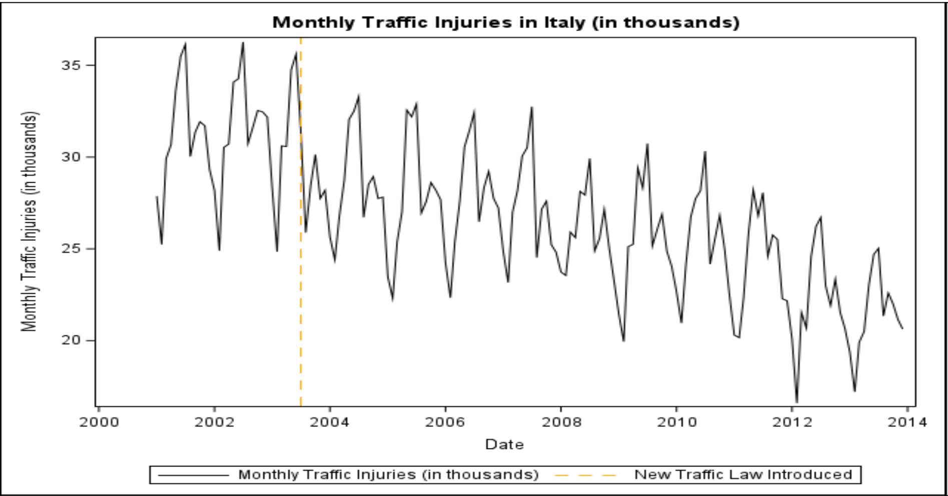
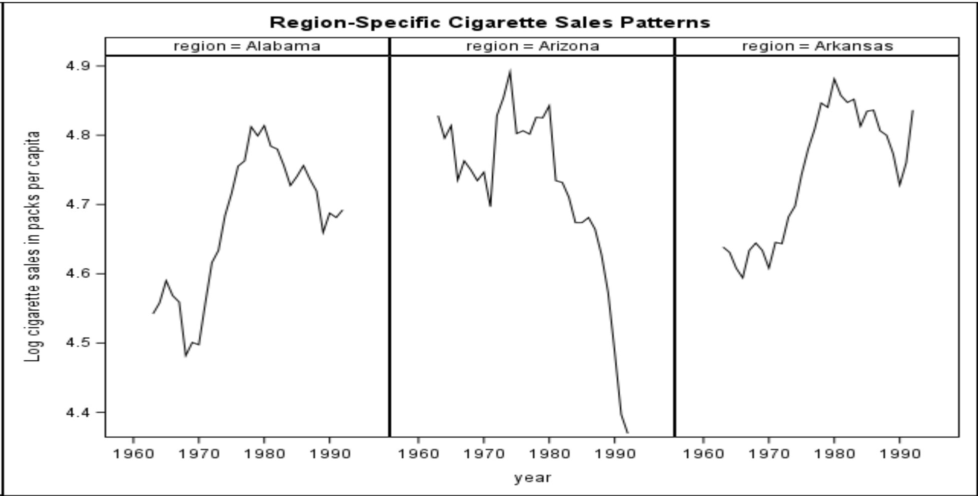
State Space Modeling of Time Series and Longitudinal Data by PROC SSM and PROC UCM

