Paper 262-2010

Applications of Proc ESM in SAS® 9.2

Anders Milhøj, University of Copenhagen, Denmark

ABSTRACT

In this paper the various facilities of Proc ESM in SAS® 9.2 will be demonstrated using a dataset of 87789 registrations of the speed of passing cars on a highway near Copenhagen. Proc ESM (for Exponential Smoothing Models) was introduced in SAS® version 9.2 as a procedure corresponding to more modern time series analysis than its successor Proc Forecast as the smoothing parameters are estimated and not fixed. Moreover Proc ESM has facilities to accumulate transactional data in various ways and has a broad flexibility in transforming the time series to be forecasted.

INTRODUCTION

The availability of fast modern computer equipment has made more efficient predicting methods possible for practical use. These facilities include the possibility of analyzing transactional data with many observations and the possibility of estimating many parameters efficiently in order to derive the best forecasting procedure. In the SAS® system many of these possibilities have for some years been included in special packages for forecasting, but from SAS® version 9.2 some of these features are available in form of a new procedure Proc ESM in the SAS® ETS package.

The various features of Proc ESM are demonstrated by analyses of various aspects of the same data set of automatically registered speed measurements on a high way near Copenhagen. The traffic is a combination of two main transport needs with one component being the drivers living outside Copenhagen travelling in an out of the city for work while the other part is the transit traffic from Germany to Sweden. The data set consists of speed measurements for four days in kilometers pr. hour giving a total of 87789 observations in summer 2002. Each speed measurement is registered by the exact time of the day in seconds.

ACCUMULATION

As a first description of the dataset the number of cars passing each minute is plotted by an application of Proc ESM

```
ods graphics;
proc esm data=sas2009.hastighed plot=(modelforecasts);
id datetime interval=minute accumulate=nobs ;
forecast speed;
run;
ods graphics off;
```

A seasonal effect is clearly seen as the number of passing cars of course depends on the hour of the day with only few cars at night while many persons are driving for work in the mornings. It seems that the maximum is about 23 cars passing in a minute and it is also clear that at night the traffic is low so in fact for some minutes no cars are passing.

This call of Proc ESM also fits the default simple exponential smoothing model, but only a few forecasts are plotted so they are hardly visible. The predicted values are displayed on the graph as a full line and the actual values mostly fall within the forecast limits. If the interval is changed from minute to hour in the above program the picture becomes more informative. Here also predictions 24 hours ahead are given as a rather uninformative constant line - more meaningful predictions are derived in the next section.

