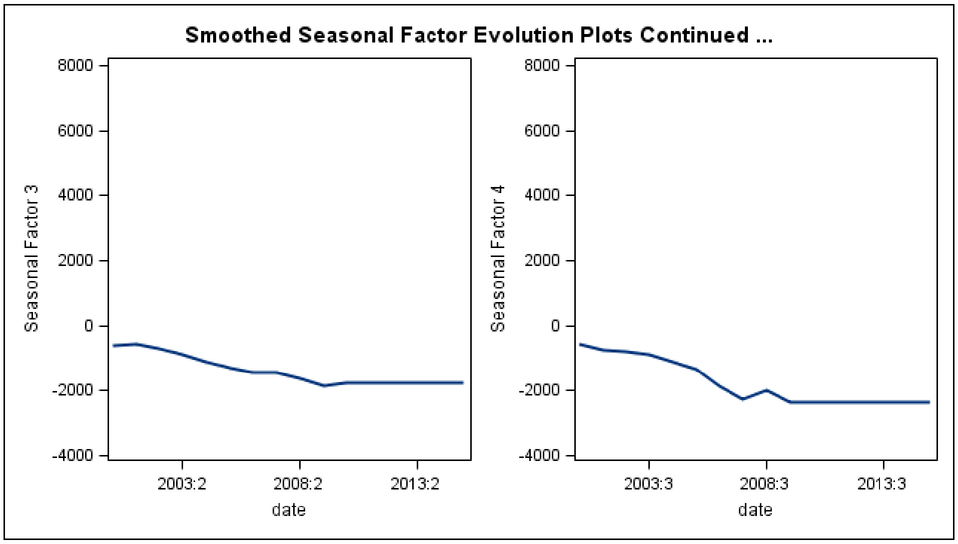
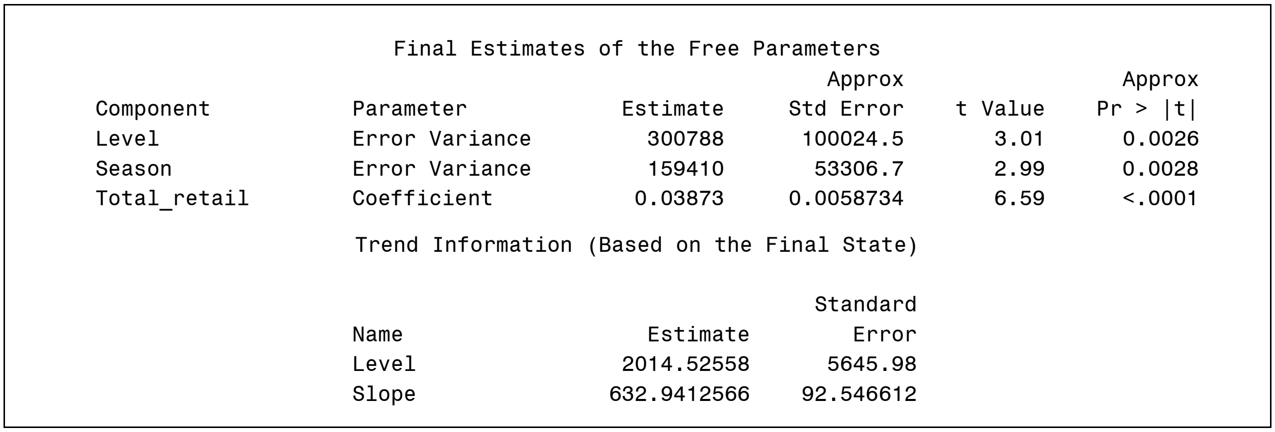


Figure 7 Seasonal factors for second and third quarters.

INCLUSION OF AN INDEPENDENT VARIABLE

In the following application the various possibilities offered by Proc UCM for inclusion of regression terms in the model will be demonstrated. This is done by inclusion of the total retail sales as an independent variable when modeling the volume of E-Commerce. In a first attempt the independent variable is included with a constant regression coefficient as in usual linear regression models. The independent variable is simply specified in the model statement. Of course more than just a single independent variables could be included by the same syntax as other regression procedures in SAS®, e.g. Proc Reg.

```
proc ucm data = sasts.E_commerce;
      id date interval = quarter;
      model E_commerce=Total_retail;
      level plot=smooth;
      slope plot=smooth var=0 noest;
      season length = 4 plot=smooth;
      outlier ;
      estimate plot=(panel residual);
      forecast lead=24 plot=forecasts alpha=0.1;
run;
```



```
                          Final Estimates of the Free Parameters
                                                              Approx                   Approx
Component             Parameter              Estimate      Std Error    t Value     Pr > |t|
Level                 Error Variance           300788       100024.5       3.01       0.0026
Season                Error Variance           159410        53306.7       2.99       0.0028
Total_retail          Coefficient             0.03873      0.0058734       6.59       <.0001

                      Trend Information (Based on the Final State)

                                                           Standard
                      Name                   Estimate         Error
                      Level                2014.52558       5645.98
                      Slope               632.9412566     92.546612
```

