Paper SAS214-2014

When Do You Schedule Preventive Maintenance? Multivariate Time Series Analysis in SAS/ETS®

Kenneth Sanford, SAS Institute Inc.

ABSTRACT

Expensive physical capital must be regularly maintained for optimal efficiency and long-term insurance against damage. The maintenance process usually consists of constantly monitoring high-frequency sensor data and performing corrective maintenance when the expected values do not match the actual values. An economic system can also be thought of as a system that requires constant monitoring and occasional maintenance in the form of monetary or fiscal policy. This paper shows how to use the SSM procedure in SAS/ETS to make forecasts of expected values by using high-frequency multivariate time series. The paper also demonstrates the functionality of the new SASEFRED interface engine in SAS/ETS.

INTRODUCTION

Time series data are observations that are collected on the same sampling unit over regular intervals. One of the most common uses of time series data is to extract information about underlying components, which might be regular seasonal or temporary cycles in the data. A time series might also be characterized by an underlying trend. Early detection of an underlying trend in the data might provide an early warning system for catastrophic failure. This paper shows how you can use the SSM procedure in SAS/ETS in the context of a multivariate time series to extract a common trend and to signal opportunities for intervention.

One application of early detection of an underlying trend is preventive maintenance of an economic system. This paper uses that trend, extracted from a multivariate time series data set, to offer moments for possible intervention. The main analytical tool in this study is a linear state space model that is estimated by the Kalman filter and smoother (KFS). This paper uses a multivariate time-series data set to look for an underlying trend. It also showcases another recent enhancement to SAS/ETS, the SASEFRED engine, which enables users to easily and freely dynamically extract data from the Federal Reserve Economic Data (FRED) repository and then use these data as part of a SAS program. This paper uses data from FRED to identify candidate time periods when “preventive maintenance” could be performed.

The following section provides a high-level overview of linear state space models and presents both the framework for identifying the underlying trend of the economic system and a mechanism for identifying candidate intervention points. The next section examines the data set and discusses the use of the SASEFRED access engine. The subsequent section discusses the model findings and potential uses. The final section concludes and offers potential extensions.

OVERVIEW OF STATE SPACE MODELING

The (linear) state space model is described in the literature in a few different ways and with varying degree of generality. The description in this section loosely follows the description given in Durbin and Koopman (2012, chap. 6, sec. 4) and provides the novice with a primer on this method of modeling.

Suppose that observations on the following variables are collected in a sequential fashion (indexed by a numeric variable, τ): the vector **y** = $(y_1, y_2, \ldots, y_q)$, which denotes the q-variate response values, and the k-dimensional vector **x**, which denotes the predictors. Suppose that the observation instances are $\tau_1 < \tau_2 < \cdots < \tau_n$, where n indexes the length of the time series vector. The possibility that multiple observations are recorded at a particular instance is not ruled out, and the successive observation instances do not need to be regularly spaced. That is, $(\tau_2 - \tau_1)$ does not need to equal $(\tau_3 - \tau_2)$. For $t = 1, 2, \ldots, n$, suppose p_t denotes the number of observations that are recorded at instance τ_t and let $p_t \geq 1$. For notational simplicity, an integer-valued secondary index *t* is used to index the data so that *t* = 1 corresponds to $\tau = \tau_1, t = 2$ corresponds to $\tau = \tau_2$, and so on. Consider the following model:

$$\begin{aligned} \mathrm{Y}_t &= \mathbf{Z}_t \boldsymbol{\alpha}_t + \mathrm{X}_t \boldsymbol{\beta} + \boldsymbol{\epsilon}_t && \textit{Observation equation} \\ \boldsymbol{\alpha}_{t+1} &= \mathrm{T}_t \boldsymbol{\alpha}_t + \mathrm{W}_{t+1} \boldsymbol{\gamma} + \mathbf{c}_{t+1} + \boldsymbol{\eta}_{t+1} && \textit{State transition equation} \\ \boldsymbol{\alpha}_1 &= \mathbf{c}_1 + \mathrm{A}_1 \delta + \mathrm{W}_1 \boldsymbol{\gamma} + \boldsymbol{\eta}_1 && \textit{Initial condition} \end{aligned}$$