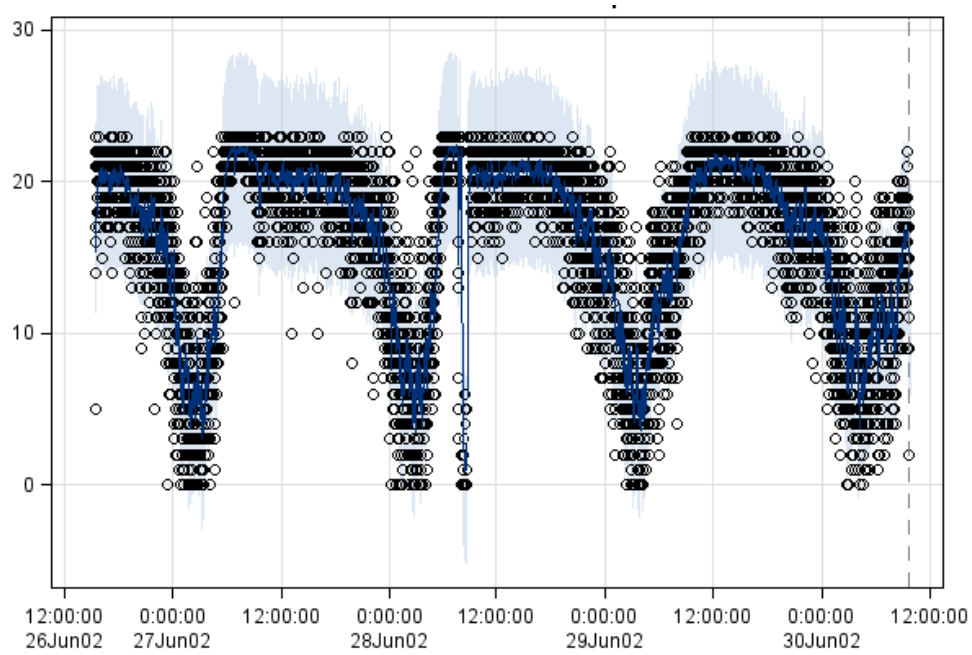

Figure 1. The number of passing cars in one minute

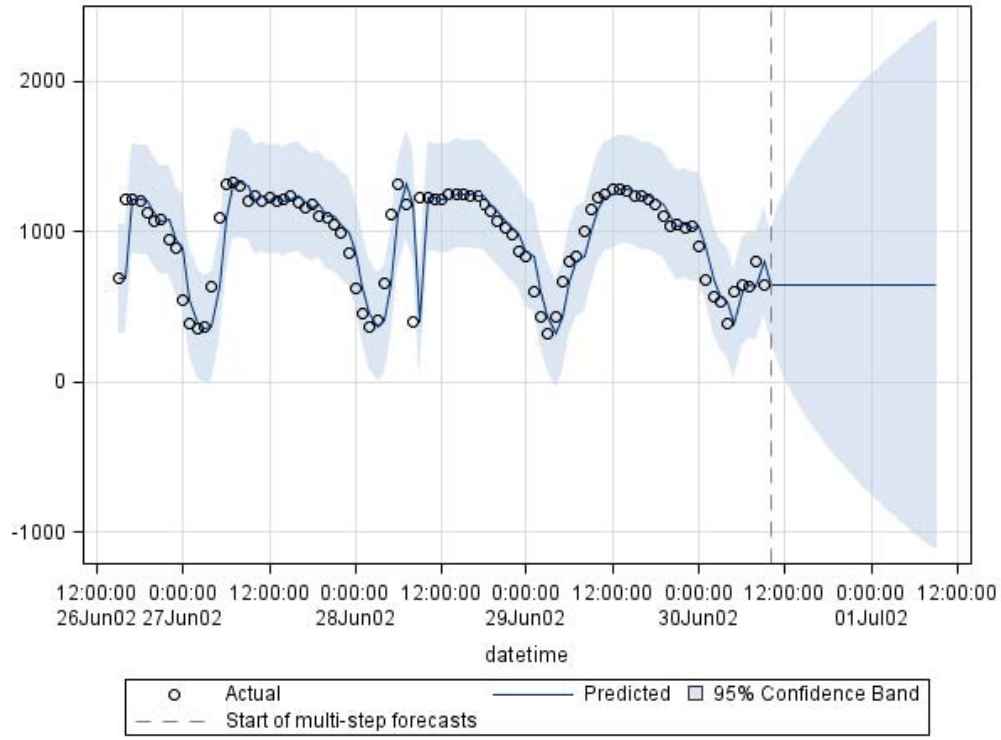

Figure 2. The number of cars passing one hour

The average speed each minute is derived by the program

```
ods graphics;
proc esm data=sas2009.hastighed lead=120 plot=(modelforecasts forecastsonly);
id datetime interval=minute accumulate=average ;
forecast speed/model=double;
run;
ods graphics off;
```

The speed is seen to drop significantly in the morning rush hours the first two days but as the last two days are Saturday and Sunday the traffic is more constant each hour. At night a few very fast cars are observed.

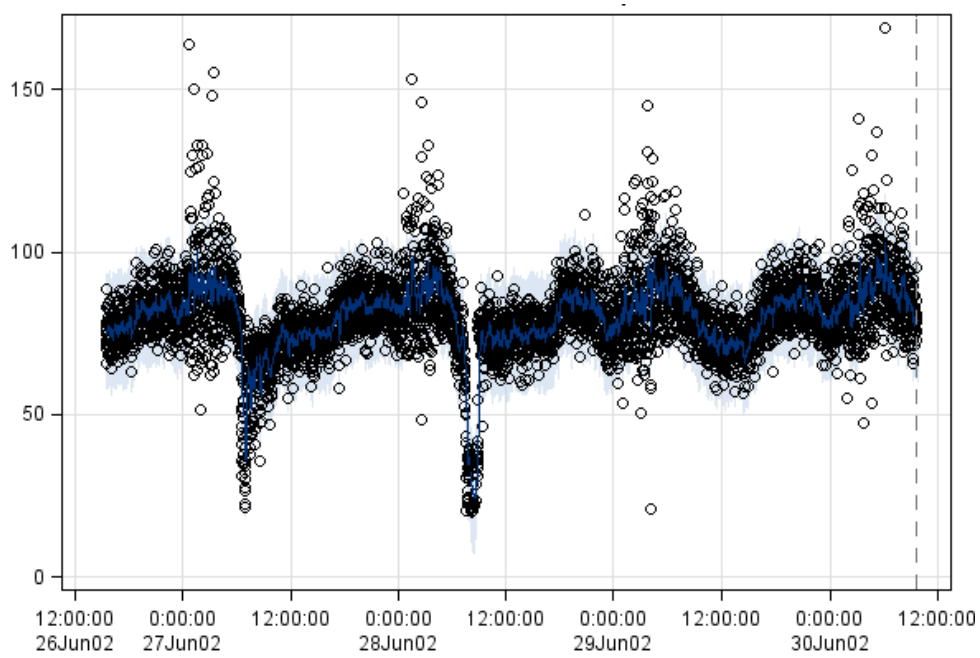

Figure 3. The average speed for one minute

The other facilities for accumulation, minimum, maximum, median and standard deviation could also be relevant when analysing this time series as the minimum speed gives a picture of crowdedness and the maximum speed is a picture of the respect for the laws as a speed restriction exists. The standard deviation could be seen as an indicator of the risk for accidents as a high standard deviation appears if some cars are slow while others fast competing on the same lane.

It looks as the automatic speed measurement equipment is unable to register speed below a certain limit around 20 km/h.