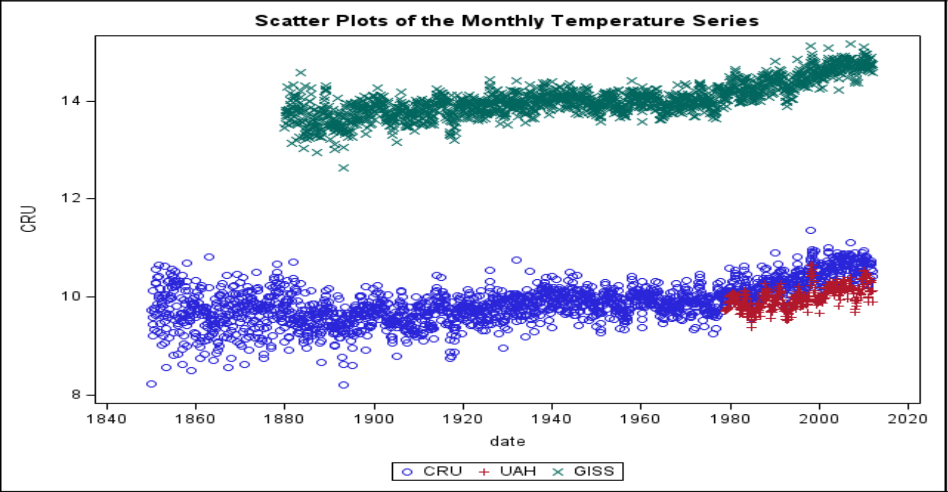
Rajesh Selukar

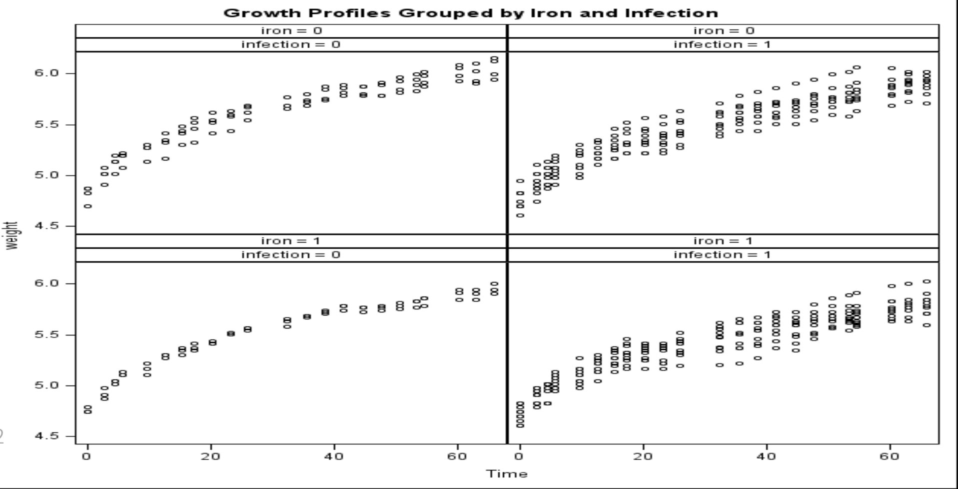
Rajesh.Selukar@sas.com

Time Series and Longitudinal Data









Linear State Space Models (SSMs)

- SSMs are MIXED effects models that are customized for sequential data:
  - the fixed and random effects can vary with time
  - the time-varying effects, α_t, are called states, which evolve autoregressively

$$
\begin{aligned}
Y_t &= X_t \beta + Z_t \alpha_t + \epsilon_t, && \epsilon_t \sim N(0, \Sigma_t) && \textit{Observation Equation} \\
\alpha_{t+1} &= T_t \alpha_t + \eta_{t+1}, && \eta_t \sim N(0, Q_t) && \textit{State Evolution Equation} \\
\alpha_1 &= A\,\delta + \eta_1, && \eta_1 \sim N(0, Q_1) && \textit{Initial Condition}
\end{aligned}
$$

- SSMs incorporate ARIMA (or Box-Jenkins) models and many others as special cases.
- SSMs are also called:
  - Unobserved components models (UCMs)
  - Structural time series models
- All models discussed in this lecture have this of state-space form

SSM-Based Decomposition of Response Curves

- Like an ANOVA (analysis of variance) model, an SSM decomposes the response curves into different components:

  Response curve = trend curve + regression effects + periodic patterns + … + noise

- The terms in the model represent different aspects of the response curve:
  - The trend represents time-varying intercept and is often modeled as a smooth curve
  - Regression effects reflect changes in the response curve due to external effects
  - Periodic patterns are usually associated with the seasonal fluctuations
  - The noise can be simple Gaussian white noise or could be an Autoregressive Moving Average (ARMA) noise, which captures short-term temporal correlation
  - There can be many other types of patterns
- A Model is formulated by choosing suitable patterns from a large library of commonly needed patterns.

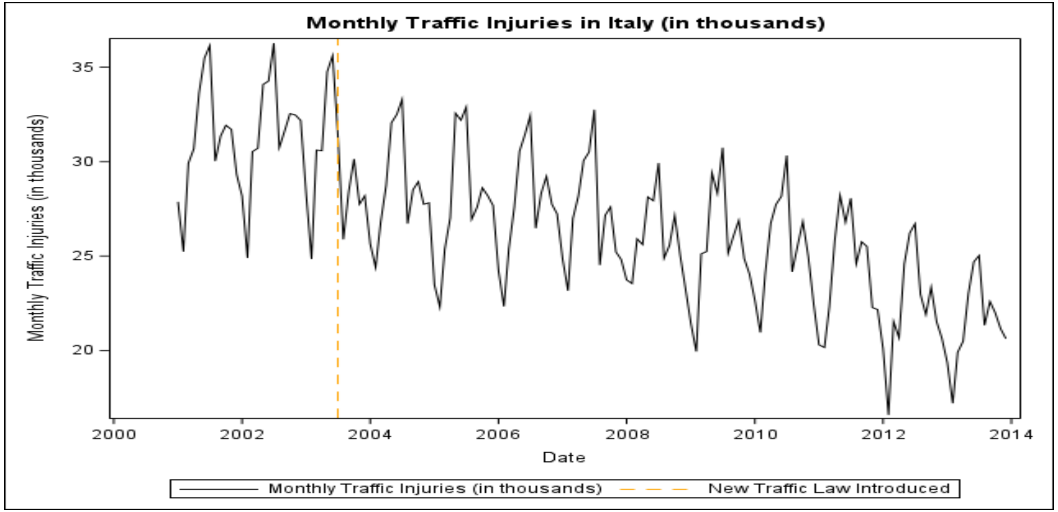
PROC UCM, PROC SSM and PROC CSSM

- PROC UCM and PROC SSM are part of SAS/ETS.
- PROC UCM is used for modeling univariate time series.
- PROC SSM is more general. It is used for modeling:
  - Univariate and multivariate time series
  - Panels of Univariate and multivariate time series
  - Univariate and multivariate longitudinal data
- PROC CSSM, which is part of SAS VIYA/Econometrics, is a Cloud-enabled version of PROC SSM.
- PROC CSSM has more features and is more performant.
- PROC SSM programs can be converted to PROC CSSM with very minor changes.

