Detecting and Adjusting Structural Breaks in Time Series and Panel Data

Rajesh Selukar

SAS Institute Inc., Cary, NC
rajesh.selukar@sas.com

June 12, 2017

Outline

Three sessions of about 50 minutes each:

- Session 1
  - Example snippets
  - Change-point analysis based on state space modeling
- Session 2 and most of session 3
  - Illustrative examples
- Last few minutes of session 3
  - Brief discussion of change-point analysis by other methods

State space modeling in this workshop is done by using PROC SSM, a procedure in SAS/ETS® software

You Do Change-Point Detection and Adjustment So You Can:

Answer questions like:

- How effective was a recent advertising campaign?
- What is the impact of an EPA policy on water quality in different rivers?

Improve forecast accuracy

Properly estimate trends, business cycles, seasonal effects and so on

Monitor industrial processes, environmental variables, etc (only briefly discussed in this workshop)

Change-Point Detection and Adjustment for Sequential Data

Change-point detection and adjustment can be a difficult problem

Change in the normal behavior of the data can be of different types. For example, the changes can occur in

- the trend pattern (shift in the level, slope, or something else)
- in the periodic behavior
- the relationship with the predictors (regression coefficients, transfer function relationships)
- the volatility pattern
- the relationship between response variables (multivariate modeling)
- ...

The Change-Point Analysis Methodology in this Workshop

Based on linear state space modeling

Applicable to a variety of sequential data types (continuous response)

- univariate and multivariate time series
- panel of time series
- longitudinal data

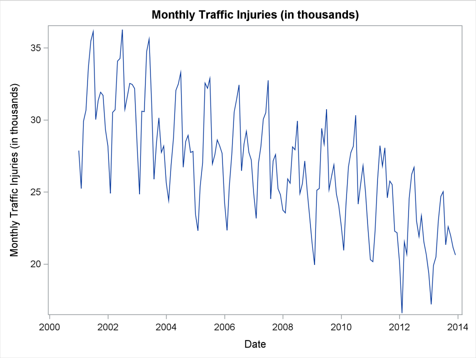
Can detect and adjust a variety of change types. For example, changes in

- the trend pattern
- the seasonal or other periodic patterns
- the regression and transfer function relationships
- ...

Not particularly suitable for detecting change in volatility

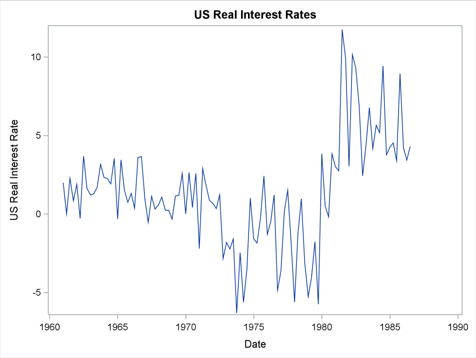
Notice Any Breaks in the Normal Behavior of These Time Series?

Monthly Numbers of Traffic Accidents in Italy



Response = Trend + Season + Error

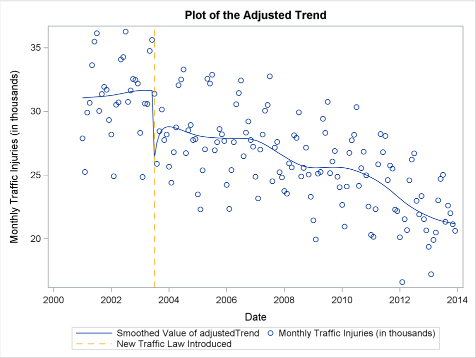
Quarterly Real Interest Rates in US



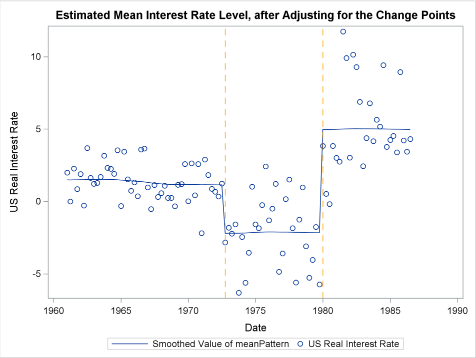
Response = Trend + Error

Results of Modeling and Change-Point Analysis

Trend after Adjusting for the July 2003 Change

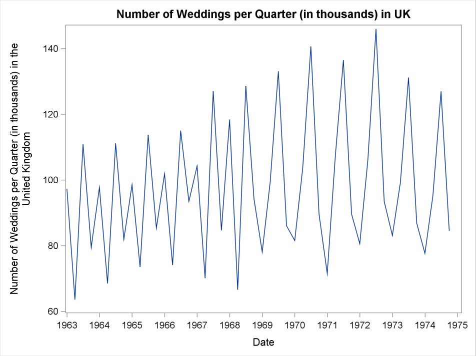


Adjusted Interest Rate Level



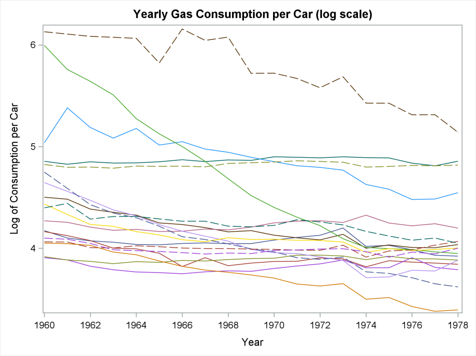
Notice Any Changes in the Normal Behavior?

Number of Weddings in UK (Quarterly)



Response = Trend + Season + Error

Gas Consumption per Car in 18 OECD Countries (Yearly)



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

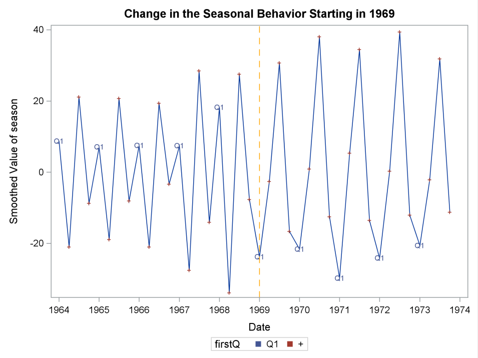
μ_t is global trend and η_{it} is correction for ith country

Results of Modeling and Change-Point Analysis

A change in the seasonal pattern detected in 1969
No major changes detected in the trend