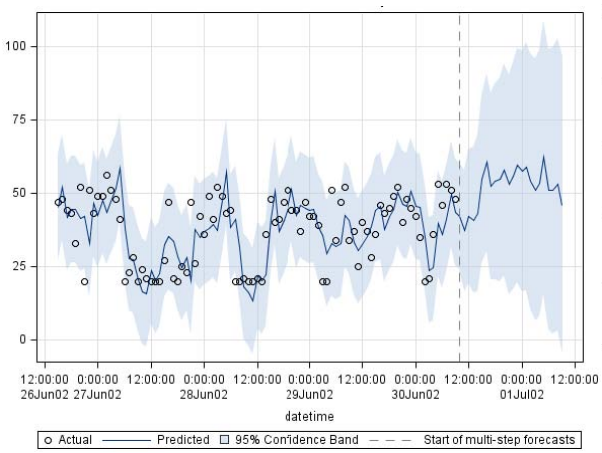

Figure 4. Minimum speed for one hour

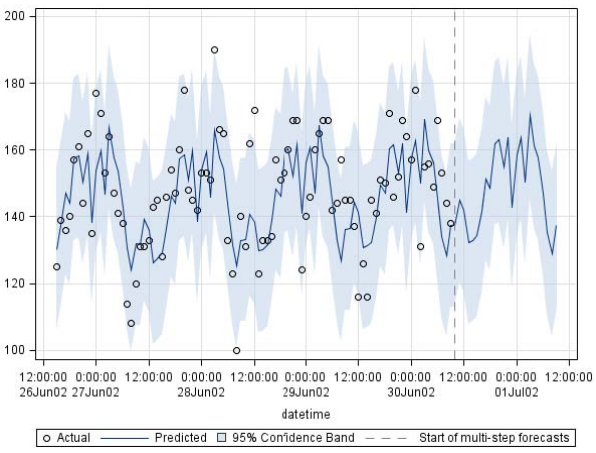

Figure 5. Maximum speed for one hour

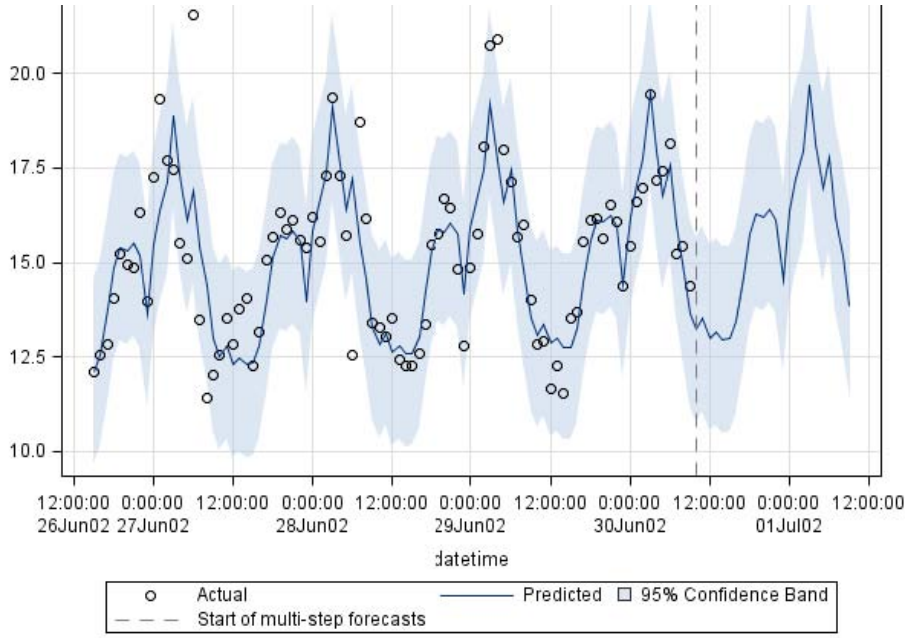

Figure 6. Standard deviation for speed one hour

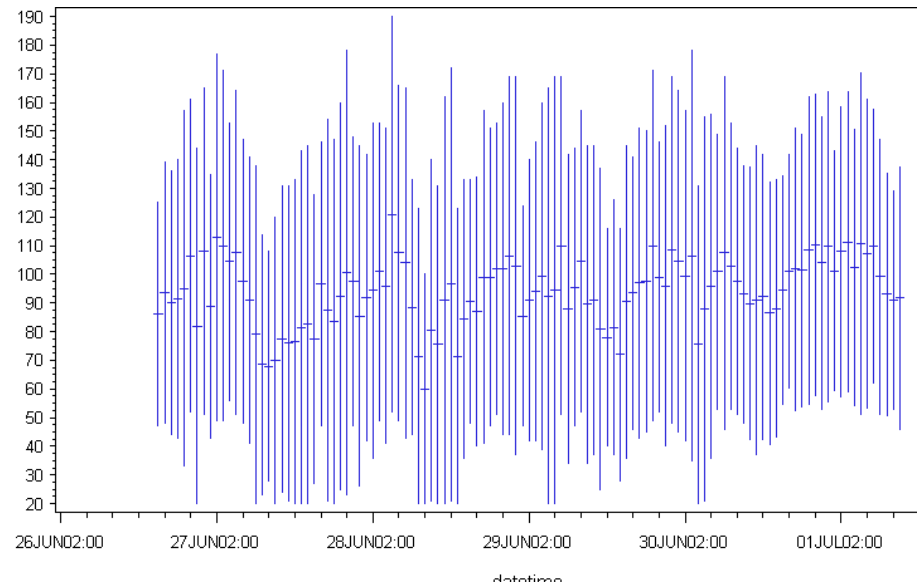

Figure 7. High-low plot for speed one hour

The accumulated series could of course be saved for further analysis and in this way a "High-Low" diagram of the speed each hour is easily generated. One obvious conclusion from this plot is, that queuing is not a problem for a full hour as the maximum speed never falls below 100 km/h. The fact that the minimum speed never exceeds 50 km/h is easily explained by slow drivers like trucks etc.

FORECASTING

If no seasonality is present the technique of exponential smoothing is often applied. The idea is that the series varies around some smooth curve that might be considered as the true level which may be time varying. The actual observations apart from this true level are also affected by other irregularities which mean that the true level is unobserved. When the smoothed level is denoted $\bar{X}_t$ a formula for updating the level by a moving average has the form

$$\bar{X}_t = (1-\alpha)\bar{X}_{t-1} + \alpha X_t = \bar{X}_{t-1} + \alpha(X_t - \bar{X}_{t-1}),$$

where α is a smoothing constant 0 < α < 1. This formula provides us with a present value of the smoothed component defined as an average of the previously estimated value of the smoothed component and the present observed value of the series. The smaller the value of α, the smoother the estimated level $\bar{X}_t$ becomes. The value α = 0 corresponds to the extreme situation of the level being absolutely constant, while the value α = 1 simply lets the smoothed value equal the observed value so no smoothing is performed. The value α = 1 corresponds to a model which in statistical terms would be denoted a random walk, or at least a situation where a differencing is needed in order to obtain stationarity.