Univariate Time Series Analysis by PROC UCM

- PROC UCM enables you to easily specify models that capture:
  - A wide variety of trend patterns (random walk, local linear trend, etc.)
  - A variety of periodic effects, with parsimonious handling of long seasonal patterns
  - Different types of regression effects, including
    - Time-invariant and time-varying regression coefficients
    - Nonlinear effects by using splines
    - Lagged regression effects (transfer-function)
  - Lagged-response effects (differencing, lagged response terms)
  - White and colored noise (ARMA) noise
- Rich diagnostics: Residual analysis, Information criteria, structural break detection, …

A Transfer-Function Model for the Monthly Italian Traffic Accident Data
Example 10 in the UCM Doc and a Case Study in Pelagatti (2015)



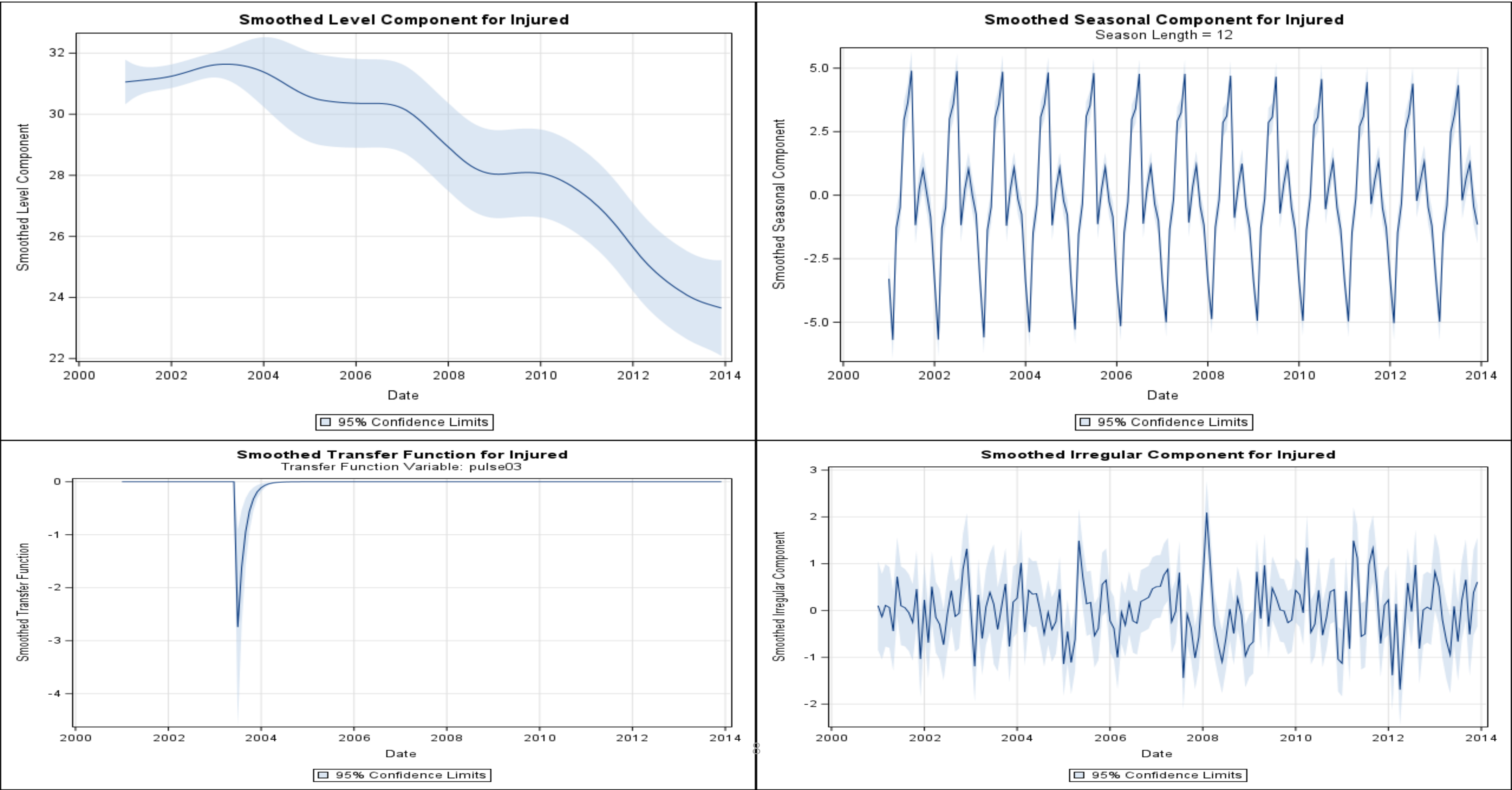

```
proc ucm data=Italy;
  id date interval=month;
  model Injured = shift03;
  level variance=0 noest;
  slope;
  season length=12 type=trig;
  tf pulse03 den=1 tfstart=0 plot=smooth;
  irregular;
  estimate plot=(panel residual) like=marginal;
  forecast plot=decomp;
run;
```