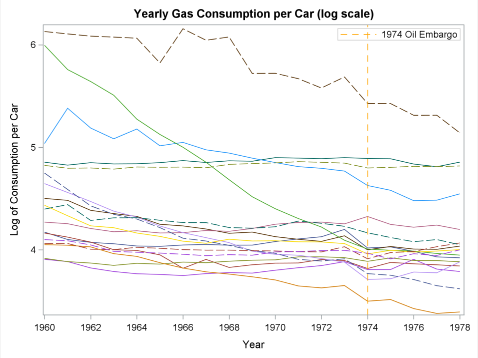
A shock to the global trend (μ_t) detected in 1974
A shock also detected for Greece (η_{7t}) in 1961

Before 1969, First-Quarter Weddings Were Beneficial Taxwise



A tax law change in 1969 made the wedding date tax neutral

Gas Consumption per Car in the 18 OECD Countries



The 1974 oil embargo temporarily affected gas consumption

SSM: A Flexible Framework for Modeling Sequential Data

$y_t = \mathbf{X}_t\, \beta + \mathbf{Z}_t\, \boldsymbol{\alpha}_t + \epsilon_t$	$\epsilon_t \sim \mathrm{N}(0, \sigma^2)$	Observation equation
$\boldsymbol{\alpha}_t = \mathbf{T}\, \boldsymbol{\alpha}_{\mathbf{t-1}} + \mathbf{W}_\mathbf{t}\, \gamma + \boldsymbol{\eta}_\mathbf{t}$	$\boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{Q})$	State transition equation
$\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}$	appropriate	Initial condition

State space model (SSM) generalizes the standard regression model

- β time-invariant regression coefficients (design matrix $\mathbf{X}_t$)
- $\boldsymbol{\alpha}_t$ (states) time-varying regression coefficients (design matrix $\mathbf{Z}_t$)
- γ "state" regression coefficients (design matrix $\mathbf{W}_t$)

The state vector $\boldsymbol{\alpha}_t$ contains time-varying regression effects, such as time-varying intercept, time-varying seasonal indices, and so on

Most common time series models, such as ARIMAX, UCM, and panel data models, can be formulated as SSMs

When $\{y_t\}$ Are Modeled As a SSM, What Is Usually Computed?

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

Usually, the design matrices $\mathbf{X}_t$, $\mathbf{Z}_t$, and $\mathbf{W}_t$ are part of the data (known). If $\mathbf{T}, \mathbf{Q}$, and σ^2 are not fully known, their unknown elements are estimated (usually by maximum likelihood). Assuming all these matrices are known, the famous Kalman filter and smoother algorithm provides

- Estimates of the regression coefficients: $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\boldsymbol{\alpha}}_t$ for all t
- Estimates of the disturbances: $\hat{\boldsymbol{\eta}}_t$ and $\hat{\epsilon}_t$ for all t
- Forecasts and interpolations of $y_t, \forall t$
- The likelihood of the data, $\{y_t, 1 \leq t \leq T\}$, and various goodness of fit measures

You Can Also Detect Breaks in the Normal Behavior of $\{y_t\}$

$$y_t = \mathbf{X}_t\,\beta + \mathbf{Z}_t\,\boldsymbol{\alpha}_t + \epsilon_t \qquad \text{Observation equation}$$

$$\boldsymbol{\alpha}_t = \mathbf{T}\,\boldsymbol{\alpha}_{\mathbf{t-1}} + \mathbf{W}_\mathbf{t}\,\gamma + \boldsymbol{\eta}_\mathbf{t} \qquad \text{State transition equation}$$

De Jong, P., and Penzer, J. (1998). *Diagnosing Shocks in Time Series.* JASA

Assuming that the time series follows an SSM, De Jong and Penzer describe an efficient algorithm to detect perturbations in

- the observation equation—additive outliers
- the state transition equation—structural breaks

This algorithm is the main tool behind the change-point analysis methodology in this workshop

Detecting Unusual Response Values (Additive Outliers)

At each $t, 1 \leq t \leq T$, the De Jong-Penzer algorithm provides the *delete-one-residual* (and its standard error):

$$AO_t = y_t - E(y_t|\mathbf{Y}^t)$$

AO_t is 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 $\mathbf{Y}^t$.

Time points with "large" $AO_t/stdError(AO_t)$ are additive outliers.

The SSM procedure reports significant AOs by default.

In the remainder of the workshop, AOs are discussed only briefly.

Detecting Unexpected Changes in the State Transition

$$\boldsymbol{\alpha}_t = \mathbf{T}\,\boldsymbol{\alpha}_{\mathbf{t-1}} + \mathbf{W_t}\,\boldsymbol{\gamma} + \boldsymbol{\eta}_{\mathbf{t}}$$

The state transition changed at $t = t_0$ if, for some vector $\boldsymbol{\delta}$,

$$\boldsymbol{\alpha}_t = \mathbf{T}\,\boldsymbol{\alpha}_{\mathbf{t-1}} + \mathbf{W_t}\,\boldsymbol{\gamma} + \mathbf{I}_{\mathbf{t=t_0}}\boldsymbol{\delta} + \boldsymbol{\eta}_{\mathbf{t}}$$

where $\mathbf{I}_{t=t_0}$ is the identity matrix if $t = t_0$, and is zero otherwise.

At each $t, 1 \leq t \leq T$, the De Jong-Penzer algorithm provides:

- $\hat{\boldsymbol{\delta}}$, an estimate of $\boldsymbol{\delta}$, and its standard error
- chi-square statistic, $\boldsymbol{\tau}_t$, to test the hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}$

Time points with significant $\boldsymbol{\tau}_t$ are change-point candidates

Simple Example: Detecting Change in a Regression Coefficient

Simulated $y_t, z_t, 1 \leq t \leq 60$, that satisfy the relation

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

The regression coefficient α_t changes over time:

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

Suppose the standard regression $(\alpha_t \equiv \alpha)$ is taken as initial model:

$$y_t = \mu + z_t\, \alpha + \epsilon_t, \;\; \epsilon_t \sim \mathrm{N}(0, \sigma^2)$$

Question: Can we detect the changes in α_t?

