Paper SAS456-2017

# Detecting and Adjusting Structural Breaks in Time Series and Panel Data Using the SSM Procedure

Rajesh Selukar, SAS Institute Inc.

## ABSTRACT

Detection and adjustment of structural breaks are an important step in modeling time series and panel data. In some cases, such as studying the impact of a new policy or an advertising campaign, structural break analysis might even be the main goal of a data analysis project. In other cases, the adjustment of structural breaks is a necessary step to achieve other analysis objectives, such as obtaining accurate forecasts and effective seasonal adjustment. Structural breaks can occur in a variety of ways during the course of a time series. For example, a series can have an abrupt change in its trend, its seasonal pattern, or its response to a regressor. The SSM procedure in SAS/ETS® software provides a comprehensive set of tools for modeling different types of sequential data, including univariate and multivariate time series data and panel data. These tools include options for easy detection and adjustment of a wide variety of structural breaks.

This paper shows how you can use the SSM procedure to detect and adjust structural breaks in many different modeling scenarios. Several real-world data sets are used in the examples. The paper also includes a brief review of the structural break detection facilities of other SAS/ETS procedures, such as the ARIMA, AUTOREG, and UCM procedures.

## INTRODUCTION

Monitoring changes in the normal behavior of a sequential process is an important problem in many fields. For example, timely detection of unexpected changes in sensor readings is important in industrial process control, evaluating the impact of an advertising campaign on product sales is important to marketers, and the impact of a volcanic eruption on atmospheric levels of carbon dioxide might interest climatologists. In some cases the change points are already known—the start date of an advertising campaign or the date of a volcanic eruption—and the main interest is in determining the nature of change at these change points. In other cases, both the change points and the nature of change at these points need to be determined. The nature of change in the process behavior can be of several different types; a misrecorded observation, a shift in the level or slope of the mean function, a change in seasonal characteristics, a change in the relationship with the input variables, and so on. After the change points and the nature of the change are determined, appropriate actions can be taken to account for these changes. This process of finding the change points and making the necessary adjustments is called change-point analysis.

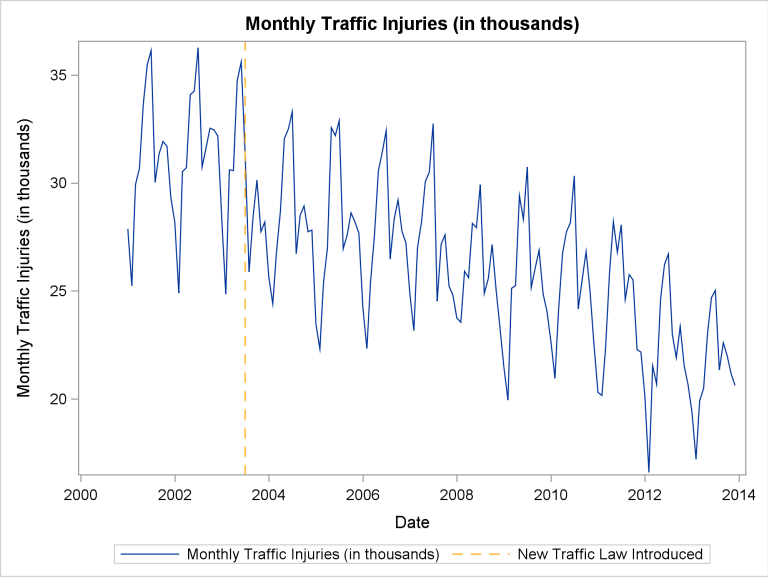
All change-point analysis methodologies are based on assuming a statistical model to describe the data generation process. In some change-point problems, the data generation process can be modeled by a relatively simple model, such as a mean plus error model. However, in most other situations, much more complex models are needed to describe the data generation process. This paper describes a general-purpose change-point analysis methodology, based on the SSM procedure, that assumes that the observation process follows a (linear) state space model (SSM). This model class is quite large and covers most commonly used data generation models in the change-point analysis literature. In particular, the data generation process could follow a model such as a piecewise linear regression model, a univariate or multivariate ARIMAX model, a univariate or multivariate unobserved components model (UCM), or any of the various panel data models. (You can get a sense of the scope of the data generation models that PROC SSM can handle by a quick review of the examples in the section “Examples: SSM Procedure” in the *SAS/ETS User’s Guide* (SAS Institute Inc. 2016).) Although this aspect is not emphasized in this paper, you can use the same methodology for change-point analysis of longitudinal data (see Selukar 2015 for examples of using PROC SSM to model hierarchical, longitudinal data).

### Impact of the Point System on Motor Vehicle Injuries in Italy

This example is based on a case study described in Pelagatti (2015, chap. 9, sec. 1). In July 2003, Italy introduced a new traffic monitoring system with the aim of improving traffic safety. The case study tried to answer the question,

was the monitoring system effective in reducing the number of traffic injuries? The time series plot in Figure 1 shows monthly traffic injuries for the span of January 2001 to December 2013. Visual inspection of the plot clearly shows that the series is seasonal and has an overall downward trend, which appears to be more pronounced after the intervention. The remainder of this section shows you how to use the SSM procedure to statistically evaluate the impact of this intervention.

**Figure 1** Monthly Traffic Injuries in Italy




The analysis begins with fitting a reasonable baseline model to the data that accounts for the most obvious features of the series but does not explicitly account for the intervention under consideration (July 2003). Based on the visual inspection of the series, the following model appears to be a reasonable baseline candidate:

$$y_t = \mu_t + \psi_t + \epsilon_t$$

Here $y_t$ denotes the response series (monthly traffic injuries, in thousands), $\mu_t$ denotes the trend component (modeled as an integrated random walk), $\psi_t$ denotes the monthly seasonal component, and $\epsilon_t$ is the noise component (modeled as a sequence of independent, identically distributed, zero-mean, Gaussian random variables). The components $\mu_t$, $\psi_t$, and $\epsilon_t$ are assumed to be statistically independent. The following statements show you how to use the SSM procedure to fit this model to the data:

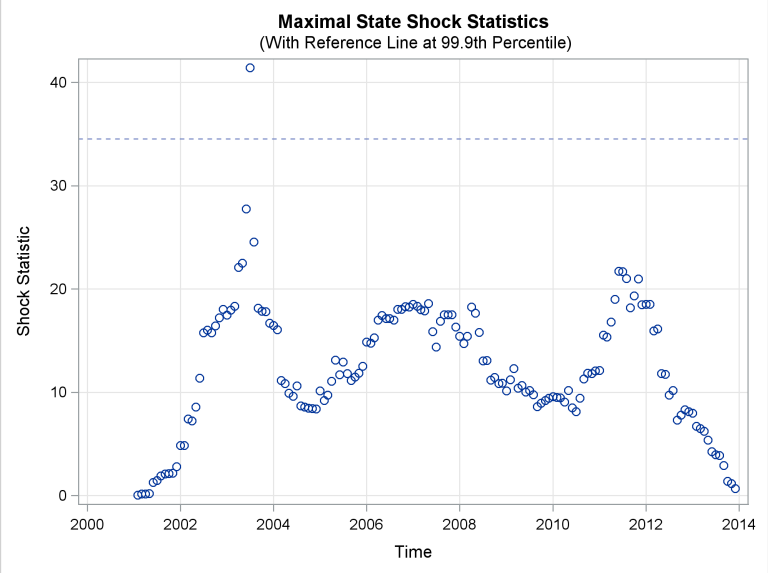
```
proc ssm data=Italy plot=maxshock;
   id date interval=month;
   trend irw(ll) variance=0 checkbreak;
   state seasonState(1) type=season(length=12)
       cov(g) checkbreak(overall);
   comp season = seasonState[1];
   irregular wn;
   model injured=irw season wn;
run;
```

