Model-Based and Semi-Parametric Estimation of Time Series Components and Mean Square Error of Estimators

Michael Sverchkov

(Bureau of Labor Statistics)

Based on: Pfeffermann, D., and Sverchkov, M. (2014). Estimation of mean square error of X-11-ARIMA and other estimators of time series components. *Journal of Official Statistics,* **30**, No.4, pp. 811 - 838

Content

We define seasonal and trend components under which the X-11 ARIMA estimators of them are almost unbiased at least in the central part of the series.

We define the Variance and Mean Square Error (MSE) of X-11 ARIMA and Basic State-Space Structural Model (BSM) estimators with respect to the newly defined trend and seasonal components and propose estimators for the Bias, Variance and MSE.

We investigate the behavior of the X-11 ARIMA and BSM estimators of the newly defined trend and seasonal components and their MSE estimators in a small simulation study based on real data.

Definitions

$$y_t = G_t + e_t, \qquad t = \underbrace{t_{start},\ldots,0}_{unobserved}, \underbrace{1,\ldots,N}_{y_t-observed}, \underbrace{N+1,\ldots,\infty}_{unobserved};$$

y_t - **target time series** (*observed for* $t = 1,\ldots,N$ ***only***)

G_t - **signal** (*unobserved*), e_t - **Error** (*unobserved*)

Assumptions: e_t, G_t **independent,** $E(e_t) = 0$, $Var(e_t) < \infty$

Signal and Error can be defined in several ways. We consider the following alternative definitions:

I) Signal (G_t) = Trend (T_t) + Seasonal component (S_t),

Error (e_t) = Irregular term (I_t) + Sampling error (ε_t) (Pfeffermann, 1994).

II) Signal (G_t) = Trend (T_t) + Seasonal component (S_t) + Irregular term (I_t),

Error (e_t) = Sampling error (ε_t) (Bell and Kramer 1999).

Components of interest

Seasonally Adjusted Series $A_t = T_t + I_t$, and/or **Trend** T_t.

X-11 ARIMA Estimators

X-11 ARIMA estimators of trend and seasonal components can be approximated as:

$$\hat{T}_t = \sum_{k=-(t-1)}^{N-t} \boldsymbol{w_{kt}^T} y_{t+k}\,, \qquad \hat{S}_t = \sum_{k=-(t-1)}^{N-t} \boldsymbol{w_{kt}^S} y_{t+k} \quad \Rightarrow \quad \hat{A}_t = y_t - \hat{S}_t = \sum_{k} w_{kt}^A y_{t-k}$$

Filters w_{kt}^S and w_{kt}^T defined by the X-11 ARIMA program options and length of series, N.

At **center** part of series, filters **time-invariant & symmetric;**

$$\boldsymbol{w_{kt}^T = w_k^T}\,,\ \boldsymbol{w_{-k}^T = w_k^T}\ ,\ \boldsymbol{a_T} \le t \le N - \boldsymbol{a_T}\,;$$

$$\boldsymbol{w_{kt}^S = w_k^S}\,,\ \boldsymbol{w_{-k}^S = w_k^S}\,,\ \boldsymbol{a_S} \le t \le N - \boldsymbol{a_S}\,.$$

BLS

X-11 ARIMA Estimators (cont.)

Notice: X-11 ARIMA contains “non-linear” operations:

Identification and estimation of ARIMA models,

Identification and gradual replacement of extreme observations.

We assume:

Time series under consideration already corrected for outliers,

Effects of identification and estimation of ARIMA models are negligible (verified in previous studies).

What does X-11 ARIMA estimate?

Model implies: $E(\hat{S}_t \mid \mathbf{G}) = \sum_{k=-(t-1)}^{N-t} w_{kt}^{S} G_{t-k} \overset{def}{=} \tilde{S}_t$, where $\mathbf{G} = (G_{t_{start}}, G_{t_{start}+1}, \ldots)$.

(Similarly for trend and seasonally adjusted components.)