PROC SSM Specification of the Initial Model

Regression in the SSM form (with the coefficient of z_t treated as state):

$$y_t = \mu + z_t \, \alpha_t + \epsilon_t; \quad \alpha_t = \alpha_{t-1}; \quad \alpha_1 = \alpha \;\; \text{(unknown)}$$

The *regBreak* data set contains three variables: t, y, and z.

```
proc ssm data=regBreak plot=maxshock;
   id t;
   intercept = 1.0;
   state regCoeff(1) T(I) A1(1) checkbreak;
   component zTerm = regCoeff * (z);
   irregular wn;
   model y = intercept zTerm wn;
run;
```

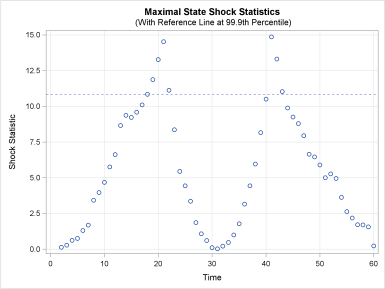
Detection of Change in the Regression Coefficient

$$y_t = \mu + z_t\, \alpha_t + \epsilon_t; \quad \alpha_t = \alpha_{t-1}; \quad \alpha_1 = \alpha$$

Fitting this model you get the usual regression output:

$$\hat{\mu} = 6.95, \quad \hat{\alpha} = 2.7, \quad \hat{\sigma}^2 = 8.58$$

Additionally, you get Chi-square statistics plot to test $H_0 : \delta = 0$ in $\alpha_t = \alpha_{t-1} + \delta$:



Strong evidence of shifts at $t = 21$ and at $t = 41$.

PROC SSM Specification for the Revised Model

$$y_t = \mu + z_t\,\alpha_t + \epsilon_t; \quad \alpha_t = \alpha_{t-1} + I_{(t=21)}\,\gamma_1 + I_{(t=41)}\,\gamma_2; \quad \alpha_1 = \alpha$$

```
proc ssm data=regBreak;
   id t;
   intercept = 1.0;
   break1 = (t = 21);
   break2 = (t = 41);
   state regCoeff(1) T(I) W(g)=(break1 break2) A1(1);
   component zTerm =  regCoeff * (z);
   irregular wn;
   model y = intercept zTerm wn;
run;
```

Fit of the Revised Model

$$y_t = \mu + z_t\ \alpha_t + \epsilon_t; \quad \alpha_t = \alpha_{t-1} + I_{(t=21)}\ \gamma_1 + I_{(t=41)}\ \gamma_2; \quad \alpha_1 = \alpha$$

Fit of this model gives reasonable estimates:

Parameter	μ	α	γ_1	γ_2	σ^2
Estimate	9.35	1.25	1.92	−1.95	0.23
True Value	10	1.00	2.00	−2.00	0.25

Change-Point Detection in a General SSM: What Changed?

$$y_t = \mathbf{X}_t\, \beta + \mathbf{Z^1}_t\, \boldsymbol{\alpha}_t^1 + \mathbf{Z^2}_t\, \boldsymbol{\alpha}_t^2 + \ldots + \mathbf{Z^k}_t\, \boldsymbol{\alpha}_t^k + \epsilon_t$$
$$\boldsymbol{\alpha}_t = \mathbf{T}\, \boldsymbol{\alpha}_{\mathbf{t-1}} + \mathbf{W_t}\, \gamma + \boldsymbol{\eta}_\mathbf{t}$$

Three types of chi-square break statistics plots could be examined:

Check for change in the full state vector ($\boldsymbol{\alpha}_t$)

- One or more components changed at a point t_0

Check for change in a particular state block ($\boldsymbol{\alpha}_t^j$), which could correspond to trend, or seasonal component, or something else

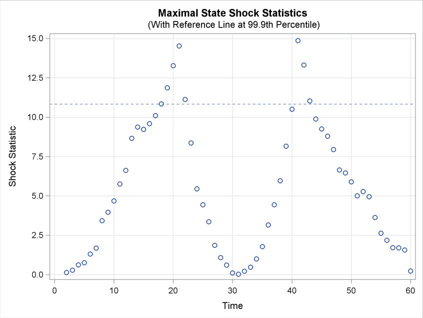
- Change happened in j-th block at a point t_0

Check for change in a particular element of a state block ($\boldsymbol{\alpha}_{it}^j$)

- A particular element of a particular component changed at a point t_0

Break Points Correspond to Significant Peaks

Recall the chi-square break statistics plot in the "Detecting Change in a Regression Coefficient" example. The time points around the peak usually correspond to the same break phenomenon and can be ignored (in the SSM procedure you can output these significant peaks by using the *BREAKPEAK* option).



Different Ways to Revise The Model To Account For Breaks

$$y_t = \mathbf{X}_t\,\beta + \mathbf{Z^1}_t\,\alpha_t^1 + \mathbf{Z^2}_t\,\alpha_t^2 + \ldots + \mathbf{Z^k}_t\,\alpha_t^k + \epsilon_t$$
$$\alpha_t = \mathbf{T}\,\alpha_{\mathbf{t-1}} + \mathbf{W_t}\,\gamma + \eta_{\mathbf{t}} \qquad \eta_{\mathbf{t}} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q_t})$$

Suppose a break is detected in one or more elements of α_t^j at t_0. You can revise the model to account for this break in a few different ways:

- Add a suitable dummy regressor in the observation equation
- Add a suitable dummy regressor in the state equation
- Suitably down-weigh $\alpha_{t_0-1}^j$ by adjusting the covariance of the disturbance terms, $\mathbf{Q_t}$, for $t \geq t_0$
- Make more involved model changes

Change-Point Detection and Adjustment: An Iterative Scheme

1. Start with a reasonable baseline model for the data.
2. Fit and diagnose the baseline model (including the detection of change points).
3. If all good STOP, else:
4. Revise the baseline model, which becomes the new baseline model. Go to step 2.

This is a fairly ad hoc process. Many things are subjective:

- How to choose the baseline model
- What is the right significance level and how to guard against multiple testing
- Which breaks to adjust for and which ones to ignore
- How to adjust the model for a particular break
- ...

In each application, this process must be properly tuned

PROC SSM Makes SSM-Based Change-Point Analysis Easy

- Designed for handling very general SSMs (more general than the ones shown so far)

- Works with a variety of sequential data types:
  - Univariate and multivariate time series
  - Panels of time series
  - Longitudinal data

- Provides all the diagnostic tools needed for an SSM-based change-point analysis

Where to Find Additional Info

- Selukar, R. S. (2017). *Detecting and Adjusting Structural Breaks in Time Series and Panel Data Using the SSM Procedure.* In Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc.
  `http://support.sas.com/resources/papers/proceedings17/SAS0456-2017.pdf`

- Selukar, R. S. (2016). *"Time Series Modeling with Unobserved Components."* In Proceedings of the International Symposium on Forecasting, Santander, Spain.
  `https://forecasters.org/wp-content/uploads/gravity_forms/7-621289a708af3e7af65a7cd487aee6eb/2016/07/Selukar_Rajesh_ISF2016.pdf`

- SAS Institute Inc. (2016). *"The SSM Procedure."* In SAS/ETS 14.2 Users Guide. Cary, NC: SAS Institute Inc.
  `http://support.sas.com/documentation/onlinedoc/ets/142/ssm.pdf`

Workshop Illustrations: General Comments

The data in these examples have been previously analyzed by others.