The estimated regression coefficient, 0.039, is clearly significant as seen in the table of parameter estimates. It tells that the volume of E-Commerce corresponds to about 4% of the total retail value. The slope component is estimated as the constant 633 which is a bit below the value 850 in the model without the independent variable. As this slope component in this model describes the increasing importance of E-Commerce the regression only describes the part of E-Commerce which is proportional to the total retail sales. This part could be due to inflation, economic growth and other factors which affects the two series the same way. The seasonal component in this model specification allows for different seasonal structures of E-Commerce and total retail sales. Similarly the significant upward trend corresponds to the more rapid increase in E-Commerce compared to total retail sales.

Another way of letting the model take into account the fact that E-Commerce in the observation period has increased its importance relative to the value of the total retail sales is to allow the regression coefficient to be time varying. This is obtained by application of the `randomreg` statement as an alternative to the fixed regression specified in the model statement in the previous code. In this specification it is natural to exclude the slope component and moreover the level is fixed by setting its variance to zero in order to avoid over parameterization. The trend is then only described by the independent variable, the total retail sales, which is upward trending and the increasing importance of the E-commerce is now described by the value of the time varying regression coefficient so the trend of E-Commerce as the dependent variable could be more steep than the trend of total retail sales, the independent variable.

```
ods graphics;
proc ucm data = sasts.E_commerce;
      id date interval = quarter;
      model E_commerce;
      randomreg Total_retail /plot=smooth;
      level plot=smooth var=0 noest;
      season length = 4 plot=smooth;
      outlier ;
      estimate plot=(panel residual);
      forecast lead=24 plot=forecasts alpha=0.1;
run;
ods graphics off;
```

The smoothed version of the time varying regression coefficient is plotted by the `plot=smooth` option. Note that this option has to be preceded by a slash (/) in the `randomreg` statement while no slashes are needed for the options in the level, slope or season statements. The plot of the regression coefficient, Figure 8, shows that the coefficient has increased from 0.03 (that is 3% of the total retail sales) in the year 2000 to 0.06 in 2010.

The seasonal component, Figure 9, changes during the years. It seems like the seasonal spike has moved from the first quarter to the fourth quarter of the year. It has to be kept in mind that the seasonality in this model is not for the value of E-Commerce itself but for the residuals in the regression using total retail sales as input variable. One possible explanation is that this could be due to problems in the registration of the volume of E-Commerce volume where the time of the actual trade in recent years is observed more immediate but the E-Commerce at Christmas sales earlier may be registered in the beginning of the new year.

MODEL FIT

The residuals are plotted along with a panel of diagnostic plots for model fit, see Figure 10 and 11 as specified by the `plot=(panel residual)` option in the `estimate` statement. The model fit seems acceptable when judged from the panel of residual plots as the normality assumption seems acceptable and no autocorrelation problems seem present.