The PROC SSM statement specifies the input data set, **Italy**, that contains the analysis variables. The ID statement specifies the time index of the observations. The TREND statement declares **irw** to be a locally linear trend component (signified by the *ll* option in **`irw(ll)`**) and the disturbance variance that is associated with the level equation to be 0, which makes it an integrated random walk. This specification corresponds to $\mu_t$ in the model. The next two statements, STATE and COMP, create the seasonal specification $\psi_t$: the STATE statement declares **seasonState** to be a state vector that is associated with a seasonal pattern of length 12, and the COMP statement declares **season** to be the seasonal component formed by appropriate linear combination of the elements of **seasonState**. The IRREGULAR statement declares **wn** to be the white noise sequence. Finally, the MODEL statement completes the model specification, which corresponds to the observation equation: $y_t = \mu_t + \psi_t + \epsilon_t$. The PLOT=(MAXSHOCK) option in the PROC SSM statement and the CHECKBREAK option in the TREND and STATE statements produce

the change-point diagnostics. The state vector that is associated with this model is 13-dimensional. It is formed by joining two subvectors: a 2-dimensional vector that is associated with the trend $\mu_t$, and an 11-dimensional vector that is associated with the monthly seasonal component $\psi_t$. The PLOT=(MAXSHOCK) results in a time series plot of chi-square statistics (with 13 degrees of freedom) that shows likely change points in the evolution of this 13-dimensional time-varying state vector. The CHECKBREAK option in the TREND and STATE statements provides similar change-point diagnostics for state vectors that are associated with $\mu_t$ and $\psi_t$, respectively. (For more information about these statements, see the section "Syntax: SSM Procedure" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).)

As a result of fitting this model, you can obtain the estimates of trend $\hat{\mu}_t$, season $\hat{\psi}_t$, and noise $\hat{\epsilon}_t$; forecasts of response values $\hat{y}_t$; and many more details, including different types of residual analyses. In the case of this model, a casual inspection of the residual plots (not shown) does not reveal the series to be very troublesome. However, the structural break diagnostic plot shown in Figure 2 (which is produced because the PLOTS=(MAXSHOCK) option is specified) clearly shows a possible break at July 2003. The diagnostics that are shown in this plot check for an unexpected change in the overall model state vector, which for this model means change in the dynamics of $\mu_t$, $\psi_t$, or both $\mu_t$ and $\psi_t$.

**Figure 2** Structural-Break Chi-Square Statistics Plot




The summary table of likely change points for the trend component, shown in Figure 3, also shows July 2003 as a likely break point for the first element of the state that underlies the trend. However, the break diagnostics that are produced by the CHECKBREAK option in the STATE statement that is associated with the seasonal component $\psi_t$ (not shown) do not show July 2003 as a break point. In summary, July 2003 is seen as the most likely change point, and the change seems to be caused by a change in the first element of the state that is associated with $\mu_t$.

**Figure 3** Discovered Break Locations

**Elementwise Break Summary for irw**

| ID | Element Index | Z Value | Pr > \|z\| |
|---|---|---|---|
| **JUL2003** | 1 | -5.55 | <.0001 |
| **JUN2003** | 1 | -3.97 | <.0001 |

Now that July 2003 has been discovered as a likely change point, the next step is to adjust the baseline model to account for this change. The following modification of the baseline model is suggested by Pelagatti (2015, chap. 9, sec. 1),

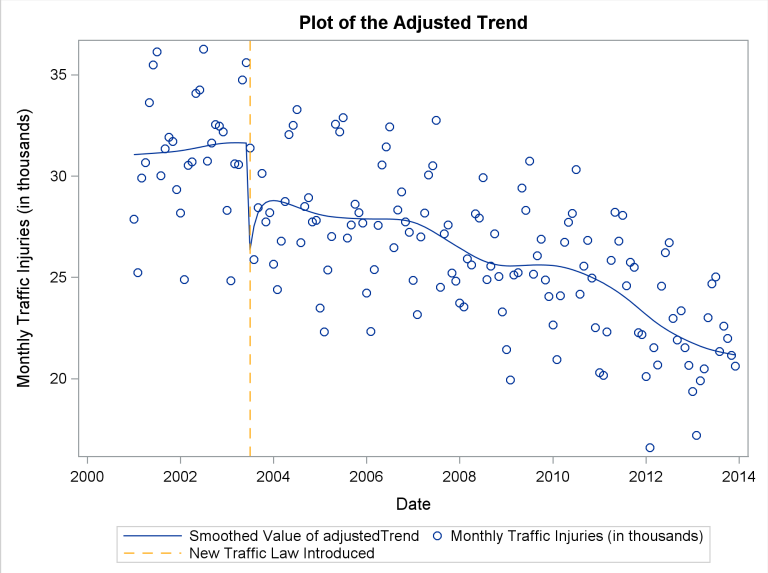
$$y_t = \text{shift\_Jul03}\ \beta + \xi_t + \mu_t + \psi_t + \epsilon_t$$

where the components $\mu_t$, $\psi_t$, and $\epsilon_t$ are the same as in the baseline model and the new terms are a regression term that is associated with **shift_Jul03** (which is 0 before July 2003 and 1 thereafter) and a transfer function term $\xi_t$ that satisfies the relation $\xi_t = \delta\xi_{t-1} + \omega$ point_Jul03, where the dummy regressor **point_Jul03** is 1 at July 2003 and 0 otherwise. The following statements show you how to fit this model to the data:

```
proc ssm data=italy;
   id date interval=month;
   shift03 = date >= '01jul2003'd;
   point03 = (date = '01jul2003'd);
   trend irw(ll) variance=0 checkbreak;
   state tfSt(1) T(g) w(g)=(point03);
   comp tfInput = tfSt[1];
   state seasonState(1) type=season(length=12)
       cov(g) checkbreak(overall);
   comp season = seasonState[1];
   irregular wn;
   model injured=shift03 tfInput irw season wn;
   eval tfSum = shift03 + tfInput;
   eval  adjustedTrend = irw + shift03 + tfInput;
   eval fullFit = irw + shift03 + tfInput + season;
   output out=forItaly pdv;
run;
```

This model passes basic diagnostic checks (not shown), and no further modifications seem necessary. Figure 4 shows the adjusted trend curve (shift_Jul03 $\beta + \xi_t + \mu_t$).

**Figure 4** Estimated Trend Pattern for Monthly Traffic Accidents in Italy




## State Space Model Notation and Terminology

In order to formalize the notation used throughout this paper, consider the following SSM:

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t & \boldsymbol{\epsilon}_t &\sim \mathrm{N}(\mathbf{0}, \sigma^2) & &\text{Observation equation}\\
\boldsymbol{\alpha}_t &= \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{W}_t\boldsymbol{\gamma} + \boldsymbol{\eta}_t & \boldsymbol{\eta}_t &\sim \mathrm{N}(\mathbf{0}, \mathbf{Q}) & &\text{State transition equation}\\
\boldsymbol{\alpha}_1 &= \boldsymbol{\alpha} & & \text{unknown} & &\text{Initial condition}
\end{aligned}
$$

Many models discussed in this paper, including the models in the section "Impact of the Point System on Motor Vehicle Injuries in Italy" on page 1, are special cases of the preceding SSM. This SSM is a dynamic version of the standard regression model in which the overall regression vector is divided into two parts: a time-invariant part, $\boldsymbol{\beta}$, and a

time-varying part, $\boldsymbol{\alpha}_t$. The observation equation shows that the response value at time $t$, $y_t$, is decomposed into three parts: $\mathbf{X}_t\boldsymbol{\beta}$ and $\mathbf{Z}_t\boldsymbol{\alpha}_t$ are the contributions from the regression variables that are associated with the time-invariant and time-varying regression coefficients, respectively; and $\epsilon_t$ is a value from a sequence of independent, zero-mean, Gaussian noise variables. The time-varying part, $\boldsymbol{\alpha}_t$, is called the state, which evolves in time as a first-order vector autoregression. The elements of $\boldsymbol{\alpha}_t$ often correspond to key features of the time series—for example, the time-varying mean, the time-varying seasonal factors, and so on. The state transition equation, which describes the time evolution of $\boldsymbol{\alpha}_t$, shows that the new instance of $\boldsymbol{\alpha}_t$ is obtained by multiplying its previous instance, $\boldsymbol{\alpha}_{t-1}$, by a square matrix $\mathbf{T}$ (called the state transition matrix) and by adding two more terms: a regression term $\mathbf{W}_t\boldsymbol{\gamma}$, where $\mathbf{W}_t$ denotes the design matrix and $\boldsymbol{\gamma}$ is the regression vector, and a random disturbance vector $\boldsymbol{\eta}_t$. The state disturbance vectors $\boldsymbol{\eta}_t$ are assumed to be independent, zero-mean, Gaussian random vectors with covariance $\mathbf{Q}$. The state transition equation is initialized at $t = 1$ with an unknown vector $\boldsymbol{\alpha}$ (which is qualitatively similar to the regression vector $\boldsymbol{\beta}$). This type of initial condition is called a *diffuse* initial condition in the state space modeling literature. Many useful time series models are obtained by appropriate choices of $\mathbf{T}$, $\mathbf{Q}$, and $\sigma^2$ and by different choices of the design matrices $\mathbf{X}_t$, $\mathbf{Z}_t$, and $\mathbf{W}_t$. You can specify, fit, and diagnose this and more complex SSMs by using the SSM procedure. (For information about the most general SSM that you can specify, see the section "State Space Model and Notation" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).)