$\Rightarrow$ $\hat{S}_t$ - unbiased estimator of $\tilde{S}_t$ but $\tilde{S}_t$ not time-invariant. For example, $\tilde{S}_N$ based on $y_1, \ldots, y_N$ is not equal to $\tilde{S}_N$ based on $y_1, \ldots, y_N, y_{N+1}, \ldots, y_{2N}$.

$\Rightarrow$ $\tilde{\mathbf{S}} = \{\tilde{S}_t,\ t = 0, \ldots, \infty\}$ **not** a seasonal component (not defined uniquely).

X-11 decomposition

Assuming t_{start} << 0 define **"X-11 trend and seasonal components"** as:

$$S_t^{x11} = \sum_{k=-a_S}^{a_S} w_k^S G_{t-k}, \quad T_t^{x11} = \sum_{k=-a_T}^{a_T} w_k^T G_{t-k}.$$

(Bell & Kramer,1999)

$\Rightarrow$ **X-11 decomposes the observed series into the "X-11-components"**

$$y_t = T_t^{x11} + S_t^{x11} + e_t^{x11}; \quad e_t^{x11} = y_t - T_t^{x11} - S_t^{x11}$$

- **X-11 estimators almost unbiased at the center part of the series with respect to this decomposition**.

X-11 decomposition (cont.)

- X-11 ARIMA in common use all over the world. Users trust the method.

  Important to study what is estimated.

- Estimator defines the parameter**; not** parameter defines the estimator.

- Conditioning on signal ⇒ viewing trend **&** seasonal effects as fixed

  population parameters ⇒ conforms with **classical sampling theory.**

- Parameters must be defined by symmetric filters.

Conditional Bias, variance and MSE of X-11 ARIMA estimators

$$\textbf{Bias}(\hat{T}_t \mid \mathbf{G}) = E[(\hat{T}_t - T_t^{X11}) \mid \mathbf{G}] = \sum_{k=-(t-1)}^{N-t} \mathbf{w}_{kt}^{T} G_{t+k} - \sum_{k=-a_T}^{a_T} \mathbf{w}_{k}^{T} G_{t+k},$$

$$\textbf{Var}[\hat{T}_t \mid \mathbf{G}] = E[\sum_{k=-(t-1)}^{N-t} w_{kt}^{T}(y_{t+k} - G_{t+k})]^2 = E(\sum_{k=-(t-1)}^{N-t} w_{kt}^{T} e_{t+k})^2,$$

$$\textbf{MSE}(\hat{T}_t \mid \mathbf{G}) = E[(\hat{T}_t - T_t^{X11})^2 \mid \mathbf{G}] = \textbf{Var}(\hat{T}_t \mid \mathbf{G}) + \textbf{Bias}^2(\hat{T}_t \mid \mathbf{G}).$$

$$\Downarrow$$

$MSE[\hat{S}_t]$ **can be estimated by estimating** $Var[\hat{S}_t \mid \mathbf{G}]$ **and** $Bias[\hat{S}_t \mid \mathbf{G}]$

- Similar expressions for any estimator $\tilde{H}_t \approx \sum_{k=-(t-1)}^{N-t} h_{kt} y_{t+k}$ with arbitrary weights $\{h_{kt}\}$,

e.g., Basic Structural Model

Bell and Kramer (1999)

BK propose **same** target components. They estimate the components by augmenting the series with sufficient **minimum mean squared error** forecasts and backcasts under the ARIMA model, such that the **symmetric filters** can be applied to the augmented series at every time point with observation.

$\hat{T}_t^{BK} = \sum_{k=-a_T}^{a_T} w_k^T y_{t+k}^*$; $y_{t+k}^* = y_{t+k}$ if y_{t+k} observed, y_{t+k}^* = forecast or backcast otherwise.

$E(y_{t+k}^* - y_{t+k}) = 0$ (**unconditionally**) $\Rightarrow$ $E(\hat{T}_t^{BK} - T_t^{X11}) = E[\sum_{k=-a_S}^{a_S} w_k^T y_{t+k}^* - \sum_{k=-a_S}^{a_S} w_k^T y_{t+k}] = \mathbf{0}$ (**unconditionally**).