Question: How did the July 2003 intervention impact the time series?

$$y_t = \mu_t + \psi_t + shift03 * \beta + \zeta_t + \epsilon_t$$

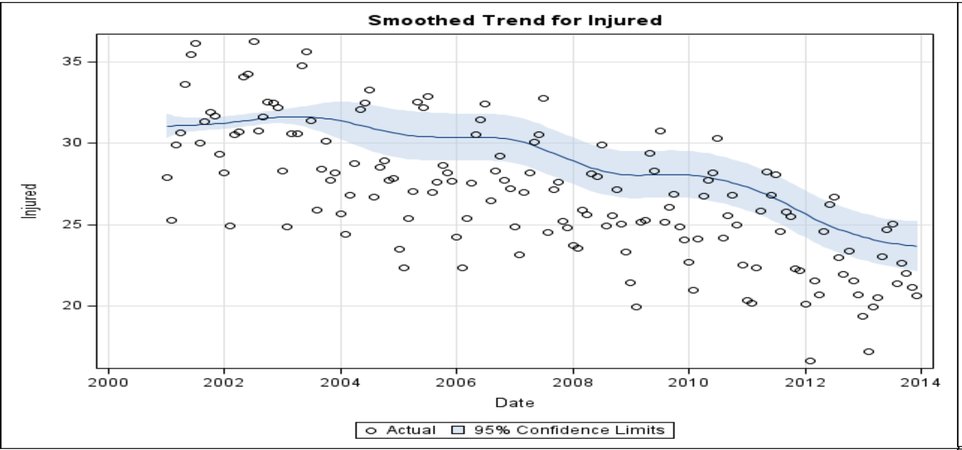
response = integrated RW trend + monthly seasonal effect + permanent shift starting on July 2003 + transient effect of July 2003 intervention + white noise

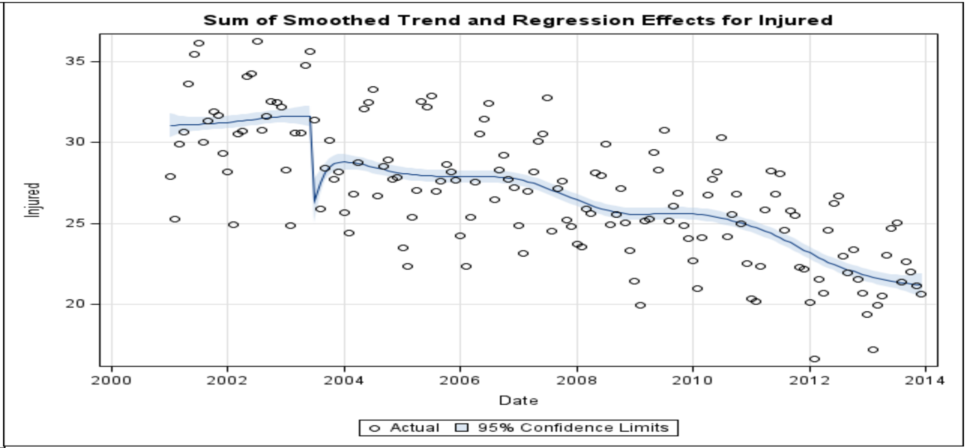
Smoothed (Full-Sample) Estimates of Model Components for the *Injured* Series



Model-Based Decomposition of the *Injured* Series

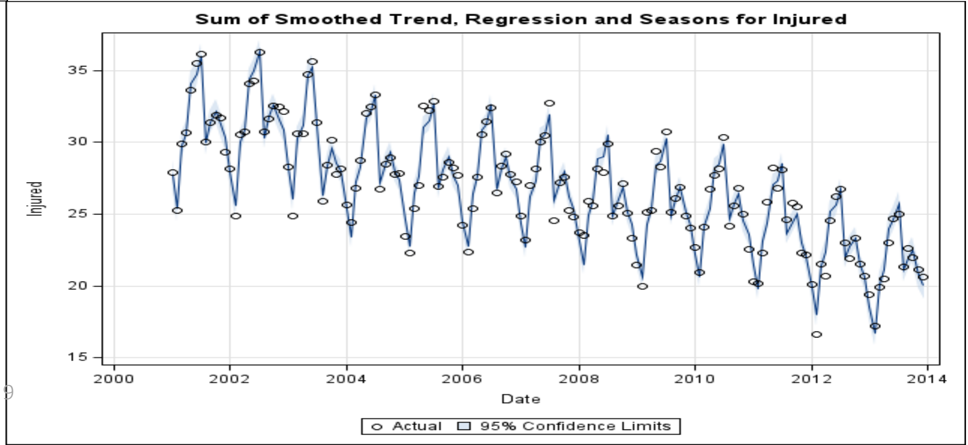
$y_t = \mu_t + \text{shift03} * \beta + \zeta_t + \psi_t + \epsilon_t$ Estimate of $\beta = -2.74$