Very often the state vector is composed of subsections that are statistically independent. For example, suppose that $\boldsymbol{\alpha}_t$ can be divided into two disjoint subsections, $\boldsymbol{\alpha}_t^a$ and $\boldsymbol{\alpha}_t^b$, that are statistically independent. This division entails a corresponding block structure to the system matrices $\mathbf{T}$, $\mathbf{W}_t$, and $\mathbf{Q}$ that govern the state equations. In this case the term $\mathbf{Z}_t\boldsymbol{\alpha}_t$ that appears in the observation equation also splits into the sum $\mathbf{Z}_t^a\boldsymbol{\alpha}_t^a + \mathbf{Z}_t^b\boldsymbol{\alpha}_t^b$ for the appropriately partitioned matrices $\mathbf{Z}_t^a$ and $\mathbf{Z}_t^b$. The model specification syntax of the SSM procedure makes building an SSM from such smaller pieces easy. Throughout this paper, the linear combinations of the state subsections (such as $\mathbf{Z}_t^a\boldsymbol{\alpha}_t^a$) that appear in the observation equation are called components. An SSM specification in PROC SSM is created by combining separate component specifications. In general, you specify a component in two steps: first you define a state subsection $\boldsymbol{\alpha}_t^a$, and then you define a matching linear combination $\mathbf{Z}_t^a\boldsymbol{\alpha}_t^a$. You do this by using a matching combination of STATE and COMP statements. For some special components, such as some commonly needed trend components, you can combine these two steps into one keyword specification by using the TREND statement. (For more information about these statements, see the section "Syntax: SSM Procedure" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).)

## SSM-BASED STRUCTURAL BREAK DETECTION AND ADJUSTMENT

The detection and adjustment of change points are an iterative process. Starting with a reasonable baseline model, it cycles through the following steps:

1. Fit and diagnose the baseline model, including the detection of likely change points and the nature of their change. You can use the SSM procedure to detect the following types of changes:
   - unusual series values, which are called additive outliers (AO)
   - unexpected change in the state vector at one or more time points
2. Adjust the baseline model so that a few of these likely change points are accounted for. Which change points are chosen for adjustment depends on the context of the problem. The adjusted model becomes the new baseline model.
3. Stop when the baseline model appears satisfactory, or else return to step 1.

### Change-Point Detection

This discussion of detecting change points assumes that the data generation process follows the fitted baseline model. The treatment of additive outliers and structural breaks that is described in this section is based on De Jong and Penzer (1998).

Let $\mathrm{AO}_t = y_t - E(y_t|\mathbf{Y}^t)$ denote the difference between the observed response value $y_t$ and its estimate or prediction by using all the data except $y_t$, which is denoted by $\mathbf{Y}^t$. A large value of $\mathrm{AO}_t$ signifies that the observed response value, $y_t$, is unusual relative to the rest of the sample (according to the fitted baseline model). In the literature, $\mathrm{AO}_t$ are referred by a few different names: delete-one cross validation errors, additive outliers, or simply prediction errors. The SSM procedure prints a summary table of extreme additive outliers by default. In addition,

you can request the plotting of the standardized additive outliers, and they can be output to a data set. (For more information, see the section "Delete-One Cross Validation and Structural Breaks" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).) The remainder of this section deals with detection of unexpected change in the state vector at one or more time points.

In the SSM procedure, you can request diagnostic tests that check for unexpected change in the state vector in a few different ways:

- The PLOT=MAXSHOCK option in the PROC SSM statement and the MAXSHOCK option in the OUTPUT statement detect unexpected change in the model state vector as a whole.
- The CHECKBREAK option in the STATE and TREND statements detects change in a specific element of the state vector (CHECKBREAK(ELEMENTWISE)) and change in the specific subsection of the state vector (CHECKBREAK(OVERALL)).

As an example, recall the change-point analysis of the baseline model in the section "Impact of the Point System on Motor Vehicle Injuries in Italy" on page 1. This model has the following state space form,

$$
\begin{array}{lll}
y_t = \mathbf{Z}\boldsymbol{\alpha}_t + \epsilon_t & \epsilon_t \sim \mathrm{N}(0, \sigma^2) & \text{Observation equation} \\
\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t & \boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}) & \text{State transition equation} \\
\boldsymbol{\alpha}_1 = \boldsymbol{\alpha} & \text{unknown} & \text{Initial condition}
\end{array}
$$

where the 13-dimensional state vector $\boldsymbol{\alpha}_t$ is formed by joining two vectors: a 2-dimensional vector (such as $\boldsymbol{\alpha}_t^{\mu}$) that corresponds to the integrated random walk trend $\mu_t$, and an 11-dimensional vector (such as $\boldsymbol{\alpha}_t^{\psi}$) that corresponds to the monthly seasonal component $\psi_t$. The vectors $\boldsymbol{\alpha}_t^{\mu}$ and $\boldsymbol{\alpha}_t^{\psi}$ themselves follow separate state transition equations that depend on the transition matrices $\mathbf{T}^{\mu}$ and $\mathbf{T}^{\psi}$ (for example) and the disturbance covariances $\mathbf{Q}^{\mu}$ and $\mathbf{Q}^{\psi}$ (for example), respectively. The transition matrix of the overall state $\boldsymbol{\alpha}_t$ has a block diagonal form ($\mathbf{T} = \mathrm{Diag}(\mathbf{T}^{\mu}, \mathbf{T}^{\psi})$), and the same holds for the disturbance covariance ($\mathbf{Q} = \mathrm{Diag}(\mathbf{Q}^{\mu}, \mathbf{Q}^{\psi})$). Similarly, the 13-dimensional design vector $\mathbf{Z}$ also splits into two blocks, $\mathbf{Z}^{\mu}$ and $\mathbf{Z}^{\psi}$, such that $\mu_t = \mathbf{Z}^{\mu}\boldsymbol{\alpha}_t^{\mu}$ and $\psi_t = \mathbf{Z}^{\psi}\boldsymbol{\alpha}_t^{\psi}$.

After this model is fitted, an important diagnostic question to consider is whether the 13-dimensional state vector experienced an unexpectedly large change at some time point $t = t_0$. Formally, denoting the change vector at $t_0$ by $\boldsymbol{\gamma}$, this question can be formulated as a test of the null hypothesis, $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, in the following perturbed transition equation,

$$\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \mathbf{I}_{(t=t_0)}\boldsymbol{\gamma} + \boldsymbol{\eta}_t$$

where $\mathbf{I}_{(t=t_0)}$ is the 13-dimensional identity matrix when $t = t_0$, and a zero matrix otherwise. Recall the structural break diagnostic plot shown in Figure 2, which is produced as a result of the use of the PLOT=(MAXSHOCK) option in the PROC SSM statement. This plot shows the test statistics (which follow a chi-square distribution with 13 degrees of freedom) that are associated with testing $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ at each time in the sample. Significant peaks in this plot indicate the likely change-point locations. These locations signify places where the state transition equation experienced significant shocks. Many times, several neighboring points around the significant peak are also statistically significant. These can often be attributed to the same change-point phenomenon (for example, see the plot in Output 1.1 in Example 1).