- Largely, the conclusions of the analyses done here agree with these earlier analyses
- The data analysis methodology, however, need not be the same

The examples are chosen with the following considerations:

- They are simple to describe and the change-points and their nature is "known"
- Different data types: univariate and multivariate time series, panel data
- Different types of change points: changes in trend, seasonal pattern, regression relationships
- Different types of adjustments

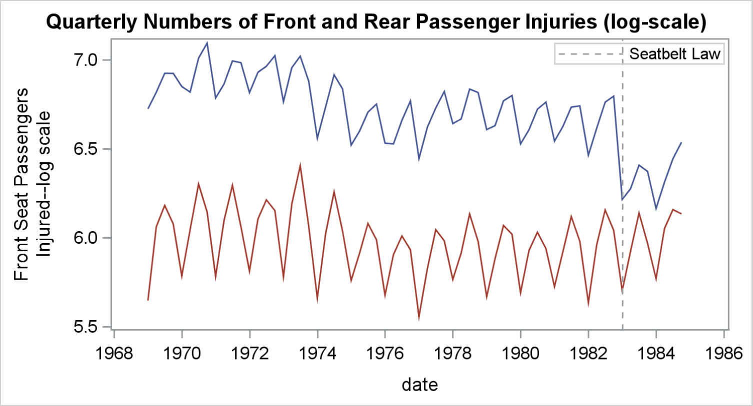
Example1: Break Detection in a Bivariate Time Series

- Based on data that have been analyzed in Durbin and Koopman (2012) and Harvey (1989).
- Quarterly data on two variables, *f_KSI* and *r_KSI*, that represent the front and rear seat passenger injuries (respectively) on the roads in UK (in log scale).
- One of the original reasons for studying these data was to assess the effect on *f_KSI* of the enactment of a seat-belt law in February 1983 that compelled the front seat passengers to wear seat belts.
- Question: How effective was the seat-belt law?

Effectiveness of the 1983 Seat-Belt Law in UK

The *f_KSI* and *r_KSI* series appear to

- move together without discernible upward or downward trend
- be seasonal



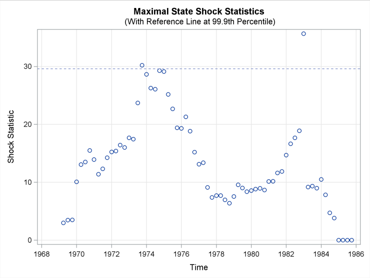
(f_KSI, r_KSI) = RW Trend + Seasonal + Error

Since (f_KSI, r_KSI) appear to move together (co-integrated), the covariance of the disturbance term of this random walk is assumed to be of lower than full rank.

```
proc ssm data=seatBelt breakpeak plots=maxshock;
    id date interval=quarter;
    state level(2) type=RW cov(rank=1) checkbreak;
    component rw1 = level[1];
    component rw2 = level[2];
    state season(2) type=SEASON(length=4);
    component s1 = season[1];
    component s2 = season[2];
    state error(2) type=WN cov(g);
    component wn1 = error[1];
    component wn2 = error[2];
    model f_KSI = rw1 s1 wn1;
    model r_KSI = rw2 s2 wn2;
run;
```

Structural Break Diagnosis

A significant break in the level of *f_KSI*—first element of the state that underlies the bivariate RW trend—is discovered in the first quarter of 1983. An earlier peak around the end of 1973 probably corresponds to the oil-embargo.



Elementwise Break Summary for level

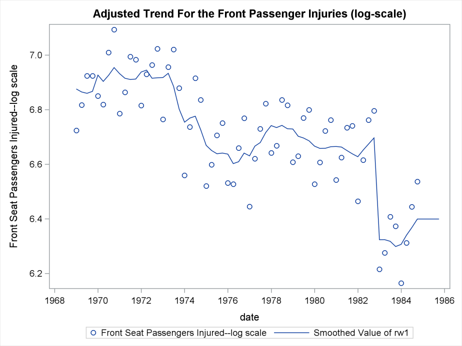
ID	Element Index	Z Value	Pr > |z|
1983:1	1	-5.85	<.0001

(f_KSI, r_KSI) = Shifted RW Trend + Seasonal + Error

```
proc ssm data=seatBelt breakpeak plots=maxshock;
    id date interval=quarter;
    Q1_83_Pulse = (date = '1jan1983'd);
    zero = 0;
    state level(2) type=RW cov(rank=1) W(g)=(Q1_83_Pulse zero);
    component rw1 = level[1];
    component rw2 = level[2];
    state season(2) type=SEASON(length=4);
    component s1 = season[1];
    component s2 = season[2];
    state error(2) type=WN cov(g);
    component wn1 = error[1];
    component wn2 = error[2];
    model f_KSI = rw1 s1 wn1;
    model r_KSI = rw2 s2 wn2;
 run;
```

Revised Model Output

- The left panel shows the plot of the estimated trend for *f_KSI*, which incorporates the 1983 shift.
- Roughly speaking, the table in the right indicates that the seat-belt law appears to have reduced the front passenger injuries by –0.408 (log scale).