- **BK** estimate $Var(\hat{T}_t^{BK} - T_t^{X11})$ over distributions of **sampling errors** and forecast & backcast **prediction errors**.
- Time series **irregulars** considered as part of the **signal**.

Where do we differ?

We **condition** on $\mathbf{G} = \{G_t, t = t_{start}, ..., \infty\}$, in which case in general $E[(\hat{T}_t^{BK} - T_t^{X11}) \mid \mathbf{G}] \neq 0$

Bias may also exist even **unconditionally** when extrapolating less than required for use of symmetric filters, depending on distribution of the signal.

- Estimation of **conditional MSE** not restricted to full forecasts and backcasts, and can be applied when estimating the target components with only **one** or **two** years of forecasts and backcasts (common case**?**) or without ARIMA extrapolations**,** or when estimating the components by fitting different models (**e.g.,** Basic Structural Model).

Where do we differ? (cont.)

We attempt to estimate the **conditional MSE**, **given the signal**.

Alternatively, when the bias estimator is obtained optimally under a model, it may be viewed as estimating the **unconditional bias** over **all possible realizations of the signal** under the model, **given** the observed series (see below).

- The proposed approach is applicable also when the signal consists of only the trend and the seasonal effect, and the time series irregular is part of the error.

Estimation of conditional MSE

Conditional variance

- When the time series irregular is part of the signal,

$$\hat{Var}(\hat{T}_t \mid \mathbf{G}) = \hat{E}\Big(\sum_{k=-(t-1)}^{N-t} w_{kt}^T \varepsilon_{t+k}\Big)^2 = \sum_k \sum_l w_{kt}^T w_{lt}^T \hat{Cov}(\varepsilon_{t+k}, \varepsilon_{t+l}).$$

Easily computable when estimators of the variances and covariances of the sampling errors are available.

- When the irregular is not part of the signal, $G_t = S_t + T_t$, $e_t = I_t + \varepsilon_t$;

$$\hat{Var}(\hat{T}_t \mid \mathbf{G}) = \sum_k \sum_l w_{kt}^T w_{lt}^T \hat{Cov}(e_{t+k}, e_{t+l}).$$

Methods for estimating $Cov(e_{t+k}, e_{t+l})$ considered in literature

(Pfeffermann 1994, Pfeffermann & Scott (1997), Chen *et al.* (2003))

Idea of estimating $Cov(e_{t+k}, e_{t+l})$**:**

$$Cov(\hat{R}_t, \hat{R}_m \mid \mathbf{G}) = Cov[\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k}, \sum_{l=-(m-1)}^{N-m} w_{lm}^R e_{m+l}] = \sum_k \sum_l w_{kt}^R w_{lm}^R Cov(e_{t+k}, e_{m+l}) \quad \textbf{(1)}$$

$$w_{0t}^R = 1 - w_{0t}^S - w_{0t}^T, w_{kt}^R = 1 - w_{kt}^S - w_{kt}^T, \; k \neq 0.$$

$$\hat{R}_t = y_t - \hat{S}_t - \hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^R y_{t+k} \quad \textbf{- observed,}$$

$$Cov(\hat{R}_t, \hat{R}_m \mid \mathbf{G}) \quad \textbf{- can be estimated}$$

Assuming I_t - stationary, $Cov(I_{t+k}, I_{t+l}) = 0, \; |k-l| > Const$ and $C\hat{o}v(\varepsilon_{t+k}, \varepsilon_{t+l})$ are available

or

assuming e_t - stationary and $Cov(e_{t+k}, e_{t+l}) = 0, \; |k-l| > Const$

(1) can be solved

Estimation of conditional bias

Within X-11 ARIMA framework, the signal can be estimated conveniently (but **not efficiently**) as:

(a) Use the model chosen by program to forecast–backcast $\boldsymbol{m} = max(a_T, a_S)$ observations.

Let $N^{aug} = N + 2m$.

(b) Estimate $\hat{G}_t^{aug} = \sum_{k=-(t-1)}^{N^{aug}-t} w_{kt}^{S,aug} y_{t+k}^{aug} + \sum_{k=-(t-1)}^{N^{aug}-t} w_{kt}^{T,aug} y_{t+k}^{aug}$; $t = -m+1, \ldots, N+m$

$y_t^{aug} = y_t$ if y_t observed, y_t^{aug} = forecast (backcast) otherwise.