Continuing with the same logic, you can test for change in specific sections or specific elements of the state. You can obtain this type of change diagnostics by using the appropriate CHECKBREAK option in the STATE and TREND statements. For example, specifying the CHECKBREAK(OVERALL) option in the STATE statement corresponds to testing $H_0 : \boldsymbol{\gamma}^{\psi} = \mathbf{0}$ in the perturbed transition equation for the 11-dimensional state subsection $\boldsymbol{\alpha}_t^{\psi}$ that corresponds to the seasonal component $\psi_t$. On the other hand, specifying the CHECKBREAK(ELEMENTWISE) option (which defaults to just CHECKBREAK) in the TREND statement corresponds to testing elementwise change in the 2-dimensional state $\boldsymbol{\alpha}_t^{\mu}$ that underlies $\mu_t$.

In the case of change in a subsection as a whole, the test statistics follow a chi-square distribution with the degrees of freedom equal to the dimension of the subsection. On the other hand, the elementwise change test statistics follow the standard normal distribution (or their squares follow a chi-square distribution with one degree of freedom). As a default, the CHECKBREAK option (with either the OVERALL or ELEMENTWISE suboption) produces a summary table of the most significant change points; for example, see the table in Figure 3. You can also print a table of test

statistics at all the time points in the sample by using the PRINT=BREAKDETAIL option in the STATE or TREND statement.

**NOTE:** Starting with SAS/ETS 14.2 (the latest version as of this writing), you can simplify the search of change points by using the new BREAKPEAKS option in the PROC SSM statement, which enables you to ignore the neighboring points around the change point even if they have significant change-point statistics, because they are usually associated with the same change-point phenomenon (for more information, see the section "Syntax: SSM Procedure" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016)). For example, the BREAKPEAKS option in the PROC SSM statement would suppress the second change point at June 2003 that is shown in the summary table in Figure 3 because it is a neighbor of the more significant change point at July 2003. The examples in this paper do not use the BREAKPEAKS option because some users might not have access to this latest SAS/ETS version.

### Change-Point Adjustment

After the likely change points and their nature have been discovered, the next step is to adjust the baseline model to account for these breaks. If you choose the baseline model carefully, the change-point detection step is somewhat mechanical. The adjustment phase, on the other hand, requires careful consideration of the overall context of the data generation process. For example, unless a good explanation is available, it is preferable not to adjust many of the additive outliers and structural breaks. For the most part, the dynamic nature of the time series models enables them to gradually adjust for minor breaks in the data generation process. Undue adjustment of change points can lead to models that fit the historical data well but fail to forecast or extrapolate properly and thus produce overly optimistic forecast confidence bands. In the cases where you do decide to adjust some change points, it is preferable to make the adjustment a few change points at a time. This is because model fitting after each adjustment often leads to changes in the old parameters and can reveal different change points that were masked in the earlier detection step. The rest of this section briefly describes how to adjust the model to account for a change point.

You can adjust the additive outliers either by setting the corresponding response value to missing or by including an appropriate dummy variable in the MODEL statement that is associated with that particular response variable. You can adjust for the change in the state element or state subsection in several different (and sometimes equivalent) ways, including the following:

- Use appropriate intervention variables in the MODEL statement. This is easiest to do when the changes in the state correspond to changes in the process mean; for example, see Example 4.
- Use appropriate dummy variables in the state transition equation as shown in some of the examples in this paper. For illustrations, see Example 1 and Example 2.
- Adjust the variance of the disturbance term in the state transition equation near the change point so that the state vector estimate before the change point is down-weighted when the Kalman filter updates the state estimate after the change point. This is illustrated in Example 3.
- Make a more complex adjustment that involves intervention variables and new components such as transfer function inputs. This is illustrated in the example in the introductory section, "Impact of the Point System on Motor Vehicle Injuries in Italy" on page 1.

## COMPARISON WITH OTHER PROCEDURES

Several other time series analysis procedures in SAS/ETS also offer change-point diagnostics. For example, you can use the ARIMA, AUTOREG, and UCM procedures for change-point analysis. In fact, the change-point diagnostics in the ARIMA and UCM procedures is largely based on the same state space model methodology that this paper describes. The change-point detection methods in PROC AUTOREG have a slightly different flavor. (For more information about the change-point detection functionality of PROC AUTOREG, see the section "Testing for Structural Change" in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).) In any event, these three procedures are applicable only in the univariate time series settings. Moreover, only limited types of structural changes—primarily change in the series trend or change in the regression parameters—can be diagnosed. PROC AUTOREG and PROC UCM also offer change-point detection based on CUSUM and CUSUMSQ statistics, which are computed using the cumulative sum and cumulative sum of squares of one-step-ahead residuals, respectively. These diagnostics are useful for change-point detection but are not very useful for identifying the nature of the change at these change points. Currently, the SSM procedure does not provide CUSUM- and CUSUMSQ-based change-point detection. Nevertheless, you can

use PROC SSM to do this type of change-point detection by computing the CUSUM and CUSUMSQ statistics as a postprocessing step because the one-step-ahead residuals can be output to a data set. The VARMAX procedure (which models multivariate time series) and the PANEL procedure (which analyzes panel data) do not currently offer formal change-point diagnostics.

## EXAMPLES

The examples in this paper are chosen to explain the SSM-based change-point analysis methodology in a variety of situations. The examples are relatively simple, and the change points are known in advance, making it easier to understand the main principles without lengthy explanations.

Note that the documentation of the SSM procedure also has some examples that you might find useful: detection of shift in the trend of a multivariate time series (Example 8) and detection of break in the transfer function component of an ARIMAX model (Example 10) in the SSM procedure chapter in the *SAS/ETS User's Guide* (SAS Institute Inc. 2016).

### Example 1: Detecting Change in the Regression Coefficient over Time

The data generation process—piecewise linear regression (also called the broken-stick model)—that this example considers is quite simple, but it is the basis of change-point detection in many natural phenomena that involve phase transition. In order to explain the main idea, the simplest version of this model is sufficient. The same methodology can be used in more general cases. For example, Das et al. (2016) describe a change-point detection methodology for regression parameters in a broken-stick model for longitudinal data; you can use the methodology described in the current paper in their data generation setting as well.

The following DATA step statements create a data set, **regBreak**, that contains the simulated values of variables **z, y, alpha,** and **t**, for $1 \leq t \leq 60$:

```
data regBreak;
   do t=1 to 60;
      if t > 20 & t <= 40 then alpha=3;
      else alpha=1;
      z = 2*ranuni(1) + 2;
      y = 10.0 + alpha*z + 0.5*rannor(1);
      output;
   end;
run;
```

The variables **y** (response), **z** (predictor), and **t** (time index) are observed, whereas **alpha** is assumed to be latent. The data are generated so that **y** satisfies the relation

$$y_t = 10.0 + z_t\alpha_t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, 0.25)$$

and the time-varying (piecewise constant) regression coefficient $\alpha_t$ has the following form:

$$\begin{aligned} \alpha_t &= 3.0 \ \text{ for } \ 21 \leq t \leq 40 \\ &= 1.0 \ \text{ otherwise} \end{aligned}$$

The goal is to detect the changes in the regression coefficient $\alpha_t$ over time, starting with the baseline model $y_t = \mu + z_t\alpha + \epsilon_t, \ \ \epsilon_t \sim \mathrm{N}(0, \sigma^2)$, which assumes that $\alpha_t$ is constant over time.