Estimate of the State Equation Regression Vector

State	Element Index	Estimate	Standard Error	t Value	Pr > |t|
level	1	-0.408	0.0259	-15.74	<.0001

Example 2: Break Detection in an Industrial Process

- Based on a case study in Box and Jenkins (1976)—data set *seriesJ*.
- The data set *seriesJ* contains sequentially recorded measurements on two variables: x, the input gas rate, and y, the output CO_2.
- For the output CO_2, Box and Jenkins suggest the model

$$y_t = \mu + f_t + \xi_t$$

where μ is the intercept, ξ_t is a zero-mean noise term that follows a second-order autoregressive model (that is, $\xi_t \sim \text{AR}(2)$) and f_t follows a transfer function model

$$f_t \;=\; \delta f_{t-1} + \gamma_1 x_{t-3} + \gamma_2 x_{t-4} + \gamma_3 x_{t-5}$$

$$y_t = \mu + f_t + \xi_t$$

```
proc ssm data=Seriesj(firstobs=6) breakpeak;
   id obsIndex;
   intercept = 1;
   /*--- Transfer function specification ---*/
   parms delta /lower=-0.9999 upper=0.9999;
   state tfstate(1) T(g)=(delta) W(g)=(x3 x4 x5) a1(1)
      checkbreak;
   component tfinput = tfstate[1];
   /*--AR(2) noise specification --*/
   trend ar2(arma(p=2)) checkbreak;
   model y = intercept tfinput ar2 ;
   eval modelCurve = intercept + tfinput;
run;
```

Break Diagnostics for the Transfer Function Model

- Some evidence to suggest that the relationship between the output CO_2 (y) and the input gas rate (x) starts to break down towards the latter part of the sample: observation number 199 onward.

Elementwise Break Summary for tfstate

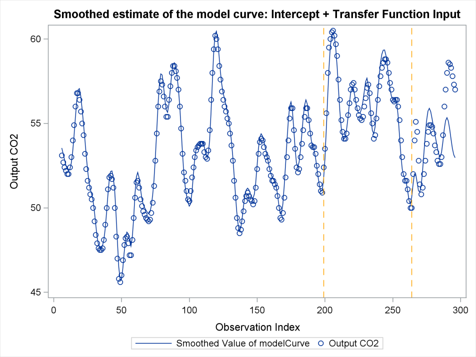
ID	Element Index	Z Value	Pr > |z|
264	1	-5.03	<.0001
199	1	4.62	<.0001
235	1	3.15	0.0017

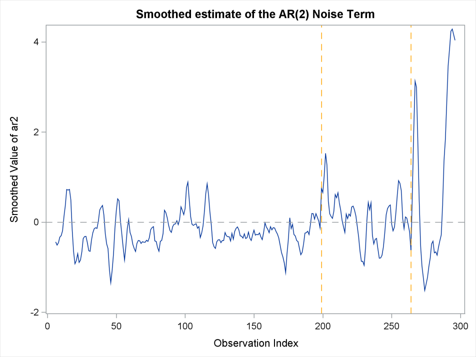
Elementwise Break Summary for ar2

ID	Element Index	Z Value	Pr > |z|
264	2	6.09	<.0001
264	1	-5.82	<.0001
199	1	4.16	<.0001

Component Estimates for the Transfer Function Model

- The left panel shows the estimate of $\mu + f_t$ and the right panel shows the estimate of the AR(2) noise ξ_t. The dashed vertical lines show the breaks at the observation numbers 199 and 264.
- This type of information could be useful in quality control settings. The process could be stopped at the first indication of the change in the relationship between x and y.



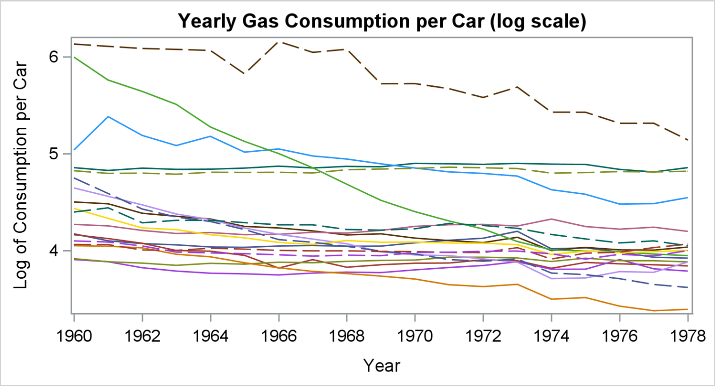


Example 3: Gas Consumption Per Car in the OECD Countries

- These data, a panel of 18 yearly time series, are discussed in Baltagi (2013)
- The variables in the data set, Gas, are
    - year
    - country
    - lgaspcar: log of gasoline consumption per car
    - lincomep: log of per capita income
    - lrpmg: log of real price of gasoline
    - lcarpcap: log of per capita number of cars
    - cIndex (an integer between 1 and 18 that uniquely identifies each country)
- lgaspcar is the response variable and lincomep, lrpmg, and lcarpcap are the predictors
- The goal of this analysis is to see whether the oil embargo of 1973–1974 had any effect on the gasoline consumption per car (lgaspcar).

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

- $\mathbf{X}_{it}\beta$ regression term involving lincomep, lrpmg, and lcarpcap
- μ_t global trend, modeled as an integrated random walk
- η_{it} country-specific deviation from the global trend, modeled as random walk
- ϵ_{it} white noise



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

```
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;
 run;
```

Break Diagnostics for the Panel Data Example

- The left table shows the break summary for the global trend (μ_t) and the right table shows the break summary for the country-specific deviation curves (η_{it}).
- A break for μ_t is detected in 1974 and a break is detected in 1961 for the deviation curve for Greece (7th country). Although not shown here, the 1961 break for Greece also appears as a significant additive outlier.

Elementwise Break Summary for gTrend

ID	Element Index	Z Value	Pr > |z|
1974	1	-4.36	<.0001

Elementwise Break Summary for rwTrend

ID	CROSS= Variable	CROSS= Array Index	Element Index	Z Value	Pr > |z|
1961	country7	7	1	7.33	<.0001

$lagaspcar_{it} = \mathbf{X}_{it}\beta + \mu_t + \eta_{it} + \text{adjustments} + \epsilon_{it}$

```
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);
    component shiftGTrend = irw[1];
    trend rwTrend(rw) cross(matchparm)=(CountryArray) nodiffuse;
    irregular wn;
    model lgaspcar=ao7 lincomep lrpmg lcarpcap shiftGTrend
       rwTrend wn;
    eval fitPattern = ao7 + lincomep + lrpmg + lcarpcap + shiftGTrend
        + rwTrend;
 run;
```

Regression Estimates in the Revised Model

- The left table shows the regression effects in the observation equation:
  - As expected, the coefficients of real price of gasoline and per capita number of cars are negative (–0.193 and –0.610) and the coefficient of per capita income (0.155) is positive.
  - The 1961 shock (0.257) for Greece appears positive and is quite significant.
- The right table shows the state regression effect that corresponds to the 1974 correction to μ_t. It shows a one-time dip of –0.0626 in μ_t.