In order to start the algorithm the initial value $\bar{X}_1$ is defined as $\bar{X}_1 = X_1$. Forecasting past the last observation X_T is simply performed by letting the prediction be defined as

$$\hat{X}_{T+i} = \bar{X}_T$$

This gives a constant prediction which could be sufficient in many situations, but also of course should be refined in situations with trending or seasonal behaviour.

If a trend is present the idea of simple exponential smoothing is extended to include a linear trend which is fitted by formulas similar to the smoothing above including other smoothing parameters. Seasonal components are also estimated by exponential smoothing. The seasonal effects, i.e. an effect of the month of may, is updated by a weighted average of a previous defined May-effect and the present value for the month of May. In the formula below the series is decomposed as a sum of a smoothed level component $\bar{X}_t$ and a seasonal component S_t as seen by the formula in which the length of the seasonal cycle is denoted p.

$$\bar{X}_t = (1-\alpha)\bar{X}_{t-1} + \alpha\left(X_t - S_{t-p}\right)$$

$$S_t = (1-\delta)S_{t-p} + \delta\left(X_t - \bar{X}_t\right).$$

If the seasonal smoothing weight δ is close to zero the seasonal structure is fixed corresponding to an application of seasonal dummy variables in a regression model, while values of δ closer to one allows for time varying seasonal structures. Many other versions of these formulas exist and Proc ESM offers a broad spectrum of these possibilities.

In the previous procedure , Proc Forecast, the smoothing parameter α is fixed at the value 0.2 in the case of simple exponential smoothing and similarly other fixed values were applied for the other smoothing parameters in the more advanced versions of the algorithm. These values were found as a result of experiences by forecasting many real time series. A new feature in Proc ESM is that these values are optimized in order to provide the best fit to the actual being forecasted. This is done by minimizing sum of squares of the prediction errors in the data period. This algorithm corresponds to the methods applied by more advanced time series modelling procedures and hence Proc ESM could be seen as a efficient easy-to-use alternative to more refined forecasting techniques.

In the application of forecasting the number of cars passing the parameter α is close to zero for short time intervals while it almost attains the value one for longer time intervals as seen by the table for time intervals in the span from 15 seconds to one hour.

Period in minutes	Estimate of α	Standard error	Residual mean square error	R^2
0.25	0.05003	0.001502	1.371	0.46
0.50	0.08709	0.002775	2.074	0.60
1.00	0.15244	0.005100	3.187	0.72
2.00	0.28382	0.009467	4.911	0.81
3.00	0.38644	0.013102	6.525	0.84
4.00	0.49680	0.016848	8.136	0.86
5.00	0.58049	0.020067	9.596	0.87
10.00	0.86052	0.030788	19.039	0.87
30.00	0.99900	0.049397	79.314	0.74
60.00	0.99900	0.071553	183.347	0.64

Simple exponential smoothing or perhaps double exponential smoothing including a trend, could be applied for short forecasting horizons, say up to 20 minutes, but for longer horizons they are unrealistic as the predictions form either a constant or an everlasting trend. Instead for longer horizons, say for several hours, seasonal models following a daily pattern have to be applied. These models incorporate the fact that traffic is light at night time and heavy in the mornings by a seasonal component which could be interpreted as dummies for each hour of the day. As for this data set no trend component is relevant for longer horizons the forecast should be based on the seasonal and a level. This is provided by the option seasonal as contrary to the additive or multiplicative Winter methods which include trends. For the number of cars passing each hour this is done by the program.

```
ods graphics;
proc esm data=sas2009.hastighed lead=24 plot=(modelforecasts);
id datetime interval=hour accumulate=nobs;
forecast speed/method=seasonal;
run;
ods graphics off;
```

In this situation the smoothing parameter α for the level is estimated as 0.63 while the smoothing parameter δ for the seasonal component is reported as 0.001, which seems to be the initial value for the estimating iterations. The seasonal smoothing parameter is then preferably taken as zero which implies that a seasonal model with 24 fixed dummies - one for each hour a day - is preferable. The root mean square error of the forecasts are 137.8 which is much smaller than the standard deviation 306.0 for the original data series. As a measure of fit the value $R^2 = 0.79$ is reported.

Seasonal Exponential Smoothing Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Level Weight	0.63345	0.07100	8.92	<.0001
Seasonal Weight	0.0010000	0.16438	0.01	0.9952