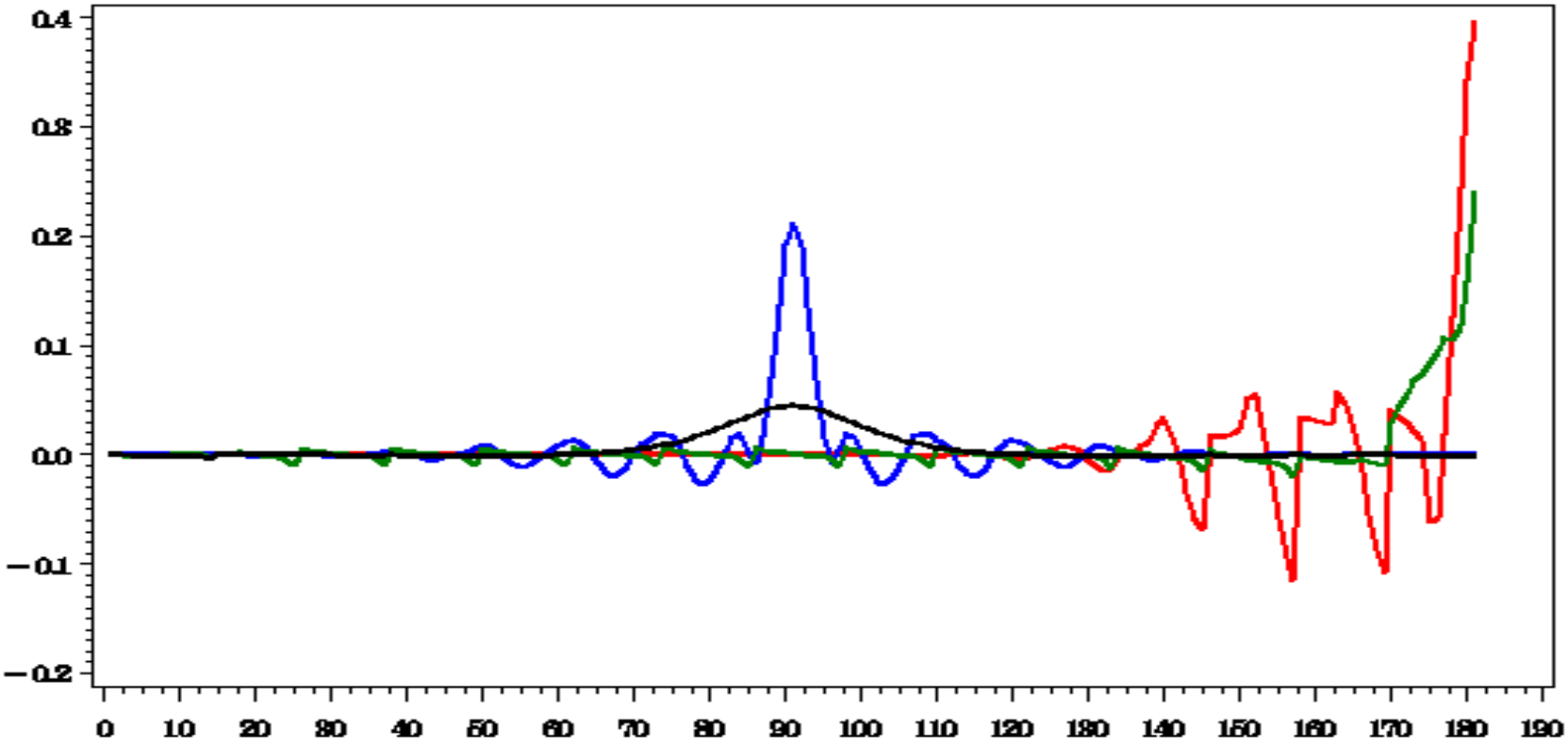
$$\boldsymbol{Bi\hat{a}s}[\hat{T}_t \mid \mathbf{G}] = \hat{E}[(\hat{T}_t - T_t^{X11}) \mid \mathbf{G}] = \sum_{k=-(t-1)}^{N-t} w_{kt}^T \hat{G}_{t+k}^{aug} - \sum_{k=-a_T}^{a_T} w_k^T \hat{G}_{t+k}^{aug} ; \quad t = 1, \ldots, N.$$

Estimation of conditional bias (cont.)