The following list describes these equations:

- The *observation equation* describes the relationship between the $(p_t * q)$-dimensional response vector Y_t and the unobserved vectors $\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t,$ and $\boldsymbol{\epsilon}_t$. The q-variate responses are vertically stacked in a column to form this $(p_t * q)$-dimensional response vector $\mathbf{Y}_t$. The *m*-dimensional vectors $\boldsymbol{\alpha_t}$ are called states, the *k*-dimensional vector $\boldsymbol{\beta}$ is the regression coefficient vector associated with predictors **x**, and the $(p_t * q)$-dimensional vectors $\boldsymbol{\epsilon_t}$ are called the *observation disturbances*. The matrices $\mathbf{Z}_t$ (of dimension $(q * p_t) \times m$) and $\mathbf{X}_t$ (of dimension $(q * p_t) \times k$) correspond to the state effect and the regression effect, respectively. The elements of $\mathbf{X}_t$ are assumed to be fully known. The states $\boldsymbol{\alpha}_t$ and the disturbances $\boldsymbol{\epsilon}_t$ are random sequences. It is assumed that ϵ_t is a sequence of independent, zero-mean, Gaussian random vectors with diagonal covariances, where the diagonal elements are denoted by $\sigma^2_{i,t}, i = 1, 2, \ldots., q * p_t$.
- The state sequence $\boldsymbol{\alpha_t}$ is assumed to follow a Markovian structure that is described by the state transition equation and the associated initial condition.
- The *state transition equation* postulates that a new instance of the state, $\boldsymbol{\alpha_{t+1}}$ **,** is obtained by multiplying its previous instance, $\boldsymbol{\alpha}_t$, by an *m*-dimensional square matrix $\mathbf{T}_t$ (called the state transition matrix) and by adding three more terms: a known nonrandom vector $\mathbf{c}_{t+1}$ (called the state input); a regression term $\mathbf{W}_{t+1}\boldsymbol{\gamma}$, where $\mathbf{W}_{t+1}$ is an *m* x *g*-dimensional design matrix that contains fully known elements and $\boldsymbol{\gamma}$ is the *g*-dimensional regression vector; and a random disturbance vector $\boldsymbol{\eta_{t+1}}$. The *m*-dimensional state disturbance vectors η_t are assumed to be independent, zero-mean, Gaussian random vectors with covariances $\mathbf{Q}_t$ (not necessarily diagonal).
- The *initial condition* describes the starting condition of the state evolution equation. The starting state vector, $\boldsymbol{\alpha_1}$, is assumed to be partially diffuse: it is the sum of a known nonrandom vector $\mathbf{c}_1$, a mean-zero Gaussian vector η_1, and the terms $\mathbf{A}_1\boldsymbol{\delta}$ and $\mathbf{W}_t\gamma$. $\mathbf{A}_1\boldsymbol{\delta}$ represents the contribution from a *d*-dimensional diffuse vector $\boldsymbol{\delta}$ (a diffuse vector is a Gaussian vector that has infinite covariance). The observation and state regression vectors ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively) are also assumed to be diffuse. The *m* x *d* matrix $\mathbf{A}_1$ is assumed to be completely known.
- The observation disturbances $\boldsymbol{\epsilon}_t$ and the state disturbances $\boldsymbol{\eta}_t$ (for t ≥ 1) are assumed to be mutually independent. Either the elements of the matrices $\mathbf{Z}_t$, $\mathbf{T}_t$, and $\mathbf{Q}_t$ and the diagonal elements of the observation disturbance covariances $\sigma^2_{i,t}$ are assumed to be completely known, or some of them can be functions of a small set of unknown parameters (to be estimated from the data). Suppose that this unknown set of parameters is denoted by $\boldsymbol{\theta}$**.**
- The *d*-dimensional diffuse vector $\boldsymbol{\delta}$ from the initial condition together with the observation and state regression vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ constitute the overall (*d* + *k* + *g*)-dimensional diffuse initial condition of the model.