Much more output is available, such as:

- Estimates of all model parameters
- Residual and structural break diagnostics
- Likelihood-based information criterion
- Hold-out-sample-based model selection statistics
- Extrapolation and interpolation of the response series and model components



Analysis of More General Sequence Data by PROC SSM/PROC CSSM

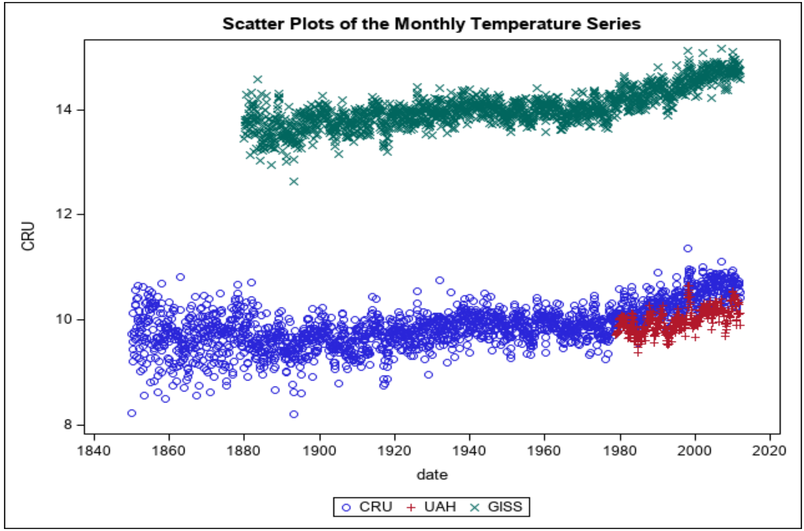
- With PROC SSM/CSSM you can:
    - Model sequential data, such as univariate and multivariate time series, univariate and multivariate longitudinal data, and hierarchical data that result from multi-level, multi-subject, longitudinal studies.
    - Specify very general SSMs:
        - Compose a complex model from smaller pieces
        - Key-word support to specify commonly needed model pieces
        - Rich language to explicitly specify the SSM system matrices, if needed
- Rich diagnostics: Residual analysis, Information criteria, structural break detection, …
- In addition, PROC CSSM also supports scoring-based what-if-analysis, and ongoing monitoring of streaming data

Joint Modeling of Monthly Temperature from Three Weather Data Sources
Example 12 in the SSM Doc

CRU: Climate Research Unit Univ, from Jan 1850.
GISS: Goddard Inst for Space Studies, from Jan 1880.
UAH: Global satellite data from Dec 1978.
For all these series the readings, in centigrade, end at Jan 2012.

A model proposed by Ansley and de Jong (2015) in *Inferring and Predicting Global Temperature Trends*:

$$GISS_t = \mu_t + \qquad colored\ noise_{1t}$$
$$CRU_t = \mu_t + \beta_{cru} + colored\ noise_{2t}$$
$$UAH_t = \mu_t + \beta_{uah} + colored\ noise_{3t}$$