Signal can be estimated more **efficiently** by extracting the models for the trend and seasonal effects using **signal extraction**, and then estimate within the observation period, and forecast and backcast under the extracted models.

- Estimators of components are in this case **minimum MSE** (**MMSE**) under the models.
- **MMSE** estimator is the conditional expectation given the observed series. The bias estimator is then the **unconditional expectation** of the bias over all possible realizations of the signal **given the observed series**.
- When predicting the signal many time points ahead, estimators may perform **very badly**.
- The filter weights $\downarrow 0$ fast when moving away from the time point of interest $\Rightarrow$ possibly large biases of estimators of the signal for distant time points may have little effect on the bias of the bias estimator.

Figure 1. **Central** and **end** weights when estimating the **trend** by **default X-11** and under **Basic Structural Model**.



---- Central X-11 ; ---- Central BSM

---- End point X-11 ; ---- End point BSM

Estimation of conditional MSE

$$\hat{MSE}(\hat{T}_t \mid \mathbf{G}) = \hat{Var}(\hat{T}_t \mid \mathbf{G}) + \hat{Bias}^2(\hat{T}_t \mid \mathbf{G}).$$

Generally, **over-estimator** because,

$$E[\hat{Bias}^2(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}]$$

$$= \{E[\hat{Bias}(\hat{T}_t \mid \boldsymbol{G})] \mid \boldsymbol{G}\}^2 + Var[\hat{Bias}(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}] \geq \{E[\hat{Bias}(\hat{T}_t \mid \boldsymbol{G})] \mid \boldsymbol{G}\}^2.$$

Overestimation **corrected** by subtracting $\hat{Var}[\hat{Bias}(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}]$.