The first step is to fit the baseline model to the data and study the break diagnostics. Depending on the chosen state space form, you can specify the regression model in PROC SSM in a few different ways. The following state space form of the regression model treats the regression coefficient of **z** as part of the state, which makes it convenient to test the possibility that this coefficient might change over time:

$$\begin{aligned} y_t &= \mu + z_t\alpha_t + \epsilon_t && \text{Observation equation} \\ \alpha_t &= \alpha_{t-1} && \text{State transition equation} \\ \alpha_1 &= \alpha && \text{Initial condition} \end{aligned}$$

On the other hand, the intercept $\mu$, the coefficient of the intercept variable (identically equal to 1), is treated as a simple regression coefficient (not part of the state) because there is no plan to test the possibility that it might change

over time. Note that this is merely a restatement of the usual regression model, because the initial condition $\alpha_1 = \alpha$ and the state transition equation $\alpha_t = \alpha_{t-1}$ together imply that $\alpha_t = \alpha \forall t$; that is, $\alpha_t$ is a time-invariant constant. The following statements show you how to specify this model in PROC SSM:

```
proc ssm data=regBreak plots=(maxshock);
   id t;
   one = 1.0;
   state zCoeff(1) T(I) A1(1) checkbreak;
   comp zTerm = zCoeff * (z);
   irregular error;
   model y = one zTerm  error;
   output break(maxpct=5);
run;
```

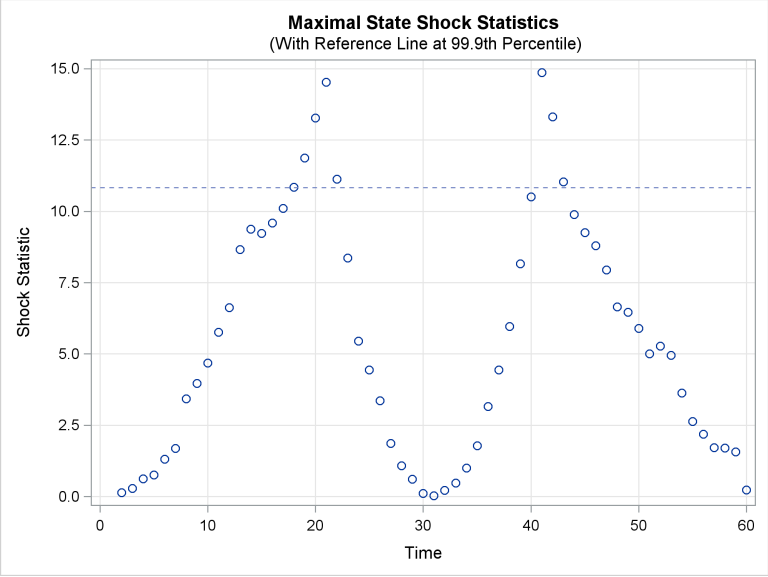
The PROC SSM statement specifies the input data set, **regBreak**, that contains the analysis variables; the PLOTS=(MAXSHOCK) option is associated with the structural break diagnostics. The ID statement specifies the time index of the observations, and the DATA step programming statement declares **one** to be a constant variable with the value 1.0. The STATE statement declares **zCoeff** as a one-dimensional state subsection with identity as the state transition matrix (T(I) option), zero disturbance error (signified by the absence of the options that specify the disturbance covariance), and a diffuse initial condition (A(1) option). In effect, the **zCoeff** specification corresponds to the state transition equation and the initial condition for $\alpha_t$. Furthermore, the CHECKBREAK option prints the break diagnostics summary. The COMP statement creates the $z_t \alpha_t$ term to be used later in the observation equation (MODEL statement). The IRREGULAR statement creates **error** as a noise term to be used in the MODEL statement. The state space model specification is completed by specifying the observation equation. Finally, the BREAK option in the OUTPUT statement limits the maximum number of break locations to be displayed in the break summary table to 3 (5% of 60, the number of time points in the data set).

After running these statements, you obtain the usual output associated with fitting a standard regression model; for example, the parameter estimates turn out to be $\hat{\mu} = 6.95, \hat{\alpha} = 2.7$, and $\hat{\sigma}^2 = 8.58$. In addition, and what is more relevant for this discussion, you also get structural break information about the state **zCoeff**. This structural break analysis is based on a series of significance tests that determine whether the postulated state transition for **zCoeff** ($\alpha_t = \alpha_{t-1}$) is in fact perturbed at some point $t = t_0$. That is, for some shift of size $\gamma$, the new state transition relation is

$$\alpha_t = \alpha_{t-1} + I_{(t=t_0)}\gamma$$

where $I_{(t=t_0)}$ is a dummy variable that takes a value of 1 if $t = t_0$ and 0 otherwise. The plot in Output 1.1, which results from specifying the PLOTS=(MAXSHOCK) option, shows the chi-square test statistics that are associated with the hypothesis $H_0 : \gamma = 0$ for each possible change point in the sample ($t_0 = 1, 2, \ldots, 60$).

**Output 1.1** Structural-Break Chi-Square Statistic Plot



Based on this plot, it is easy to see that 21 and 41 are the most likely break points. The summary table in Output 1.2 reports these two significant peaks as the likely break points for **zCoeff**.

**Output 1.2** Discovered Break Locations

**Elementwise Break Summary for zCoeff**

| ID | Element Index | Z Value | Pr > \|z\| |
|---|---|---|---|
| **41** | 1 | -3.85 | 0.0001 |
| **21** | 1 | 3.81 | 0.0001 |
| **42** | 1 | -3.65 | 0.0003 |

Based on this analysis, you can postulate a modified state transition equation for **zCoeff**:

$$\alpha_t = \alpha_{t-1} + I_{(t=21)}\gamma_1 + I_{(t=41)}\gamma_2, \quad 2 \le t \le 60$$

This transition equation (along with the initial condition $\alpha_1 = \alpha$) implies that $\alpha_t$ remains equal to $\alpha$ until $t = 20$, then shifts to $(\alpha + \gamma_1)$ at $t = 21$ and remains there until $t = 40$. Finally, from $t = 41$ onward it remains at $(\alpha + \gamma_1 + \gamma_2)$. This model is in fact the true data generation model (with unknown parameters). The following statements show how to modify the earlier baseline model program to fit this revised model:

```
proc ssm data=regBreak;
   id t;
   one = 1.0;
   break1 = (t = 21);
   break2 = (t = 41);
   state zCoeff(1) T(I) W(g)=(break1 break2) A1(1);
   comp zTerm = zCoeff * (z);
   irregular error;
   model y = one zTerm  error;
   comp zCoeff1 = zCoeff[1];
   output out=breakFor;
run;
```

The table in Output 1.3 shows the estimated shift sizes $\hat{\gamma}_1 = 1.92$ and $\hat{\gamma}_2 = -1.95$, which are reasonably close to their true values of 2.0 and –2.0.

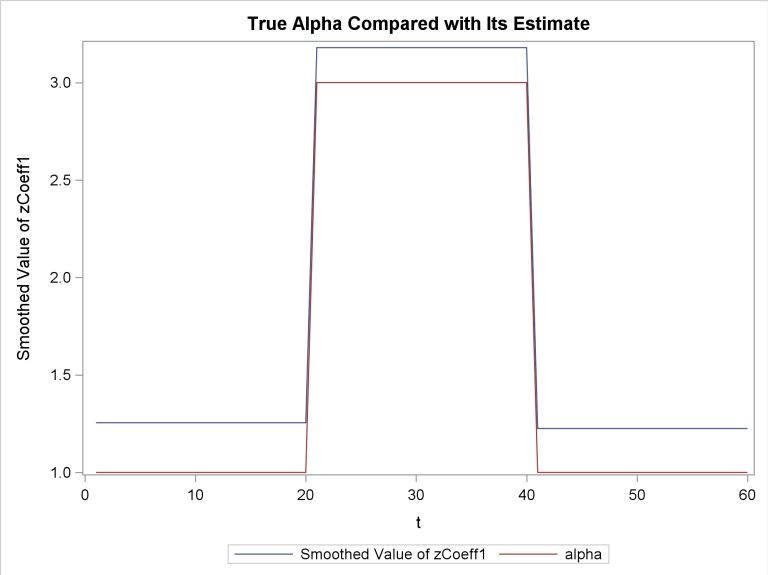
**Output 1.3** Estimated Break Sizes

**Estimate of the State Equation Regression Vector**

| State | Element Index | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **zCoeff** | **1** | 1.92 | 0.0488 | 39.40 | <.0001 |
| **zCoeff** | **2** | -1.95 | 0.0494 | -39.56 | <.0001 |

The other estimated parameters turn out to be $\hat{\mu} = 9.35, \hat{\alpha} = 1.255$, and $\hat{\sigma}^2 = 0.23$, which are also close to their true values (10.0, 1.0, and 0.25, respectively). Finally, the plot in Output 1.4 shows the comparison between the true time-varying $\alpha_t$ and its estimate.