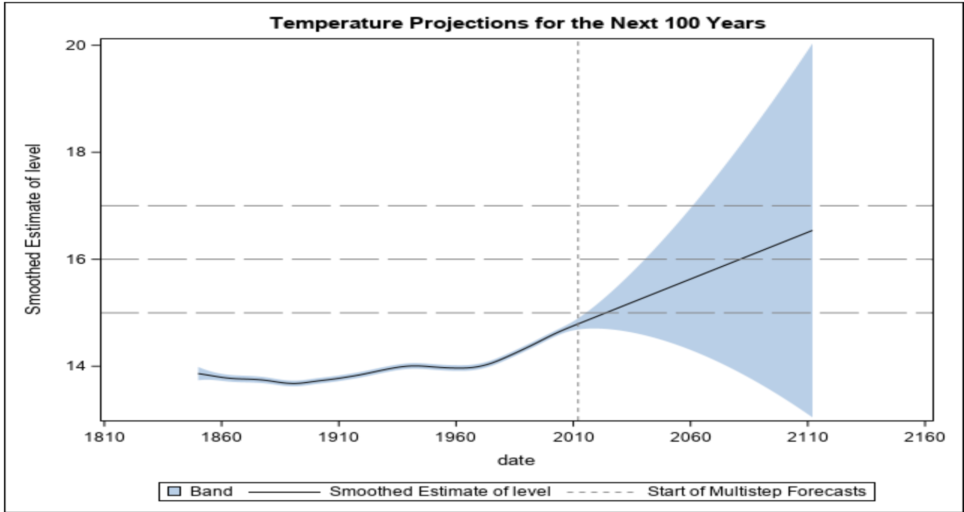


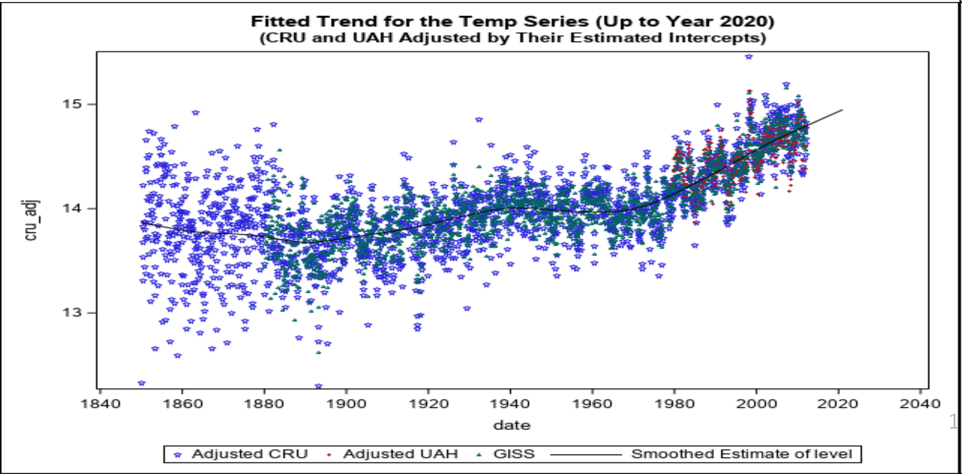
All three series share the same trend component, μ_t, which is an integrated RW, and the colored noise components for each series also share some common aspects (not shown).

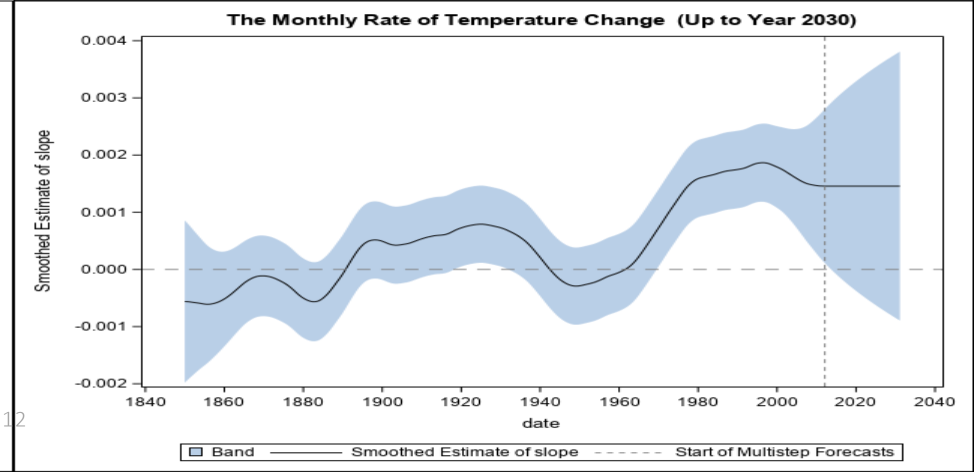
In their paper, Ansley and de Jong comment on the necessity of joint modeling of these three series for obtaining better estimate of the underlying trend and the superiority of this type of modeling over commonly used curve fitting techniques in climate science.

Model-Based Extrapolation of Temperature Trend and Its Slope

- The top right plot shows the temp trend, and the lower right plot shows the slope of the temp trend.
- The lower left plot shows the temp trend line along with the appropriately adjusted observed temp readings.
- The trend-slope plot shows that from the 1970s to 2000, the slope is positive and statistically significant but after that it becomes less so.

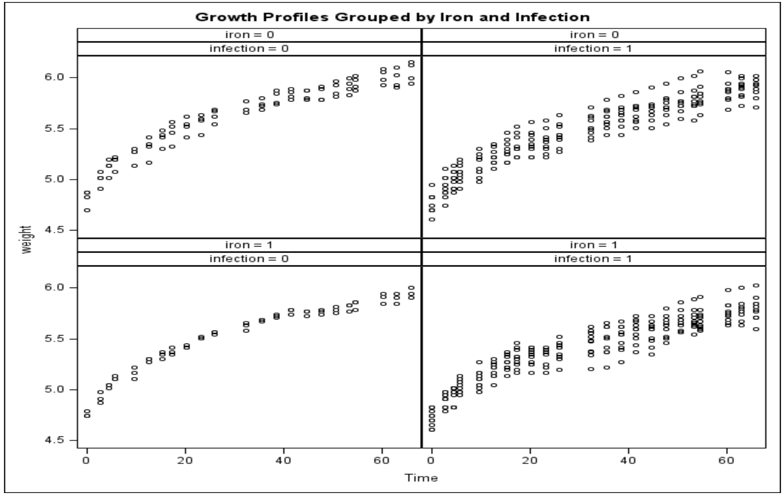






Modeling Weight Profiles of Cows in a Longitudinal Study
Slightly Altered Version of Example 4 in the SSM Doc

- Growth profiles of 26 cows are monitored over a 22-month period.
- Their weights are not measured at equally spaced time points and the spacing of the time points can be different for each cow.
- Cows are grouped in a 2 x 2 design: receiving iron dosing or not, is infected by M. paratuberculosis or not