Regression Parameter Estimates

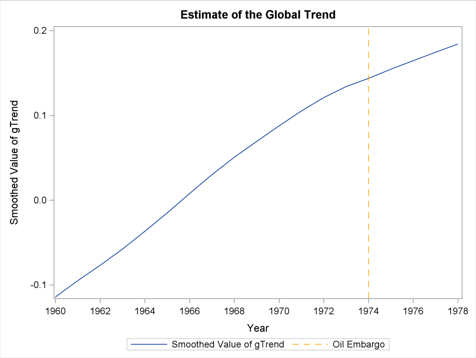
Response Variable	Regression Variable	Estimate	Standard Error	t Value	Pr > |t|
lgaspcar	ao7	0.257	0.0313	8.21	<.0001
lgaspcar	lincomep	0.155	0.0697	2.23	0.0260
lgaspcar	lrpmg	-0.193	0.0288	-6.68	<.0001
lgaspcar	lcarpcap	-0.610	0.0335	-18.21	<.0001

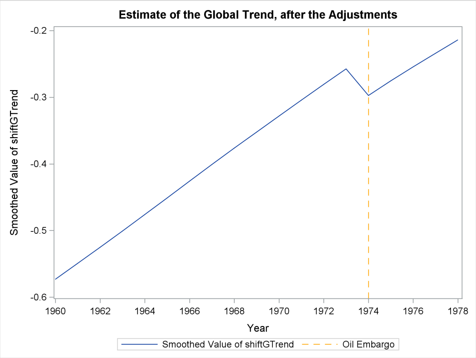
Estimate of the State Equation Regression Vector

State	Element Index	Estimate	Standard Error	t Value	Pr > |t|
irw	1	-0.0626	0.0114	-5.50	<.0001

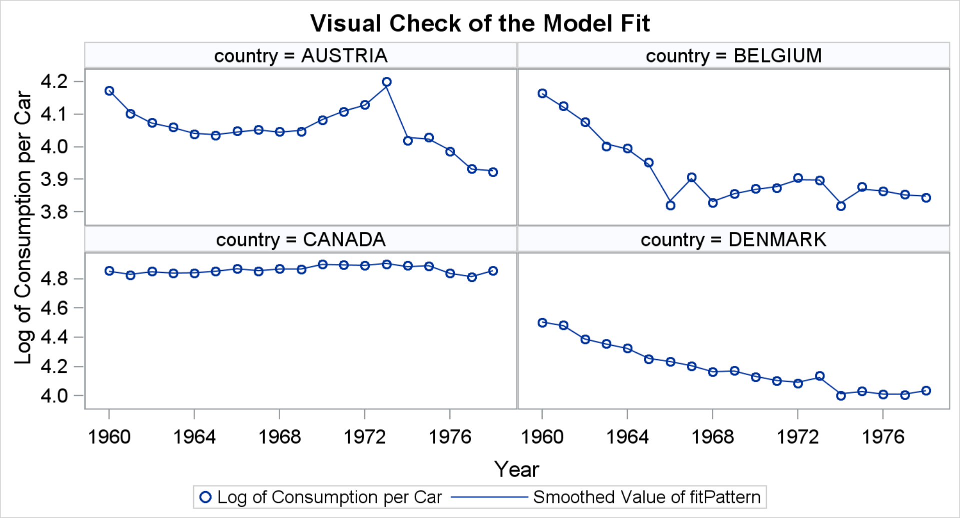
Global Trend Estimate: Before Adjustment and After Adjustment

- The left panel shows the plot of the estimate of μ_t based on the first model and the right panel shows the plot of the estimate of μ_t based on the revised model that adjusted for the breaks.

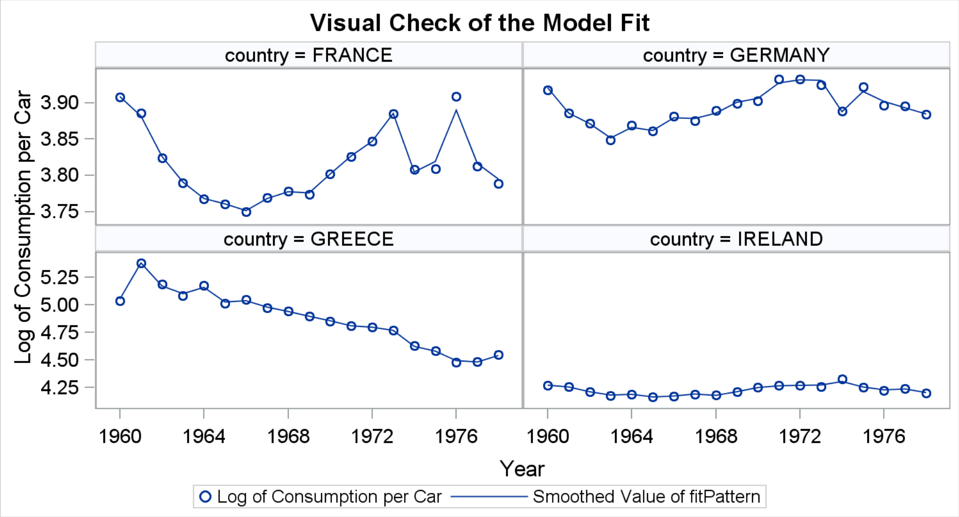




Visual Check of The Model Fit for a Few Countries

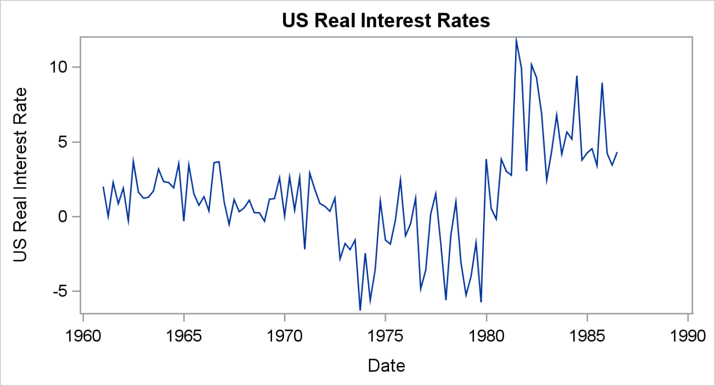


Visual Check of The Model Fit for a Few Countries



Example 4: Modeling of the Real Interest Rate in the US

- Based on an example discussed in Garcia and Perron (1996) and Bai and Perron (2003).
- Quarterly data on the real interest rate in the US.
- The interest rate series seems to hover around its mean level, which abruptly changes a few times during the course of the series. Moreover, the noise variance also seems to change with time.