**Output 1.4** Comparison of True Time-Varying Coefficient and Its Estimate




### Example 2: Detecting Change in the Common Trend of a Panel of Time Series

This example analyzes the world gasoline demand data, which are a panel of annual time series of gasoline consumption per car in 18 member countries of the Organisation for Economic Co-operation and Development (OECD). These data are discussed in Baltagi (2013). The variables in the data set, **Gas**, are **year**, **country**, **lgaspcar** (log of consumption per car), **lincomep** (log of per capita income), **lrpmg** (log of real price of gasoline), and **lcarpcap** (log of per capita number of cars). In addition, the variable **cIndex**, which contains an integer between 1 and 18 that uniquely identifies each country, is also part of the data set. The goal of this analysis is to see whether the oil embargo of 1973–1974 had any effect on gasoline consumption per car. The model

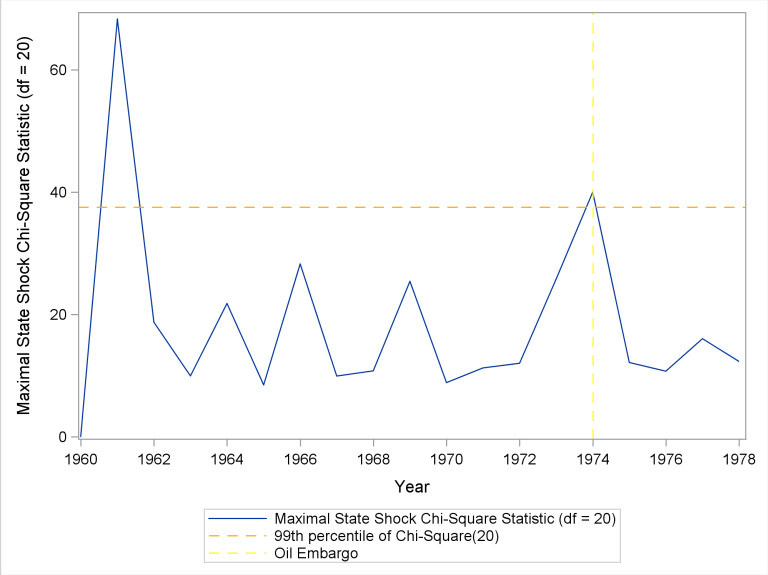
$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mu_t + \eta_{it} + \epsilon_{it}$$

decomposes $y_{it}$, the log of consumption per car in year $t$ for country with **cIndex** equal to $i$, into $\mathbf{X}_{it}\boldsymbol{\beta}$, which denotes the effect of the regressors **lincomep**, **lrpmg**, and **lcarpcap**; $\mu_t$ denotes the common time trend, which is modeled as an integrated random walk; $\eta_{it}$ denote country-specific corrections to the global trend that are modeled as random walks with nondiffuse, zero-mean, initial conditions; and $\epsilon_{it}$ denotes white noise. The following statements show you how to fit this model to the data:

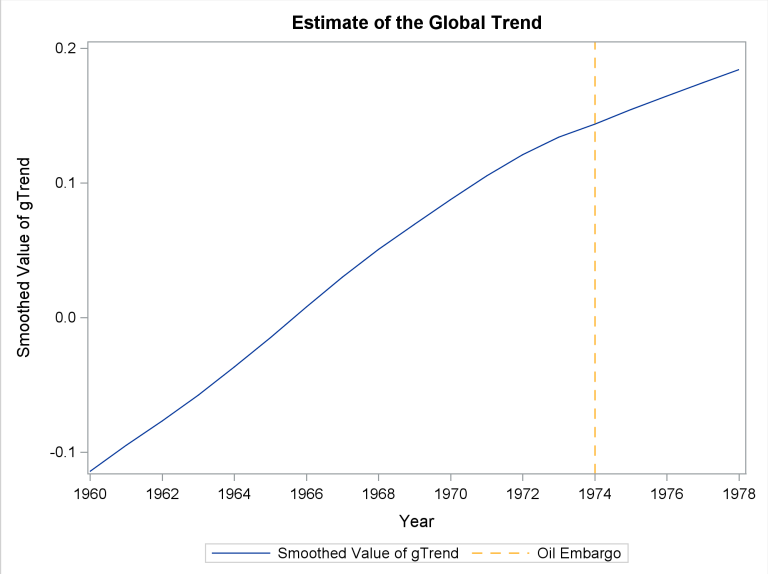
```
proc ssm data=Gas;
   id year interval=year;
   array CountryArray{18} country1-country18;
   do i=1 to 18;
       CountryArray[i] = (cindex=i);
   end;
   trend gTrend(ll) levelvar=0 checkbreak;
   trend rwTrend(rw) cross(matchparm)=(CountryArray) nodiffuse checkbreak;
   irregular wn;
   model lgaspcar=lincomep lrpmg lcarpcap gTrend rwTrend wn;
   output out=gasFor maxshock;
run;
```

The chi-square statistics plot of the change-point diagnostics shown in Output 2.1 indicates two likely change-point years: 1961 and 1974. Based on the summary table of change points for **gTrend**, which corresponds to the global trend ($\mu_t$), and **rwTrend**, which corresponds to the contribution from the country-specific deviation curves ($\eta_{it}$), you can see that the 1961 change can be attributed to Greece (CINDEX=7), and the 1974 change can be attributed to the first element of the state that corresponds to the global trend $\mu_t$. These summary tables are not shown. Output 2.2 shows the estimate of the global trend, which seems to indicate a slight change in the trend around 1973–1974.

**Output 2.1** Maximal State Shock Chi-Square Statistics Plot




**Output 2.2** Estimate of the Global Trend

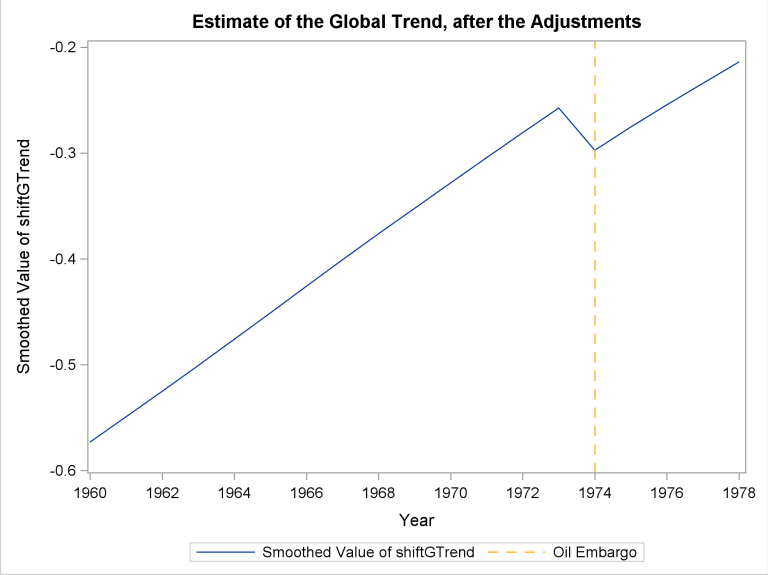


The following statements show you how to adjust the baseline model to account for these changes:

```
proc ssm data=Gas;
   id year interval=year;
   array CountryArray{18}  country1-country18;
   do i=1 to 18;
       CountryArray[i] = (cindex=i);
   end;
   ao7 = (year(year)=1961 & cindex=7);
   shift74 = (year(year) = 1974);
   zero=0;
   state irw(1) type=ll(slopecov(d)) w(g)=(shift74 zero);
   comp shiftGTrend = irw[1];
   trend rwTrend(rw) cross(matchparm)=(CountryArray) nodiffuse checkbreak;
   irregular wn;
   model lgaspcar=ao7 lincomep lrpmg lcarpcap shiftGTrend rwTrend wn;
   output out=gasFor maxshock;
run;
```

Output 2.3 shows the estimate of the common trend, after the adjustment for the oil embargo. This estimate suggests that the oil embargo did affect gasoline consumption per car but that the rate of increase in consumption per car—the slope of the common trend curve—remained constant before and after the embargo. On the other hand, the estimate of the common trend curve without the adjustment for the oil embargo, shown in Output 2.2, appears to show that the rate of increase in consumption per car slows down a little around 1973. That is, accounting for change due to the oil embargo qualitatively changes the conclusions of the data analysis.