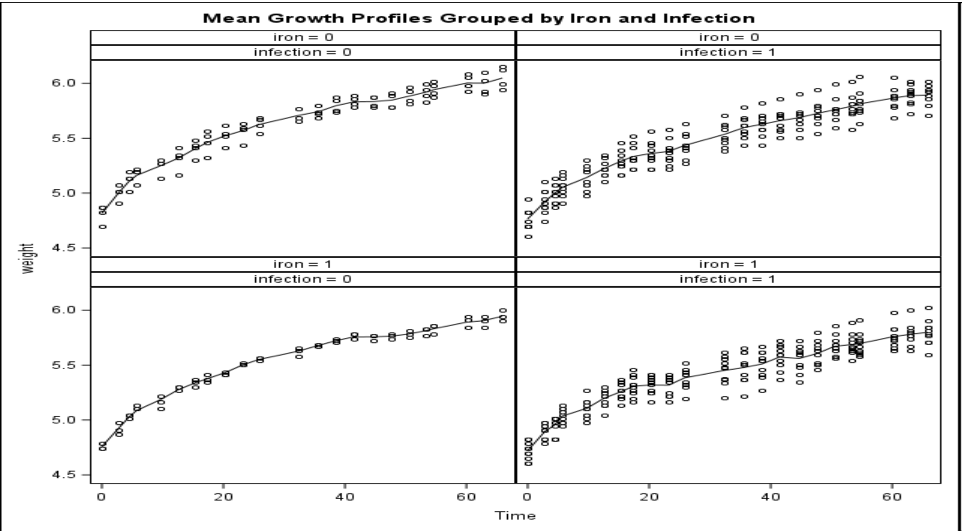
Goal of the study: compare the growth profiles of cows in each category.

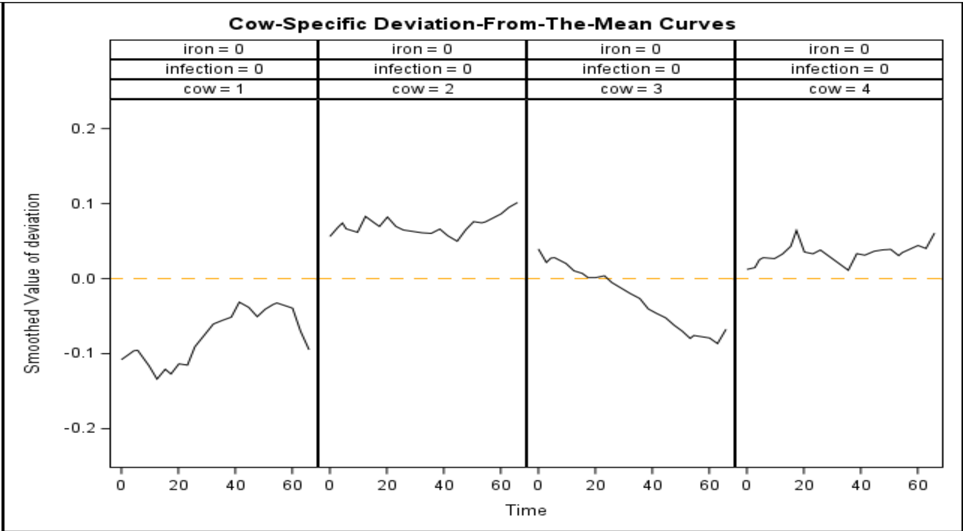
A model for the weight of j-th cow in the i-th group at time t:

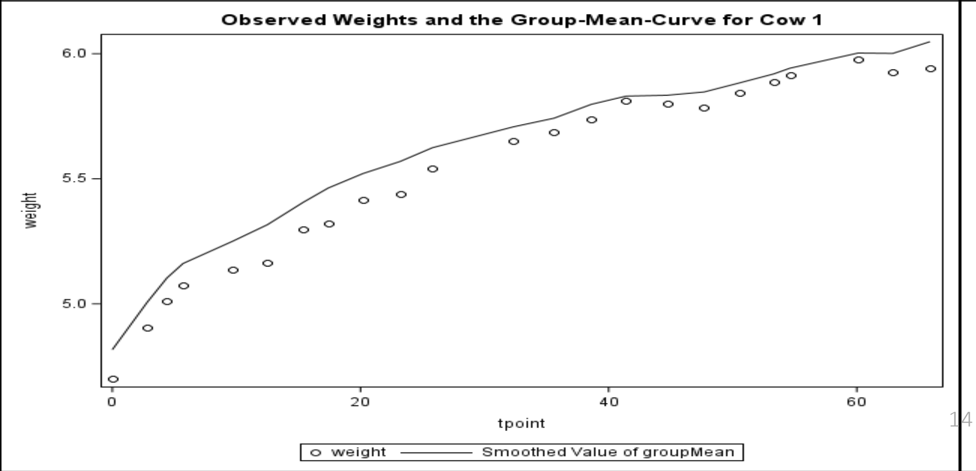
$$y_{ijt} = \mu_{it} + \psi_{ijt} + \epsilon_{ijt}$$