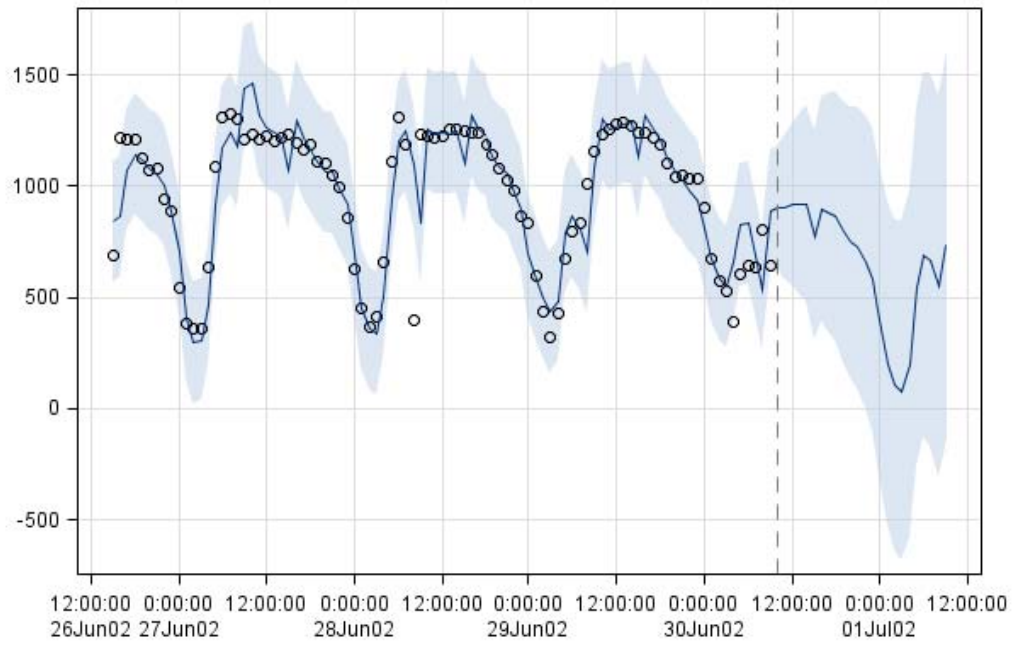

Figure 8. Forecasting of hourly number of passing cars using a seasonal model

In this situation the forecaster of course knows better as the forecasting period is a Monday and the prediction of seemingly light traffic is based on an estimation period for only four days ending in the weekend. But the seasonal - in this application hourly - component of the forecast looks right.

Forecasts of the hourly average speed are easily derived by simply changing the option "nobs" to "average" in the id statement of the procedure call. The low speed in the morning rush hours is easily seen, but as the prediction is based on the weekend traffic the morning traffic the next day, a Monday, is probably underestimated.

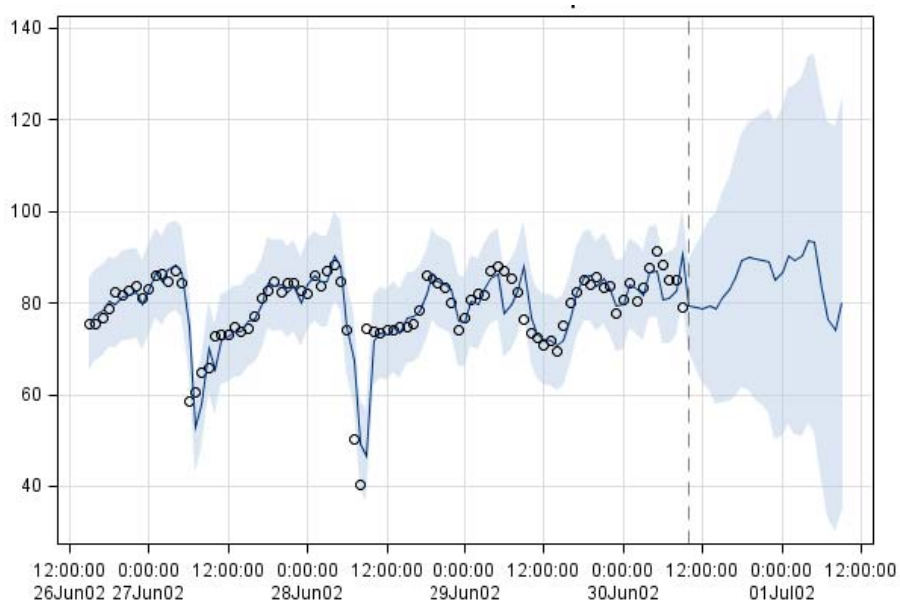

Figure 9. Forecast of average speed using a seasonal model

LOGARITHMIC TRANSFORMATION

In this section short term forecasting of the speed based on the average speed accumulated by one minute intervals are considered. Simple exponential smoothing gives the following prediction errors in the data period.