Analyzing the time series of U.S. E-Commerce using Proc Ucm

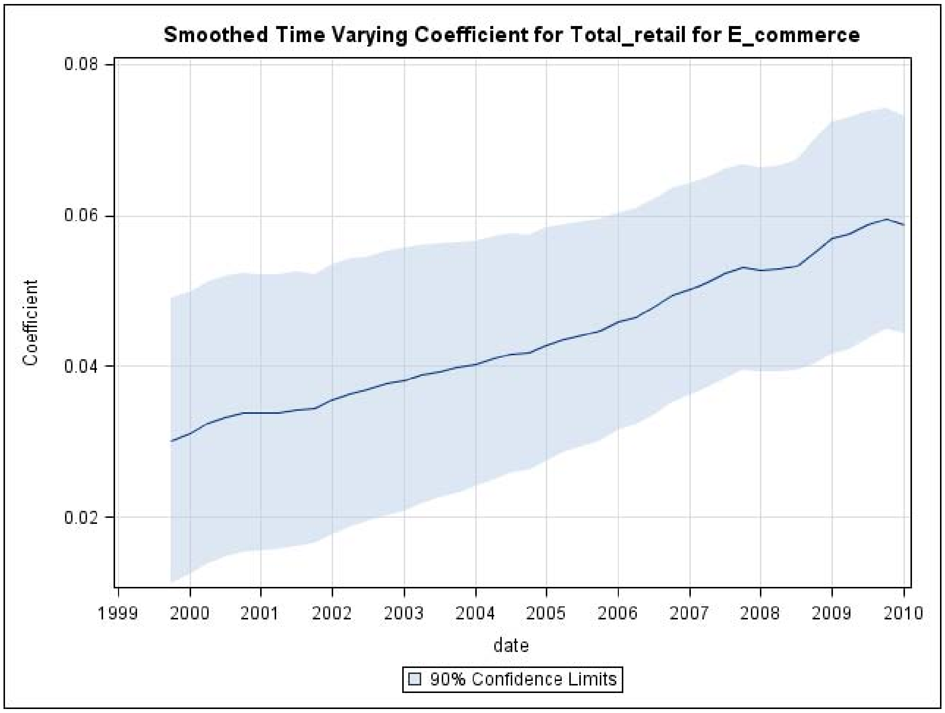


Figure 8 The time varying regression coefficient.

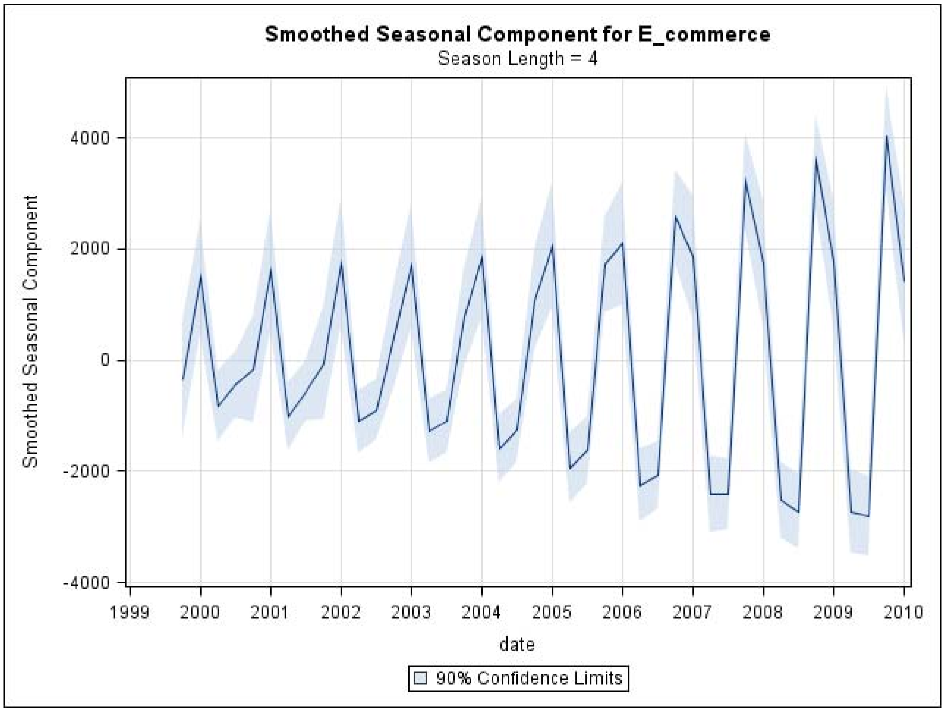


Figure 9 The seasonal component.