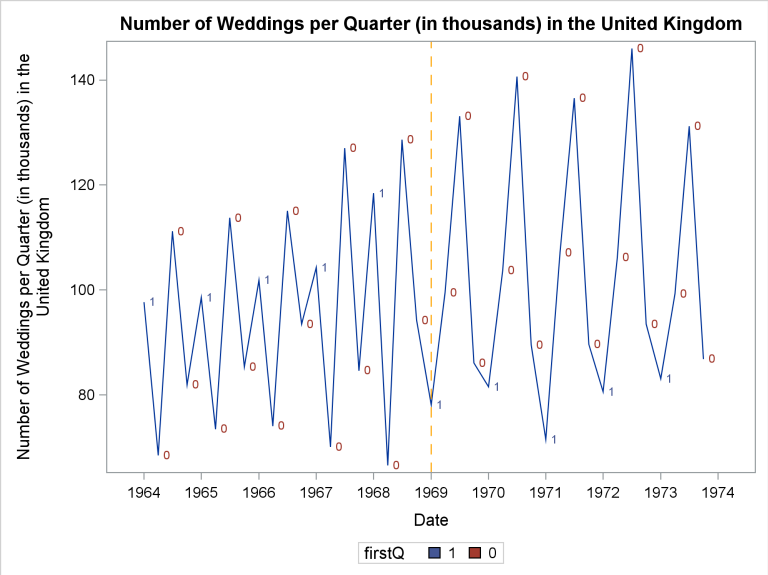
**Output 2.3** Estimate of the Adjusted Global Trend




### Example 3: Detecting Change in a Seasonal Pattern

This illustration is based on an example described in Pelagatti (2015, chap. 4, example 4.5). Until the end of 1968 in the United Kingdom, it was financially convenient to get married in the first quarter of the year. Starting in 1969, a new law, which was announced in 1968, made the taxation scheme neutral to the choice of wedding date. Output 3.1 shows a section of the data (which includes data a few years before 1969 and a few years after). You can clearly see that before this law, the number of weddings is higher in the first quarter (which is denoted by 1; other quarters are denoted by 0).

**Output 3.1** Number of Weddings per Quarter in UK




As a baseline model, the following model appears to be reasonable,

$$y_t = \mu_t + \psi_t + \epsilon_t$$

where $y_t$ = weddings, $\mu_t$ = random walk trend, $\psi_t$ = trigonometric seasonal, and $\epsilon_t$ = white noise. The following statements fit this model and search for breaks in the seasonal pattern:

```
proc ssm data=ukw;
   id date interval=quarter;
   state Season(1) type=season(length=4) cov(g) checkbreak(overall);
   comp s1 = season[1];
   trend rw(rw);
   irregular wn;
   model weddings = rw s1 wn;
run;
```

The table in Output 3.2 shows the likely break points for the seasonal pattern. As expected, they are in the latter part of 1968 and at the start of 1969.

**Output 3.2** Break Locations for the Seasonal Pattern

**Overall Break Summary for Season**

| ID | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| **1968:4** | 64.83 | 3 | <.0001 |
| **1968:3** | 64.78 | 3 | <.0001 |
| **1969:1** | 63.33 | 3 | <.0001 |

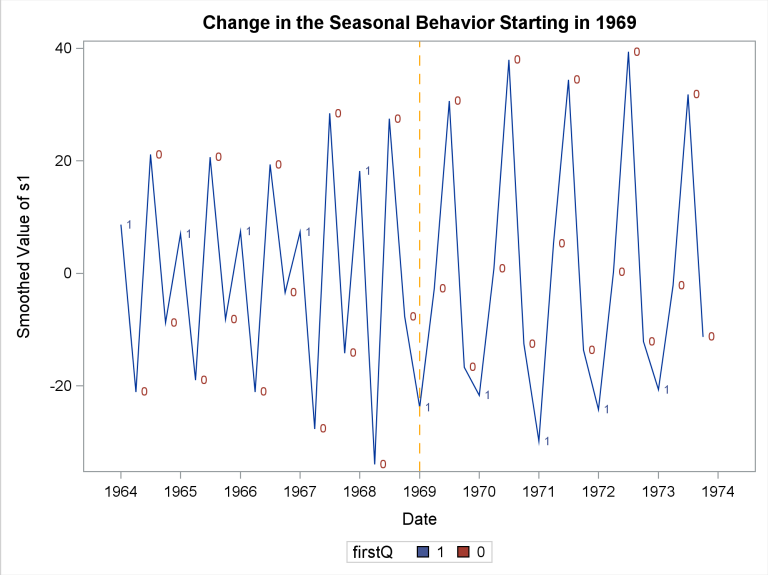
The following statements account for this change in the seasonal pattern by increasing the disturbance variance in the state transition equation of the seasonal factors during 1969. This has an effect of down-weighting the seasonal factors before 1969.

```
proc ssm data=ukw;
   id date interval=quarter;
   parms v1 v2 / lower=1.e-8;
   year69 = (year(date) = 1969);
   var = v1 + v2*year69;
```

```
   state Season(1) type=season(length=4) cov(g)=(var);
   comp s1 = season[1];
   trend rw(rw);
   irregular wn;
   model weddings = rw s1 wn;
   eval full = rw + s1;
   output out=weddingFor press;
run;
```

Output 3.3 shows the estimated seasonal pattern, which correctly reflects the change introduced in 1969.

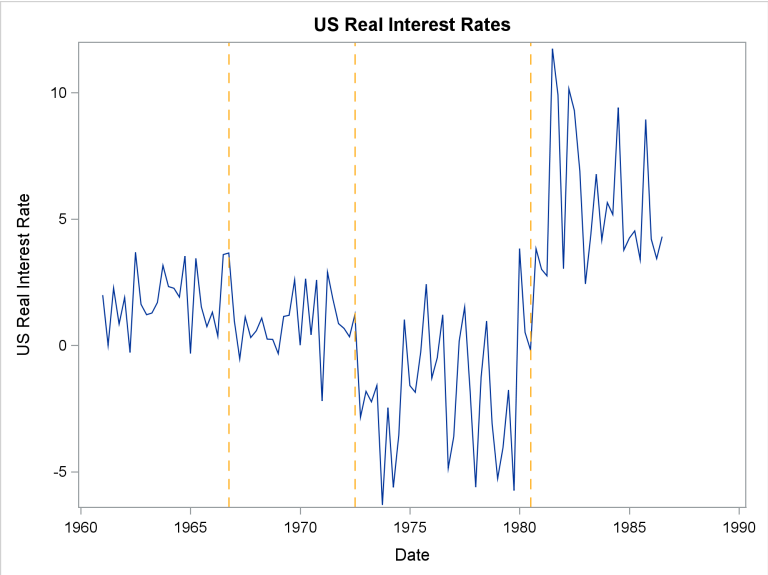
**Output 3.3** Estimated Seasonal Pattern




### Example 4: Detecting Change in the Mean When the Process Variance Is Nonconstant

The US real interest rate data set, **Interest**, that is used in this example is discussed in Garcia and Perron (1996) and Bai and Perron (2003). The plot of the quarterly data, which range from the first quarter of 1961 (1961:1) to the third quarter of 1986 (1986:3), is shown in Output 4.1.