Model for the Interest Rate Series

Based on the visual inspection of the interest rate series, a reasonable model is:

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \lambda_t)$$

- μ_t is a random walk trend
- ϵ_t is a white noise with time-varying variance λ_t
- Assuming λ_t varies smoothly with time, it could be expressed as an exponential of a cubic spline with four equally spaced knots in the observed time span:

$$\lambda_t = \exp(v1 * c1_t + v2 * c2_t + \cdots + v7 * c7_t)$$

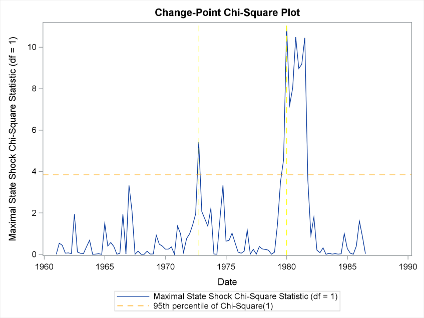
where $v1, v2, \cdots, v7$ are parameters (to be estimated using the data) and $c1_t, c2_t, \cdots, c7_t$ are bspline basis functions for a cubic spline with four internal knots.

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

```
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;
proc ssm data=interest breakpeak;
   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;
   irregular wn variance=lambda;
   model y=rw wn;
   output out=intFor break(alpha=0.05 maxnum=3);
run;
```

Break Diagnostics for the Interest Rate Series

- Two breaks are detected in the random walk trend: First quarter of 1980 and the fourth quarter of 1972.
- Bai and Perron (2003) find some evidence (rather weak) of a third break around the last quarter of 1966.



Elementwise Break Summary for rw

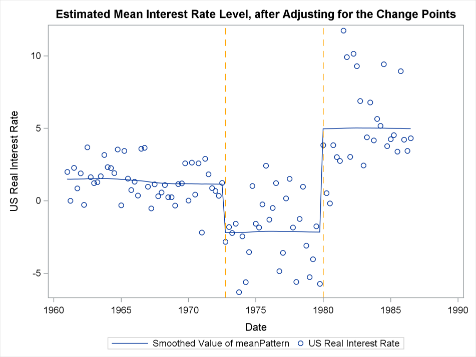
ID	Element Index	Z Value	Pr > |z|
1980:1	1	3.32	0.0009
1972:4	1	-2.33	0.0198

$$y_t = \mu_t + \text{break1} + \text{break2} + \epsilon_t, \qquad \epsilon_t \sim \mathrm{N}(0, \lambda_t)$$

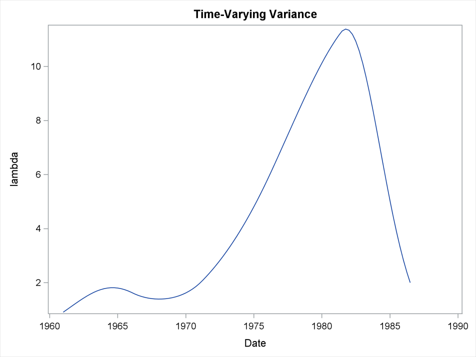
```
proc ssm data=interest;
    id date interval=quarter;
    reg1 = (date < '01oct1972'd);
    reg2 = date >= '01oct1972'd & date < '01jan1980'd;
    parms v1-v7;
    lambda = exp(v1*c1 + v2*c2 + v3*c3 + v4*c4
         + v5*c5 + v6*c6 + v7*c7);
    trend rw(rw) checkbreak;
    irregular wn variance=lambda;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

Estimate of the Adjusted Trend and the Time-Varying Error Variance

Estimate of the Adjusted Trend μ_t

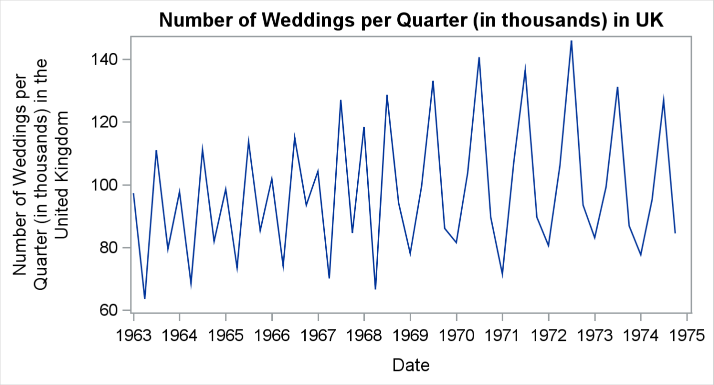


$\hat{\lambda}_t = \exp(\hat{v1} * c1_t + \hat{v2} * c2_t + \cdots + \hat{v7} * c7_t)$



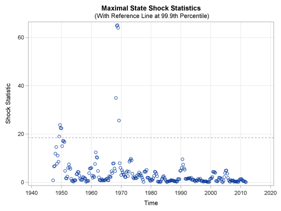
Example 5: Number of Weddings Per Quarter in the UK

- Based on Example 4.5 from Pelagatti (2016)
- Quarterly data on number of weddings (in thousands) in the UK
- A new tax rule was enacted (to come into effect at the start of 1969) that made the wedding date tax neutral. Prior to this rule, getting married in the first quarter was beneficial from the tax perspective.



Weddings = RW Trend + Seasonal + Error

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



Overall Break Summary for SeasonState

ID	Chi-Square	DF	Pr > ChiSq
1968:4	64.83	3	<.0001
1949:3	23.73	3	<.0001
1961:3	12.01	3	0.0074

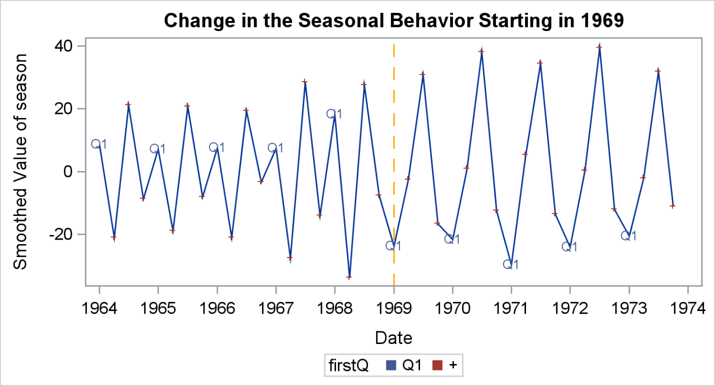
Weddings = RW Trend + Adjusted Seasonal + Error

The adjustment of the seasonal component to account for the 1969 break involves down-weighing the seasonal vector in the last quarter of 1968. This is done by temporarily increasing the state disturbance variance in all quarters of 1969 for the seasonal state equation.