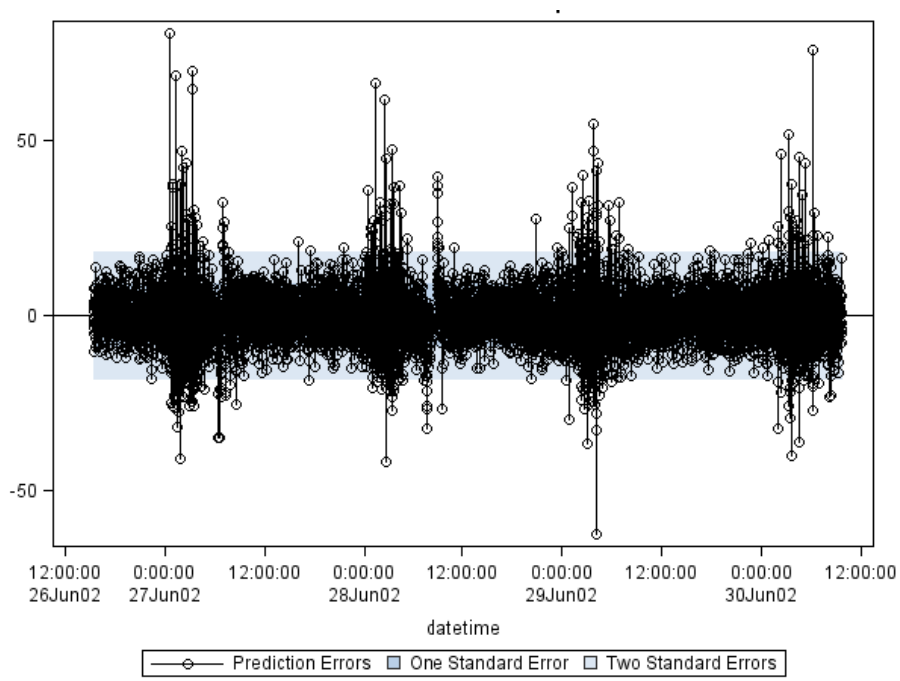

Figure 10. Forecasting errors

A clearly heteroskedastic pattern is seen, as the outliers are obviously more frequent at night than in the daytime. As is also seen from the previous analyses the average speed is also much higher at night than in the daytime. A situation like this calls for a transformation in order to provide the analyst with correct forecast limits and in order to prevent that the highly volatile observations at night dominate the estimates of the model parameters.

In this situation a logarithmic transformation seems appropriate as it considers relative errors instead of absolute errors. In order to reduce the number of observations the data set is accumulated to ten minutes time intervals. The error plot gives non constant confidence limits corresponding to a narrow interval on the absolute scale in the daytime with a heavy traffic allowing for nearly no room for fast driving. At night the prediction interval hopefully should include manly fast drivers but also some slow trucks giving rise to a much larger variation leading to broader confidence limits. A few outliers exist as it is possible to drive both fast and slow at night and as the traffic is light one or two cars can influence the average significantly.

```
ods graphics;
Title '';
proc esm data=sas2009.hastighed out=bc2 plot=all ;
id datetime interval=minute10 accumulate=average;
forecast speed/transform=log;
run;
ods graphics off;
```

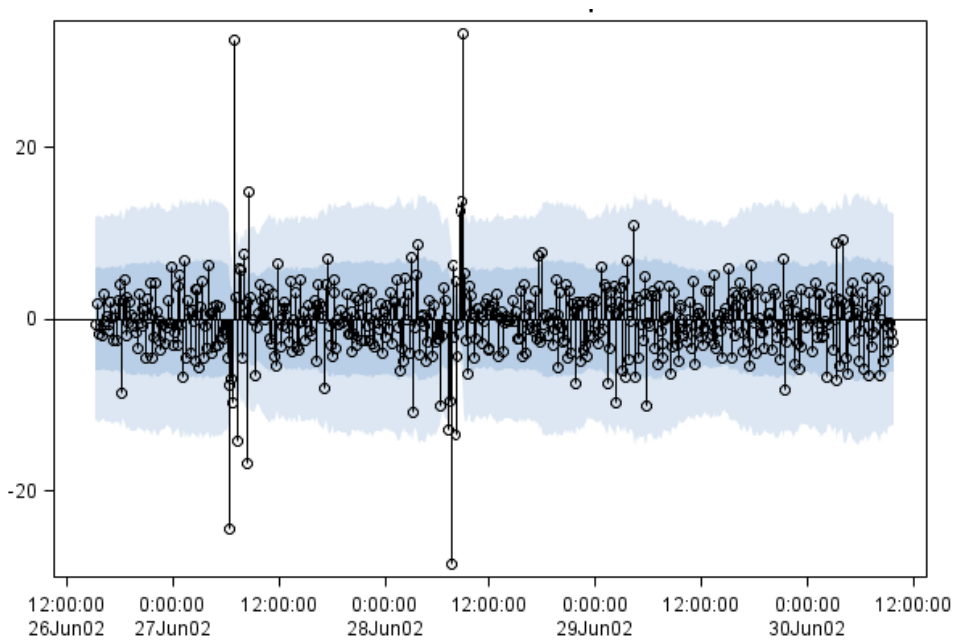

Figure 11. Forecasting errors using a logarithmic transformation

LOGISTIC TRANSFORMATIONS

In fact the speed measurements were registered in a period with road work and for the security of the workers and because of the extended risk of accidents a temporary speed restriction was imposed. This limit was set as low as 70 km/h compared to the usual limit 110 km/h and it applied for the whole day even if no work is actual done at night time. The number of cars exceeding this speed limit is rather dramatic as seen in the previous analyses. Only in the heavy morning traffic the drivers obey the speed limit but surely more as a necessity due to the rush hours than because of respect for the law. The low number of too fast drivers at night is of course only due to the low number of total drivers at night.

The number of cars and the number of drivers exceeding the speed limit is easily saved in new datasets by simple applications of Proc ESM and when these new data sets are combined the fraction of drivers exceeding the speed limit could be calculated. This fraction is forecasted by a seasonal model and 24 hour predictions are made.

```
ods graphics;
proc esm data=d lead=24 plot=(modelforecasts forecastsonly) ;
id datetime interval=hour;
forecast share/model=seasonal ;
run;
ods graphics off;
```