Although this description of the state space model might appear involved, it conveniently covers many variants of SSMs that are encountered in practice and precisely describes the most general case that can be handled by the SSM procedure in SAS/ETS.

TYPES OF SEQUENTIAL DATA USED BY THE SSM PROCEDURE

The state space model specification in the SSM procedure requires proper understanding of both the data organization and the form of the model. The SSMs that are appropriate for time series data might not be appropriate for irregularly spaced longitudinal data. PROC SSM distinguishes the following three types of data organization based on the way the observations are sequenced by the index variable (if an index variable is not specified, it is assumed that the observations are sequenced according to the observation number):

- Regular: The observations are recorded at regularly spaced intervals; that is, $\tau_1, \tau_2, \ldots, \tau_n$ are regularly spaced. Moreover, a single observation is recorded at each observation instance τ_i; that is, p_t = 1 for all *t*. Standard time series data (both univariate and multivariate) fall into this category.
- Regular with replication: The observations are recorded at regularly spaced intervals, but p_t > 1 for at least one *t*. The word replication is used loosely here; it does not mean that the multiple observations at τ_t are replications in the precise statistical sense. Panel and cross-sectional data types fall into this category. When panel data contain p cross-sections, $p_t = p$ for all *t*.
- Irregular: The observations are not recorded at regular intervals, and the number of observations p_t at each index instance can be different. Longitudinal data fall into this category.

It is not always easy to decide whether a particular model is appropriate for a particular data type. Whenever possible, the SSM procedure writes a note to the log about the possible mismatch between the specified model and the data type being analyzed.

ESTIMATION OF MODELS: FILTERING AND SMOOTHING

The Kalman filter and smoother (KFS) algorithm is the main computational tool that is used by the SSM procedure for data analysis. For proper treatment of an SSM that uses a diffuse initial condition or when regression variables are present, a modified version of the traditional KFS, called the diffuse Kalman filter and smoother (DKFS), is needed. A good discussion of the different variants of the KFS and DKFS can be found in Durbin and Koopman (2012). The DKFS that is implemented by the SSM procedure closely follows the treatment in de Jong and Chu-Chun-Lin (2003). Additional details can be found in these references and also in the chapter, “The SSM Procedure,” in the *SAS/ETS User’s Guide*.

EXAMPLE: MODELING PREVENTIVE MAINTENANCE OF THE ECONOMIC SYSTEM

The Federal Reserve Bank of Philadelphia produces a well-known index of economic indicators called the Aruoba-Diebold-Scotti (ADS) business conditions index. This index is designed to incorporate high-frequency time series data, which might be recorded at different intervals, into one measure of aggregate economic activity at a single moment in time. This type of “blended time frequency” index could be useful in the burgeoning area of forecasting called “nowcasting,” which uses nearly instantaneously recorded data to make forecasts. This example re-creates a slightly simplified version of the ADS by using data that are dynamically retrieved by the SASEFRED interface. Then, the example provides a framework for how to extract an underlying trend from those data and how to propose preventive maintenance for an economy.

DATA

The data in this analysis are intended to mirror the data in the ADS index; they consist of six different economic time series. These series consist of both stock and flow data (levels and change) and blend data of differing frequencies. Table 1 lists the variables to be used in the creation of this index. All variables are logged and differenced, except for weekly initial jobless claims, which are only logged.

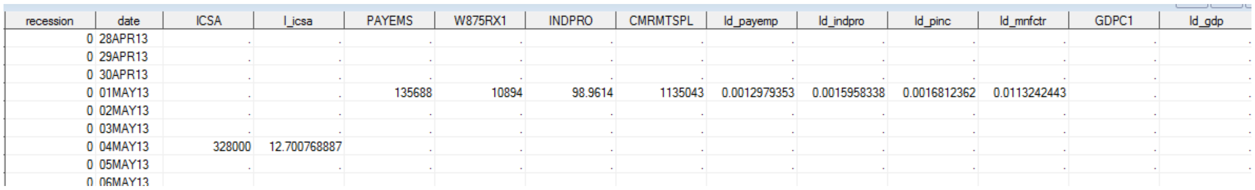
Name	FRED ID	Frequency	Description
ld_payemp	PAYEMS	Monthly	Payroll employment
ld_pinc	W875RX1	Monthly	Real personal income excluding current transfer receipts
ld_mnfctr	CMRMTSPL	Monthly	Real manufacturing and trade industries sales
ld_indpro	INDPRO	Monthly	Industrial production index
ld_gdp	GDPC1	Quarterly	Real GDP
l_icsa	ICSA	Weekly	Industrial jobless claims