```
proc ssm data=ukw;
    id date interval=quarter;
    parms v1 v2 / lower=1.e-8;
    year69 = (year(date) = 1969);
    var = v1 + v2*year69;
    state SeasonState(1) type=season(length=4) cov(g)=(var);
    component season = SeasonState[1];
    trend rw(rw);
    irregular wn;
    model weddings = rw season wn;
 run;
```

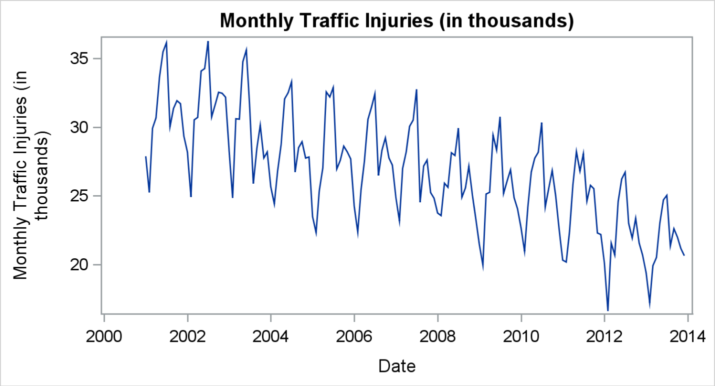
The Adjusted Seasonal Component

- Before 1969 it was tax-wise beneficial to get married in the first quarter. Therefore, the number of first quarter weddings are higher (not the highest) before 1969.
- After 1968, the wedding date had no tax significance and the first quarter wedding numbers are no longer higher.



Example 6: Modeling Motor Vehicle Injuries in Italy

- Based on Case Study # 1 from Pelagatti (2016)
- Monthly data on number of injuries due to road accidents
- A new traffic monitoring system introduced in July 2003
- Question: How effective is the new monitoring system?

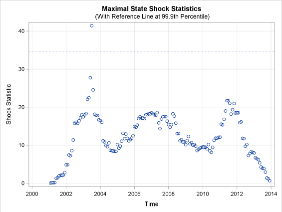


Injuries = IRW Trend + Seasonal + Irregular

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

Elementwise Break Summary for irw

ID	Element Index	Z Value	Pr > |z|
JUL2003	1	-5.55	<.0001
DEC2002	2	-2.84	0.0046

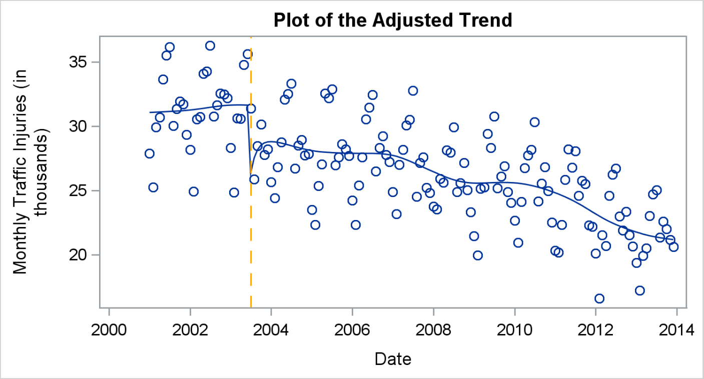


Injuries = IRW + Level Adjustment + Seasonal + Irregular

```
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);
    component tfInput = tfSt[1];
    state seasonState(1) type=season(length=12)
        cov(g) checkbreak(overall);
    component 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;
 run;
```

Estimate of the Adjusted Trend: irw + shift03 + tfInput

- There is a drop in the number of injuries soon after the introduction of the traffic monitoring system in July 2003. Moreover the injury trend seems to point downwards after this change.
- The initial drop in the mean injury level is lost slightly in the subsequent months.



Other Ways of Structural Break Detection and Adjustment

Change point detection for time series is a very large field and it is virtually impossible to summarize it in a few slides. The following list of R packages gives a rough idea about how active this field is:

- *segmented (2015), changepoint (2014), cumSeg (2012), DNAcopy (2008), strucchange (2002), cpm (2013), AutoPARM (2006), bcp (2007), ...*

The following two change-point detection approaches (with many variations) are popular:

- Model-based Break detection methods that use CUSUM and/or CUSUMSQ statistics. CUSUM and CUSUMSQ statistics are computed using the cumulative sum and cumulative sum of squares of "one-step-ahead residuals", respectively.
- Nonparametric/parametric break detection based on segmentation methods.

PROC SSM And CUSUM/CUSUMSQ-Based Change Point Detection

- The CUSUM and CUSUMSQ statistics are computed using the cumulative sum and cumulative sum of squares of one-step-ahead residuals, respectively.
- The SSM procedure provides the one-step-ahead residuals as a part of the model-fitting output.
- With only a little bit of effort, you can use these residuals to do the CUSUM/CUSUMSQ-based change point detection as a post-fit diagnostic step.
- The CUSUM/CUSUMSQ-based change point diagnostics can be complementary to the De Jong-Penzer algorithm-based change point diagnostics (which have been emphasized in this workshop).
- In particular, the CUSUM/CUSUMSQ-based change point diagnostics can be useful to detect changes in the variance of the process.

Summary

- The change point analysis discussed in this workshop treats it as a part of larger data modeling exercise
- The modeling process is based on linear state space models
- A variety of break types and a variety of sequential data types can be handled
- At the conclusion of the modeling process, one has a reasonably adequate understanding of the data generation process

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.
- De Jong, P., and Penzer, J. (1998). Diagnosing Shocks in Time Series. Journal of the American Statistical Association 93:796806.
- Durbin, J., and Koopman, S. J. (2012). Time Series Analysis by State Space Methods. 2nd ed. Oxford: Oxford University Press.
- Garcia, R., and Perron, P. (1996). An Analysis of the Real Interest Rate under Regime Shifts. Review of Economics and Statistics 78:111–125.
- Harvey, A. C. (1989). Forecasting, Structural Time Series Models, and the Kalman Filter. Cambridge: Cambridge University Press.
- Pelagatti, M. M. (2016). *Time Series Modeling with Unobserved Components.* CRC Press.