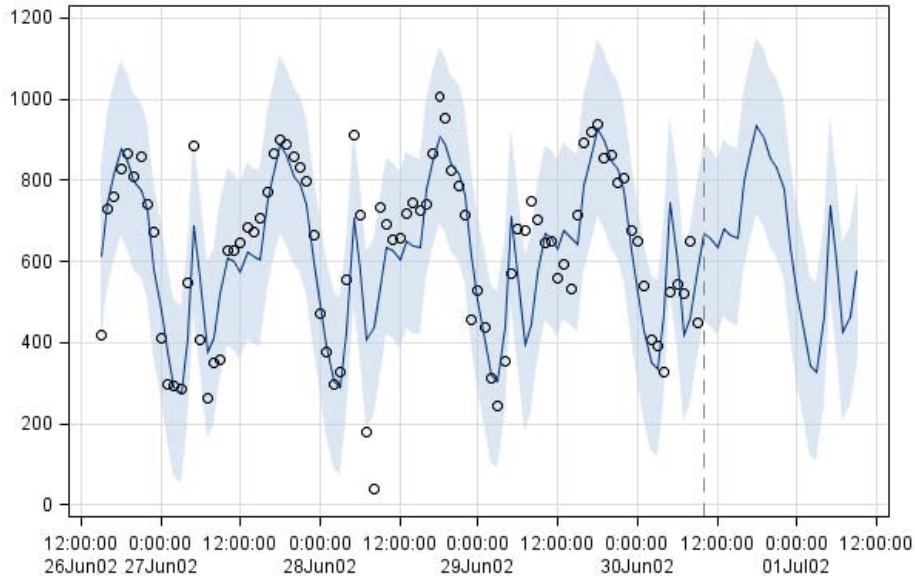

Figure 12. Seasonal model for number of too fast drivers

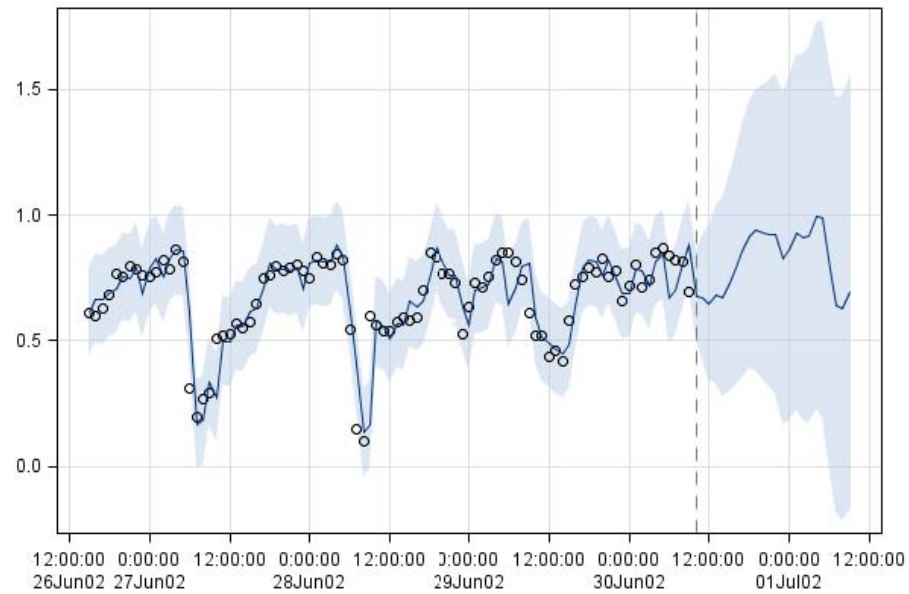

Figure 13. Forecast of share of too fast drivers

In this situation a transformation of the fractions is relevant as they by definition are numbers between zero and one and the model treats values near one half in the same way as extreme values near zero and one and hence the model is unable to fit the extreme values appropriately. Moreover the prediction limits are outside the relevant interval from zero to one for the proportion of too fast cars. These problems are solved by a logistic transformation, which transforms the fraction to a real number which is analyzed by the exponential smoothing models. The transformation, called ℓ, of the fraction, called f, is

$$\ell = \text{logistic}(f) = \frac{\exp(f)}{1+\exp(f)}$$ and the inverse transformation is $$f = \ln\left(\frac{\ell}{1-\ell}\right).$$

```
ods graphics;
proc esm data=d lead=24 plot=(modelforecasts forecastsonly) ;
id datetime interval=hour;
forecast share/model=seasonal transformation=logistic;
run;
ods graphics off;
```

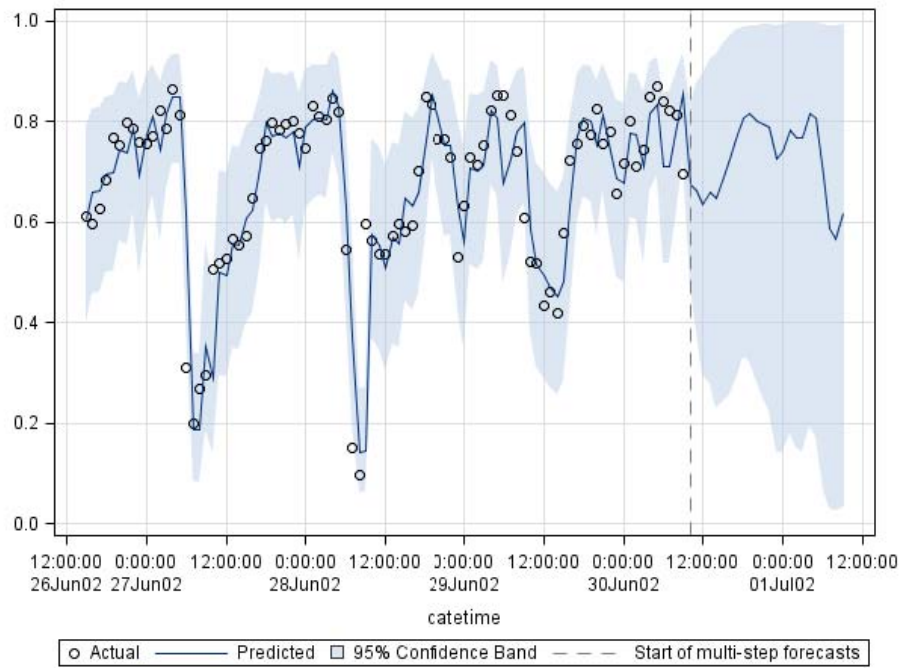

Figure 14. Forecast of share of too fast drivers using a logistic transformation

Now the forecast limits are bounded between zero and one and they have unsymmetrical confidence limits leaving less possibility for drivers to exceed the speed limit than for drivers to obey the limit. The prediction errors in the data period have a symmetric distribution when no transformation is applied while the logistic transformation allows for an asymmetric distribution. This lack of symmetry is easily understood as the part of drivers exceeding the speed limit is almost zero in the rush hours when fast driving is impossible but as the number of too fast drivers is predicted as very high at night, up to 80%, it cannot be much higher.