Table 1: Variables in Economic Index 1

The SASEFRED interface engine in SAS/ETS 13.1 provides access to the FRED database. The SASEFRED engine enables users to call the FRED database from within a SAS session, thereby ensuring that the newest data are available for analysis. For more information, see the chapter “The SASEFRED Interface Engine” in the *SAS/ETS User’s Guide*. You can use the following code to extract the monthly series by substituting your own API key in the APIKEY= option. You can request a key at (http://api.stlouisfed.org/api_key.html).

```
libname fred sasefred "%sysget(FRED)"
   OUTXML=fredex01
   AUTOMAP=replace
   MAPREF=MyMap
   XMLMAP="%sysget(FRED)fredex01.map"
   APIKEY='XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX'
   IDLIST='payems,w875rx1,cmrmtspl,indpro'
   START='1960-01-01'
   END='2013-12-01'
   FREQ='m'
   OUTPUT=1
   AGG='avg'
   FORMAT=xml;;
```

Similar SASEFRED access engine calls are necessary to extract the real gross domestic product (GDP) and initial jobless claims series, which have different frequencies. Then you need to merge these series to create a data set that has mixed frequencies. The final merged data set contains observations for the highest-frequency time dimension. For example, if the data consist of quarterly, monthly, and daily data, the final merged data set contains one observation for each day. Therefore, the final data set contains many records that have missing observations. Figure 1 shows a snapshot of the final data.



recession	date	ICSA	l_icsa	PAYEMS	W875RX1	INDPRO	CMRMTSPL	ld_payemp	ld_indpro	ld_pinc	ld_mnfctr	GDPC1	ld_gdp
0	28APR13	.	.	.	.	.	.	.	.	.	.	.	.
0	29APR13	.	.	.	.	.	.	.	.	.	.	.	.
0	30APR13	.	.	.	.	.	.	.	.	.	.	.	.
0	01MAY13	.	.	135688	10894	98.9614	1135043	0.0012979353	0.0015958338	0.0016812362	0.0113242443	.	.
0	02MAY13	.	.	.	.	.	.	.	.	.	.	.	.
0	03MAY13	.	.	.	.	.	.	.	.	.	.	.	.
0	04MAY13	328000	12.700768887	.	.	.	.	.	.	.	.	.	.
0	05MAY13	.	.	.	.	.	.	.	.	.	.	.	.
0	06MAY13												

Figure 1: Structure of Final Data Set

The state space models and the SSM procedure do not require evenly spaced, complete observations as do many time series procedures. For this reason, state space modeling is a perfect tool for mixed frequency analysis. The five time series, ld_payemp (y_1), ld_pinc(y_2), ld_mnfctr (y_3), ld_indpro (y_4), and ld_gdp (y_5) are logged, differenced versions of the underlying economic variables. Their plots (Figure 2) show these series to be hovering around a constant level with some periods of deviation from this level. The plot of the sixth series, l_icsa (y_6), which is logged but not differenced, shows a pronounced nonstationary pattern.