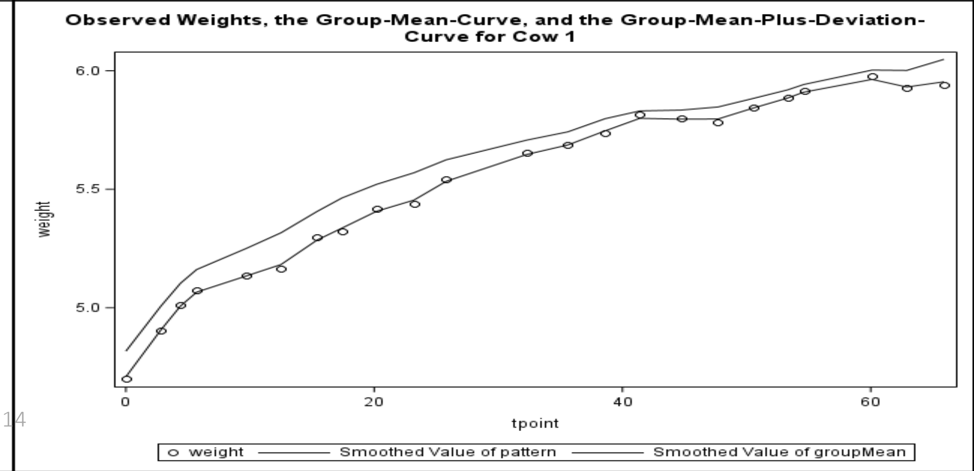
- μ_it is the group-mean-curve for the i-th group (i=1, 2, 3 4), modeled as a 2nd-order-polynomial smoothing-spline (a fixed-effect-curve). A functional fixed effect.
- ψ_{ijt} is the deviation curve of the j-th cow from the i-th mean curve, modeled as a 1st-order-polynomial smoothing-spline (a random-effect-curve). A functional random effect.
- ϵ_{ijt} is the observation error, modeled as white noise

Cow-Growth Study: Full-Sample Component Estimates



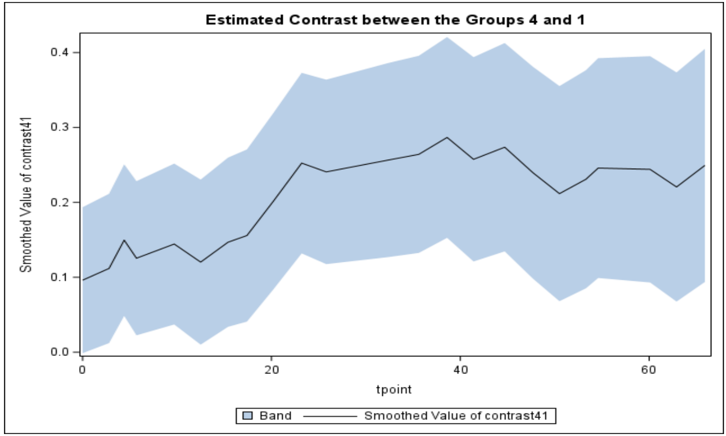






Cow Growth Study: Group-Mean Contrasts

- The curve in the right-hand side plot shows the estimate of the difference between the group-mean curves of Group 1 and Group 4, $(\mu_{1t} - \mu_{4t})$, with 95% confidence band (pointwise).
- It shows that the difference in these mean curves is statistically significant throughout the observation period.
- Group1: iron=0, infection=0, and Group4: iron=1, infection=1.
- Of course, similar contrasts can be estimated for other combinations.



- The SSMs that are used to model longitudinal data must permit arbitrary spacing of time points. Such SSMs are called continuous-time SSMs.
- Such continuous-time SSMs provide very interpretable models for multi-level, multi-subject longitudinal studies.
- For more information, see the SGF paper: [*Functional Modeling of Longitudinal Data with the SSM Procedure*] by Rajesh Selukar.

Additional Info

- The UCM procedure documentation.
- The SSM procedure documentation.
- The CSSM procedure documentation.
- Books (loosely ordered by degree of technical detail):
  - Pelagatti, M. M. (2015). *Time Series Modelling with Unobserved Components*. Boca Raton, FL: CRC Press.
  - Harvey, A. C. (1989). *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge: Cambridge University Press.
  - Durbin, J., and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. 2nd ed. Oxford: Oxford University Press.
- SGF Papers and BLOGS:
  - Selukar, R. S. (2017). "Detecting and Adjusting Structural Breaks in Time Series and Panel Data Using the SSM Procedure."
  - Selukar, R.S. (2021). "Using State Space Models for the Stability Monitoring of Streaming Data."

Closing Remarks

- Linear SSMs, which got their start in engineering, have been in use since the 1960s.
- In recent decades they have found applicability in many fields other than engineering.
- PROC CSSM, PROC SSM, and PROC UCM provide very convenient framework for SSM-based data analysis.
- Development of all these procedures remains active:
    - New features for added modeling capabilities
    - Improved scalability for handling larger and more complex SSMs
- For modeling/syntax help you can contact me at rajesh.selukar@sas.com
- Thanks for listening!!