This is also seen from the histogram of the forecast errors in the data period, where the distribution is symmetric when no transformation is applied but is skewed when the logistic transformation is applied.

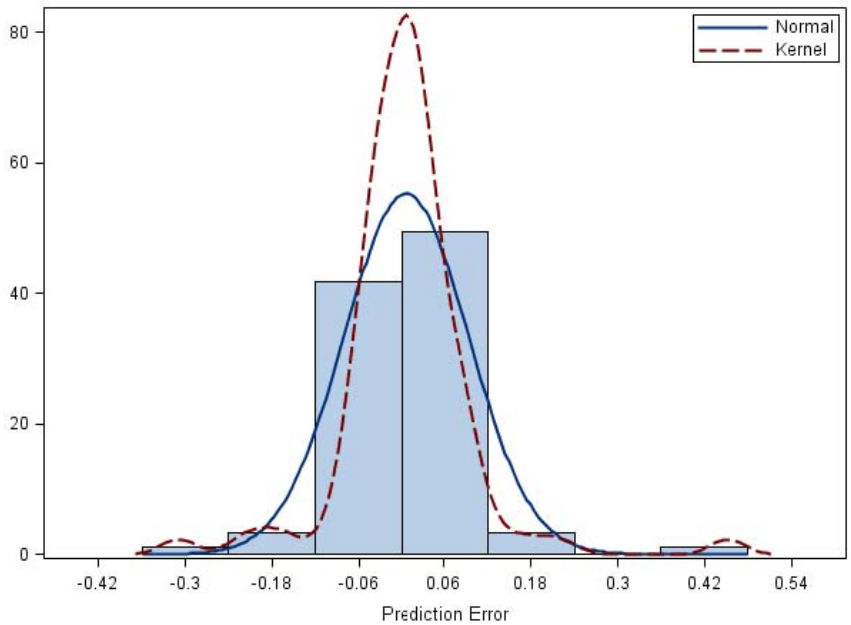

Figure 15 Histogram of forecasting errors using a logistic transformation

The fitted and predicted values are transformed back to the original scale from zero to one. As default this transformation is performed in a way such that the prediction is for the mean of the derived distribution of future observations. As the logistic transformation is highly non-linear this is obtained by a Taylor expansion. Another possibility is to use the median as a predictor, which corresponds to a simple back transformation of the prediction of the logistic transformed fraction. An application of the mean as a predictor of course gives smaller values of mean square based measured of fit, but the median often gives a more intuitive idea of a skew distribution than the mean. In this situation the predictions by the median are much larger than the predictions derived as the mean. The prediction is that the median of distribution of the proportion of cars to exceed the speed limit will be as high as 90% (with mean around 80%) for some hours at night. In more daily words this implies that the chance that more than 90% of the drivers exceed the limit is larger than one half for several hours at night.

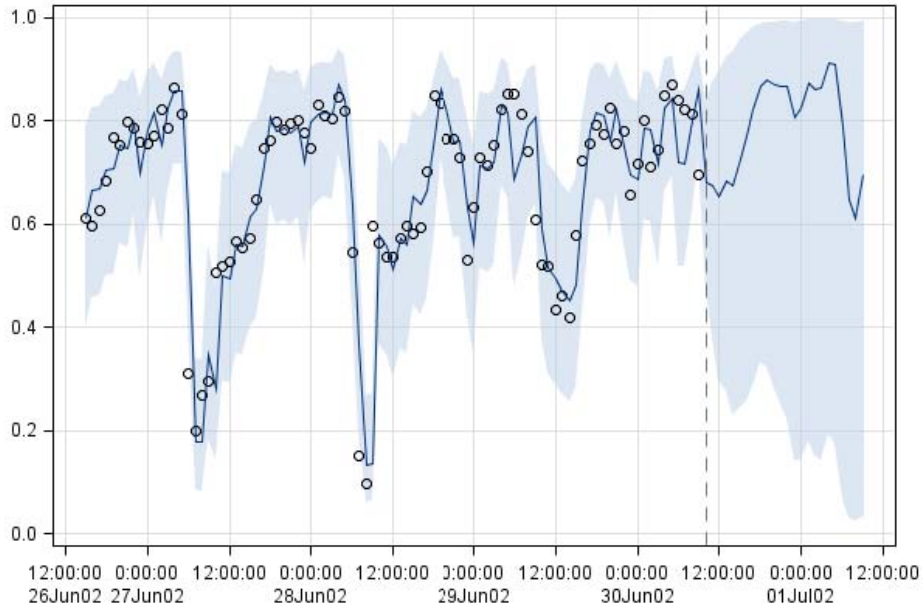

Figure 16 Forecast of median share of too fast drivers

CONTACT INFORMATION

Anders Milhøj

Department of Economics, University of Copenhagen

Øster Farimagsgade 5
DK1353 Copenhagen K
Denmark

Tlf: +45 35323265

Anders.milhoj@econ.ku.dk

http://www.econ.ku.dk/milhoj/