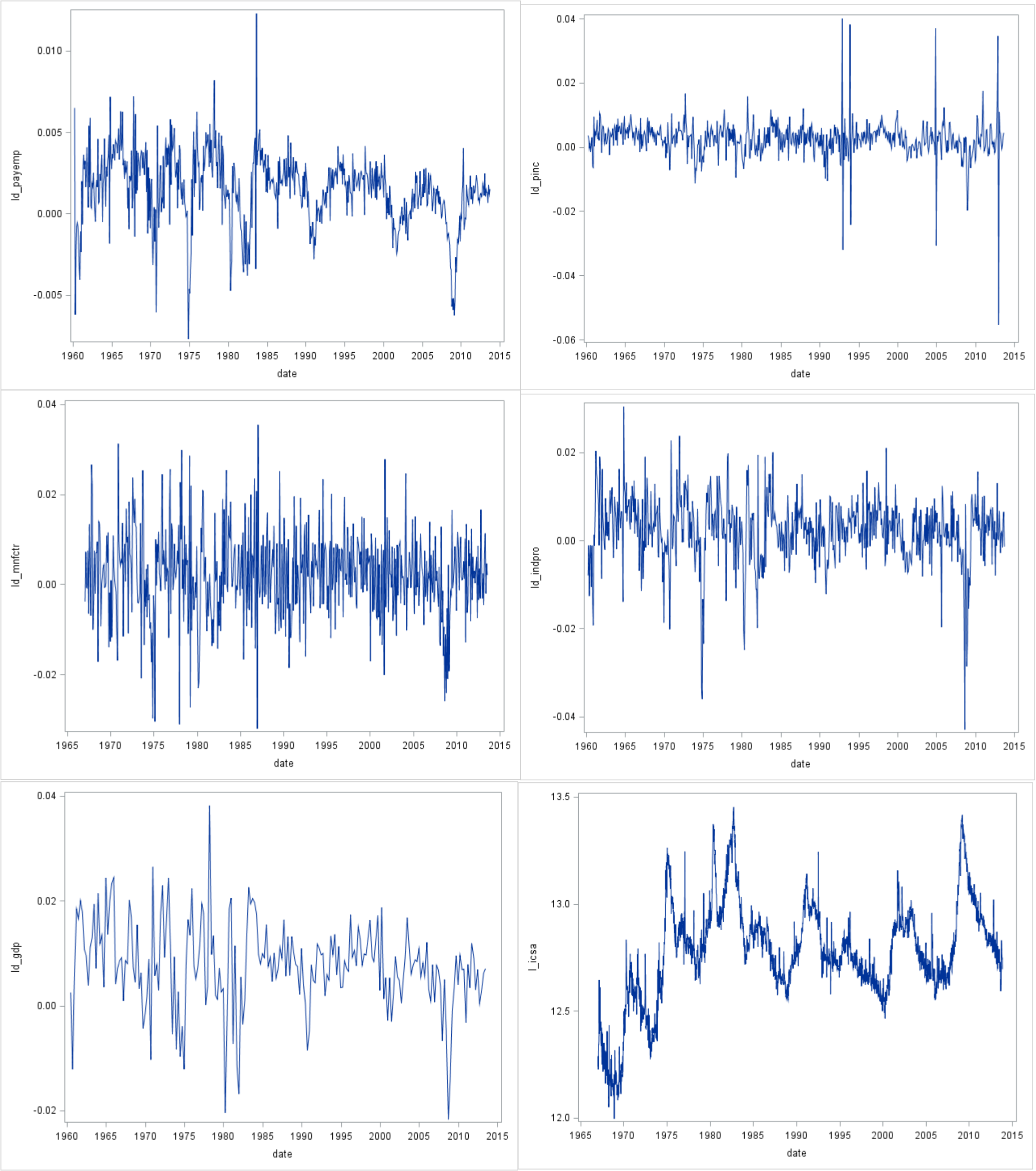


Figure 1: Graphs of ADS Components

These series can be considered as proxies for the health of the nation's economy. Each series provides information about the health of the overall system, but each measure is somewhat imperfect because of other factors. It is reasonable to assume that each series contains a component that is common to all other series and that, when appropriately weighted, conveys information about the underlying health of the economic system. This example models this common component as an integrated random walk (irw_t). The following *observation equations* formalize this system:

$$\begin{aligned} y_{it} &= intercept_i + \beta_i * irw_t + \epsilon_{it} \qquad\qquad 1 \le i \le 5 \\ y_{6t} &= \beta_6 * irw_t + \mu_t + \epsilon_{6t} \end{aligned}$$

For y_{1t} to y_{5t}, the only other terms in the model are the respective intercepts, $intercept_i$, and the random disturbances, ϵ_{it}. Because y_{6t} shows a pronounced nonstationary pattern, its model includes an additional term, μ_t, which is also modeled as an integrated random walk. For purposes of identification, the initial condition for irw_t is assumed to be zero. For the same reason, β_1 (the coefficient of irw_t in the model for y_{1t}) is assumed to be 1.