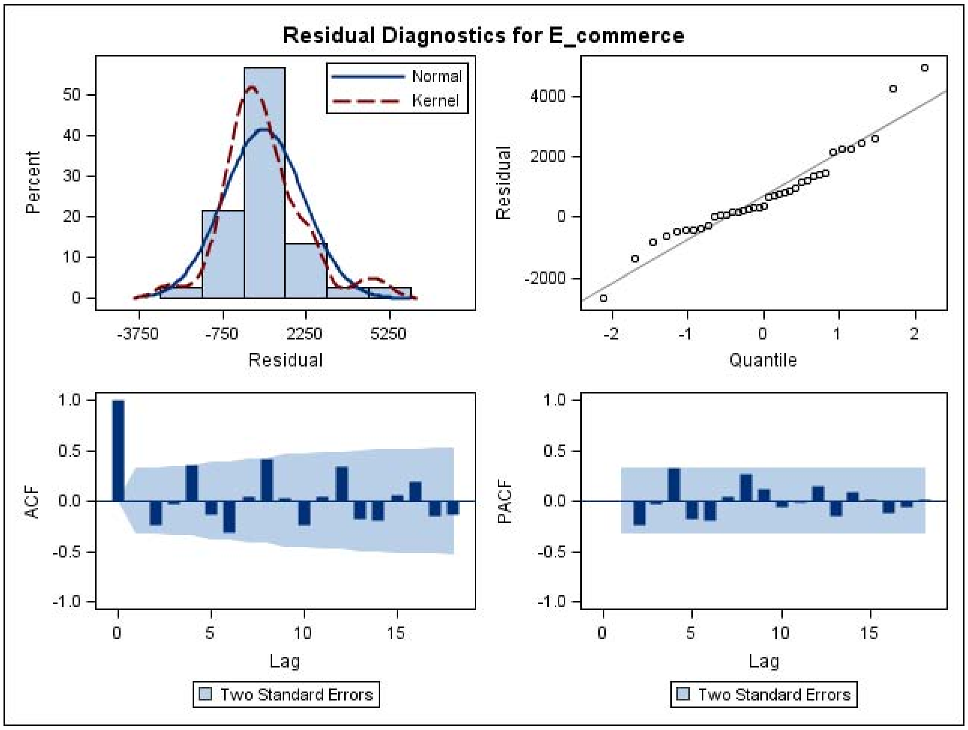


Figure 10 Residual plots.

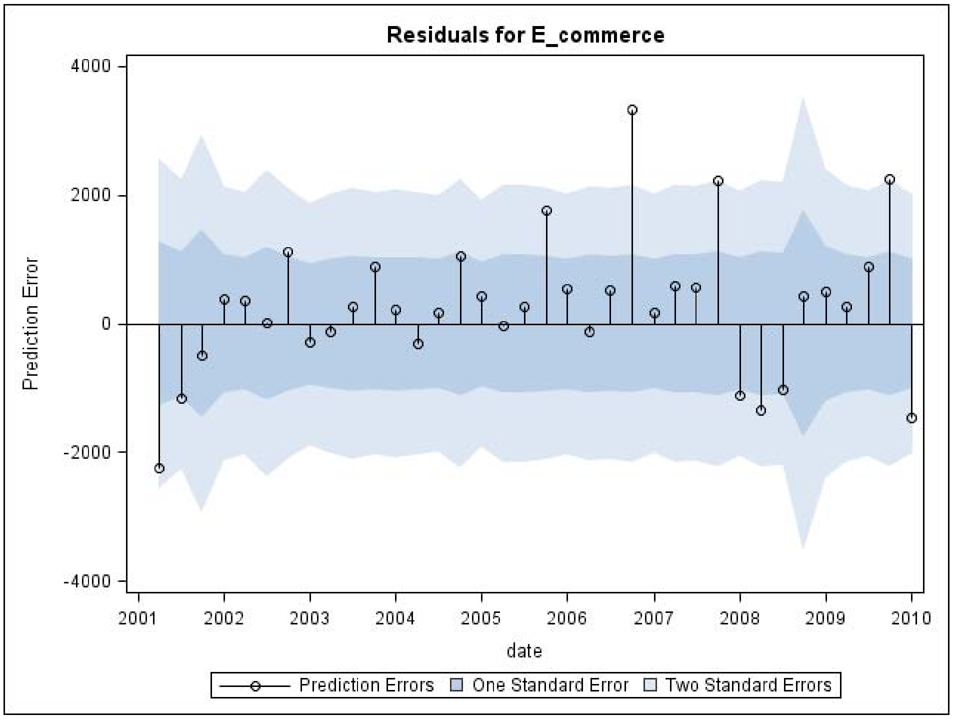


Figure 11 The residual time series.

The residual plot, Figure 11, could indicate some heteroskedasticity problems as the variance of the residuals seem to increase during the years. The residual for the first observation is numerically large but this is probably due to burn in problems and not of any econometric importance. Also the estimated seasonal component has an increasing amplitude. This is natural as seasonality and residual errors most realiastic have to be described as relative features and not in absolute terms as in all the models presented. Proc Ucm however offers no multiplicative version or other transformation possibilities like Proc Esm for exponential smoothing and Proc X12 for seasonal adjustment. In applications of Proc Ucm the only possibility for modeling multiplicity is to analyze the logarithmically transformed series and afterwards transform the results back again by the exponential function. In many situations results are needed in the original scale and results obtained by a log-transformation and a transformation back by exponentials are hard to communicate to a non mathematical audience which only demand results. All the time varying components, coefficients etc. in Proc Ucm to a large extend make these transformations superfluous as seen in in this analysis.

DISCUSSION

Proc Ucm is by this application proved to be a safe choice for a user who needs quick results while completely uninterested in genuine statistical modeling. However the results also provide a basis for more detailed modeling leading to a statistical models which could be thoroughly tested. Allowing every part of the model to evolve over time makes homogeneity and stationarity conditions unnecessary in the process towards useful results.Moreover the easy coding and the rich amount of produced graphics are very attractive features of Proc Ucm.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anders Milhøj
Department of Economics, University of Copenhagen
Øster Farimagsgade 5
DK 1353 Copenhagen K

Work Phone:+45 35323265
E-mail: anders.milhoj@econ.ku.dk
Web: www.econ.ku.dk/milhoj