**Output 4.1** Different Regimes of Real Interest Rate



Three change points, 1966:4, 1972:3, and 1980:3, that are detected by Bai and Perron (2003) are shown in the plot. They do say that the first change point, 1966:4, is not very reliable. In an earlier work that used slightly different assumptions about the data generation process, Garcia and Perron (1996) detected only the last two change points. In this example, this change-point detection problem is analyzed by assuming the baseline model

$$y_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim \mathrm{N}(0, \sigma_t^2)$$

where $\mu_t$ is a random walk trend and $\epsilon_t$ is white noise with the time-varying variance $\sigma_t^2$. This baseline model is reasonable, because the visual inspection of the series plot indicates that the interest rates within different regimes hover around a constant level and the variation around the mean level changes noticeably between the regimes. Of course, because the change points are unknown in advance, the form of the time-varying variance $\sigma_t^2$ is not clear. The following parameterization of $\sigma_t^2$, which does not assume any knowledge of the change points, is flexible:

$$\sigma_t^2 = \exp(\sum_{i=1}^{7} \nu_i \mathrm{SplineBasis}_i(t))$$

Here $\nu_i, i = 1, 2, \ldots, 7$, are unknown parameters and $\mathrm{SplineBasis}_i(t)$ are the full set of cubic spline basis functions (B-spline) with four evenly spaced internal knots within the observed time span—essentially, four equally spaced points between 1961:1 and 1986:3. Note that the number of basis functions in the full set (7) is the sum of the number of internal knots (4) and the degree of the polynomial (3). The following statements create a data set, **Spline**, that contains the desired spline basis functions (**c1**–**c7**) that are created by using the BSPLINE function in the IML procedure:

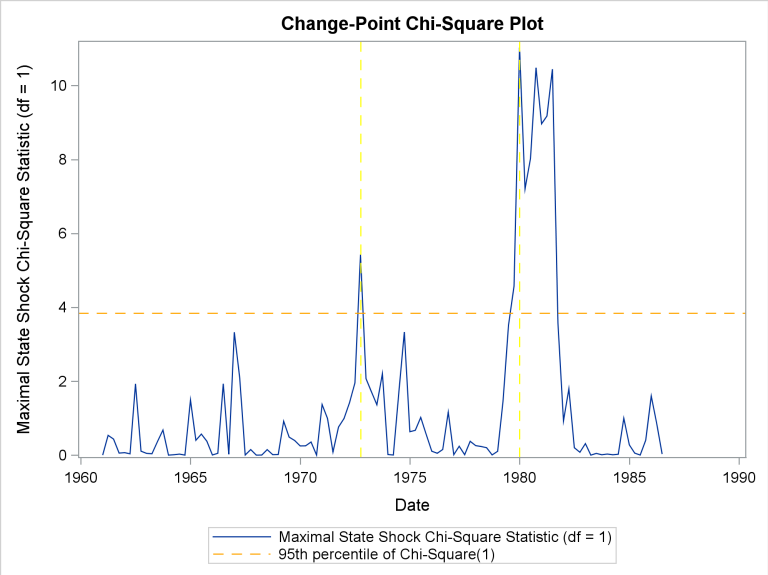
```
proc iml;
    use interest;
    read all var {date} into x;
    bsp = bspline(x, 2, ., 4);
    create spline var{c1 c2 c3 c4 c5 c6 c7};
    append from bsp;
 quit;
 data interest;
    merge interest spline;
 run;
```

The newly created basis function variables, **c1**–**c7**, are merged into the original data set, **Interest**. The following statements show you how to fit the baseline model:

```
proc ssm data=interest opt(tech=dbldog);
   id date interval=quarter;
   parms v1-v7;
   lambda = exp(v1*c1 + v2*c2 + v3*c3 + v4*c4
        + v5*c5 + v6*c6 + v7*c7);
   trend rw(rw) checkbreak print=breakdetail;
   irregular wn variance=lambda;
   model y=rw wn;
   output out=intFor maxshock  pdv;
run;
```

The change-point diagnostic plot shown in Output 4.2 indicates two change points: 1972:3 and 1980:1 (assuming a 95% level of significance). The few quarters around 1980:1 that show higher significance can be considered part of the same change-point phenomenon.

**Output 4.2** Change-Point Chi-Square Plot

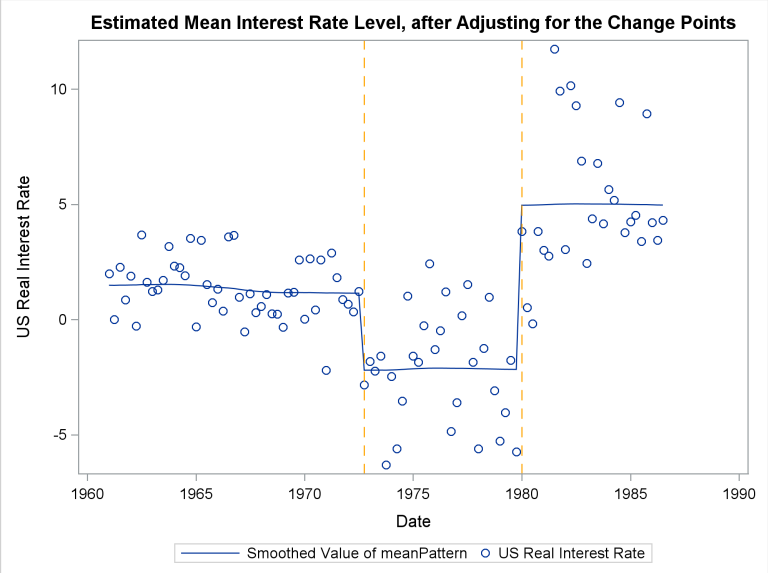



The following statements show you how to adjust the baseline model to account for these change points:

```
proc ssm data=interest opt(tech=dbldog);
   id date interval=quarter;
   reg1 = (date < '01oct1972'd);
   reg2 = date >= '01oct1972'd & date < '01jan1980'd;
   parms v1-v7;
   lambda = exp(v1*c1 + v2*c2 + v3*c3 + v4*c4
        + v5*c5 + v6*c6 + v7*c7);
   trend rw(rw) checkbreak;
   irregular wn variance=lambda;
   model y=reg1 reg2 rw wn;
   eval meanPattern = rw + reg1 + reg2;
   output out=intFor maxshock  pdv;
run;
```

The estimated mean interest rate function after adjusting for the change points is shown in Output 4.3.

**Output 4.3** Estimated Mean Interest Rate



## CONCLUSION

You can broadly divide the change-point analysis problem into three steps: (1) determine or model the normal behavior of the data generation process, (2) determine the change points and the nature of the change at these change points, and (3) adjust the data generation model to account for the discovered change points. The SSM procedure gives you an extremely flexible framework for each of these steps in change-point analysis.

## ACKNOWLEDGMENTS

The author is grateful to Ed Huddleston and Tim Arnold from the Advanced Analytics Division at SAS Institute for their valuable assistance in the preparation of this paper.

## CONTACT INFORMATION

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

Rajesh Selukar
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
rajesh.selukar@sas.com



## REFERENCES

Bai, J., and Perron, P. (2003). "Computation and Analysis of Multiple Structural Change Models." *Journal of Applied Econometrics* 18:1–22. http://dx.doi.org/10.1002/jae.659.

Baltagi, B. H. (2013). *Econometric Analysis of Panel Data*. 5th ed. Chichester, UK: John Wiley & Sons.

Das, R., Banerjee, M., Nan, B., and Zheng, H. (2016). "Fast Estimation of Regression Parameters in a Broken-Stick Model for Longitudinal Data." *Journal of the American Statistical Association* 111:1132–1143.

De Jong, P., and Penzer, J. (1998). "Diagnosing Shocks in Time Series." *Journal of the American Statistical Association* 93:796–806.

Garcia, R., and Perron, P. (1996). "An Analysis of the Real Interest Rate under Regime Shifts." *Review of Economics and Statistics* 78:111–125.

Pelagatti, M. M. (2015). *Time Series Modelling with Unobserved Components*. Boca Raton, FL: CRC Press.

SAS Institute Inc. (2016). *SAS/ETS 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

Selukar, R. S. (2015). "Functional Modeling of Longitudinal Data with the SSM Procedure." In *Proceedings of the SAS Global Forum 2015 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings15/SAS1580-2015.pdf.