The underlying economic variables of the five time series, y_{1t} to y_{5t}, are positively correlated with economic activity. For example, as payroll activity and economic activity both increase at the same time, only y_6, the variable that measures initial jobless claims, is negatively correlated with economic activity. These correlations imply that, when $\beta_1$is assumed to be 1, the estimates of $\beta_2, \dots, \beta_5$ are expected to be positive and the estimate of β_6 is expected to be negative. In the terminology of factor modeling, irw_t is called a factor and $\beta_1, \dots, \beta_6$ are called the associated factor loadings.

USING THE SSM PROCEDURE TO SPECIFY THE MODEL

The following statements call the SSM procedure to estimate the factor loadings, which are shown in Table 2.

```
proc ssm data=econ opt(tech=activeset);
   id date interval=day;
   parms beta2-beta6; parms lv1-lv8;
   avar = exp(lv7);
   wnv1 = exp(lv1); wnv2 = exp(lv2); wnv3 = exp(lv3);
   wnv4 = exp(lv4); wnv5 = exp(lv5); wnv6 = exp(lv6);
   tvar = exp(lv8);
   zero = 0;
/* --- start of model spec ----*/
   state latent(2) t(g)=(1 1 0 1) cov(d)=(zero avar);
   comp c1 = latent[1];
   comp c2 = (beta2)*latent[1];
   comp c3 = (beta3)*latent[1];
   comp c4 = (beta4)*latent[1];
   comp c5 = (beta5)*latent[1];
   comp c6 = (beta6)*latent[1];

   irregular w1 variance=wnv1; int1 = 1; /*define variance and intercept*/
   model ld_payemp = int1 c1 w1; /*model for payroll employment*/

   irregular w2 variance=wnv2; int2 = 1; /*define variance and intercept*/
   model ld_pinc = int2 c2 w2; /*model for personal income*/

   irregular w3 variance=wnv3; int3 = 1; /*define variance and intercept*/
   model ld_mnfctr = int3 c3 w3; /*model for manufacturing sales*/

   irregular w4 variance=wnv4; int4 = 1; /*define variance and intercept*/
   model ld_indpro = int4 c4 w4; /*model for industrial production index*/

   irregular w5 variance=wnv5; int5 = 1; /*define variance and intercept*/
   model ld_gdp = int5 c5 w5; /*model for gdp*/

   irregular w6 variance=wnv6 ; trend t_icsa(ll) levelvar=0 slopevar=tvar;
   /*define variance and intercept*/
   model l_icsa = c6 t_icsa w6; /*model for initial jobless claims*/
run;
```

The estimates in Table 2 are statistically significant, and their signs are economically consistent. These estimates correspond to the β_i values in the model that is specified in the preceding statements, and they transmit information about economic activity to each series.

Parameter	**Estimate** **(Standard Error)**
beta1	1 (.)
beta2	1.15 (0.1276)
beta3	1.96 (0.2390)
beta4	2.48 (0.1646)
beta5	3.27 (0.2653)
beta6	—96.42 (9.5909)

Table 2: Factor Loading Estimates

An important and valuable feature of the SSM procedure is its ability to do post-estimate processing from within the procedure. In addition to forecasting, post-estimate processing includes combining estimates and variables in various ways to produce calculations of interest. The following statements demonstrate these features and can be placed within the previous SSM procedure call:

```
/*code within PROC SSM for evaluating components*/
 eval icsaPattern = c6 + t_icsa;
 /*--index is a scaled version of the common factor--*/
 eval Index = 1000*c1;
 comp slope = latent[2];
 eval IndexSlope = 1000*slope;
 output out=forecast1 press pdv;
/*end post-estimate processing*/
```

The OUTPUT statement provides a time series with which to examine the common trend components in the data. The time series of the common trend component are combined with the markers for a recession in the United States. The National Bureau of Economic Research publishes the official beginning and ending dates for recessions. They define a recession as two consecutive quarters that show negative real growth in Gross Domestic Product (GDP). The vertical bars in the two plots in Figure 3 signal these recessions. The plot on the left shows all time periods since 1960, and the plot on the right shows only the period that led up to the last two recessions until the present time.

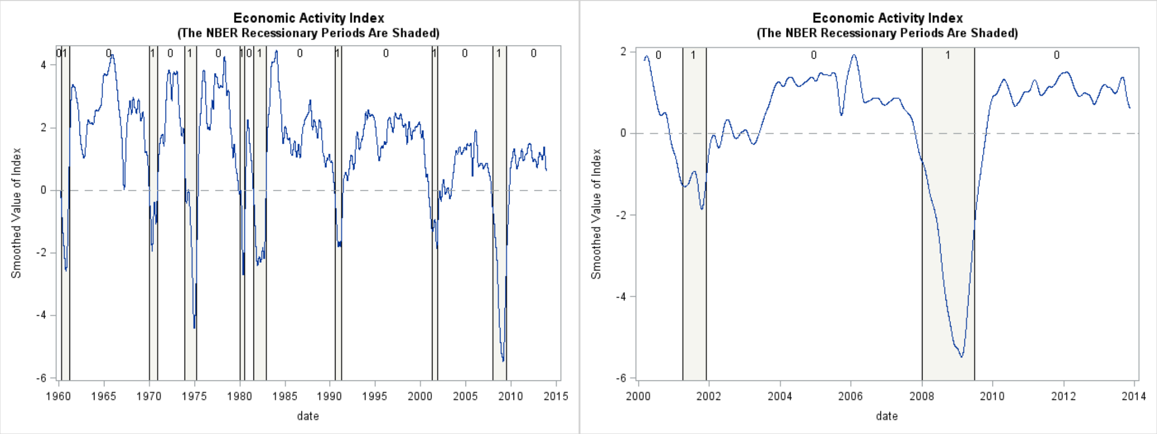


Figure 3: Graphs of Common Trends

The trend reveals a number of candidate intervention points. One intervention point might be when the index has a turning point; another might be when the index becomes negative. The plots in Figure 3 also indicate periods of recession. For the purpose of economic maintenance, lags in the typical response of any economic stimulus likely require intervention earlier than when the trend crosses into negative territory. By then it is too late. This is where expert judgment of the system must be used.

CONCLUSION

This paper shows how the new SSM procedure in SAS/ETS can be used to estimate parameters from a high-frequency multivariate time series. The tools in PROC SSM have many applications, including the ability to detect underlying trends in system data. This paper also showcases the new SASEFRED access engine in SAS/ETS, which enables users to dynamically generate data sets from the Federal Reserve Economic Data.

REFERENCES

Aruoba, B.S., Diebold, F.X., and Scotti, C. 2009. “Real-Time Measurement of Business Conditions."*Journal of Business & Economic Statistics*, 27(4), 417–427. http://www.philadelphiafed.org/research-and-data/real-time-center/business-conditions-index/

de Jong P. and Chu-Chun-Lin, S. 2003. “Smoothing with an Unknown Initial Condition." *Journal of Time Series Analysis*, 24, 141–148.

Durbin J. and Koopman, S.J. 2012. *Time Series Analysis by State Space Methods: Second Edition*.
38. Oxford University Press, Oxford.

Selukar, R. 2011. State Space Modeling Using SAS. *Journal of Statistical Software*, 41(12),1(13). http://www.jstatsoft.org/v41/i12.

Selukar, R. 2014. “PROC SSM: A New SAS Procedure for Linear State Space Modeling.” Working Paper.

ACKNOWLEDGMENTS

The author would like to thank Rajesh Selukar, the author of the SSM procedure, for his example and help with this paper. The author would also like to thank Anne Baxter for her comments and suggestions.

CONTACT INFORMATION

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

Kenneth Sanford
SAS Institute Inc.
100 SAS Campus Drive
Cary, North Carolina,
Kenneth.sanford@sas.com
www.sas.com/ets