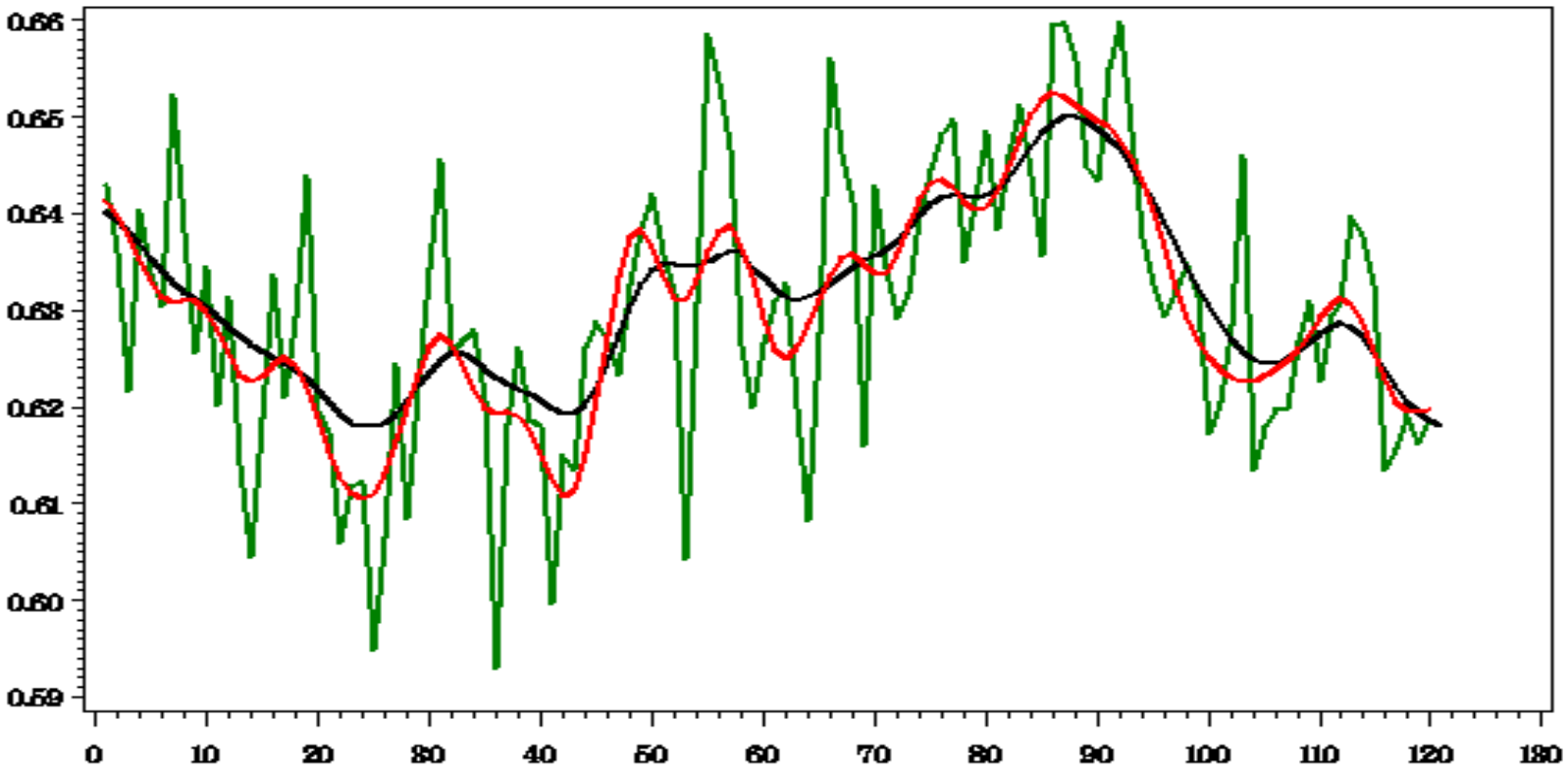
- $\hat{Bias}(\hat{T}_t \mid \mathbf{G})$ is again a **linear combination** of the observed values so the variance is estimated similarly to the estimation of the variance of the estimators of the components.

Simulation study

Simulate series from model fitted to the series **"Employment to Population Ratio in DC"**.

- **Erratic series**: X-11 ARIMA residual component explains **55%** of month to month changes and **32%** of yearly changes. Major portion of the residual is **sampling error**.

Figure 2. Employment to Population Ratio in DC, 1999-2000. Original series and trends estimated by **X-11 ARIMA** with **12** forecasts and under **BSM**.



---- Original series ; ---- Trend **X11 ARIMA** ; ---- Trend **BSM**

Basic Structural Model (BSM) used for simulation

$$y_t = Y_t + \varepsilon_t = T_t + S_t + I_t + \varepsilon_t$$

$$T_t = T_{t-1} + R_t; \quad R_t = R_{t-1} + \eta_{Rt}; \quad S_t = \sum_{j=1}^{6} S_{j,t};$$

$$S_{jt} = \cos\omega_j S_{j,t-1} + \sin\omega_j S^*_{j,t-1} + \eta_{Sjt}$$

$$S^*_{jt} = -\sin\omega_j S_{j,t-1} + \cos\omega_j S^*_{j,t-1} + \eta^*_{Sjt} \quad , \quad S_{ti} = \sum_{j=1}^{6} S_{jti}$$

$$\omega_j = 2\pi\, j/12,\ j = 1,\ldots,6$$

$I_t, \eta_{Rt}, \eta_{Sj,t}, \eta^*_{Sj,t}$ are mutually independent normal disturbances.

ε_t (***sam . err***) ~ **AR(15)** → accounts for **CPS** sampling design.

Simulation plan

S1- Generate **1,000** series $y_t^b, b = 1,...,1,000$ of length **300** from **BSM;** $y_t^b = T_t^b + S_t^b + I_t^b + \varepsilon_t^b, t = 1,...,300$. For present study, $G_t = T_t + S_t + I_t$.

Generate additional **1,000** series as: fix the signal of **2nd** series and add the sampling errors from the first 1,000 series to the fixed signal.

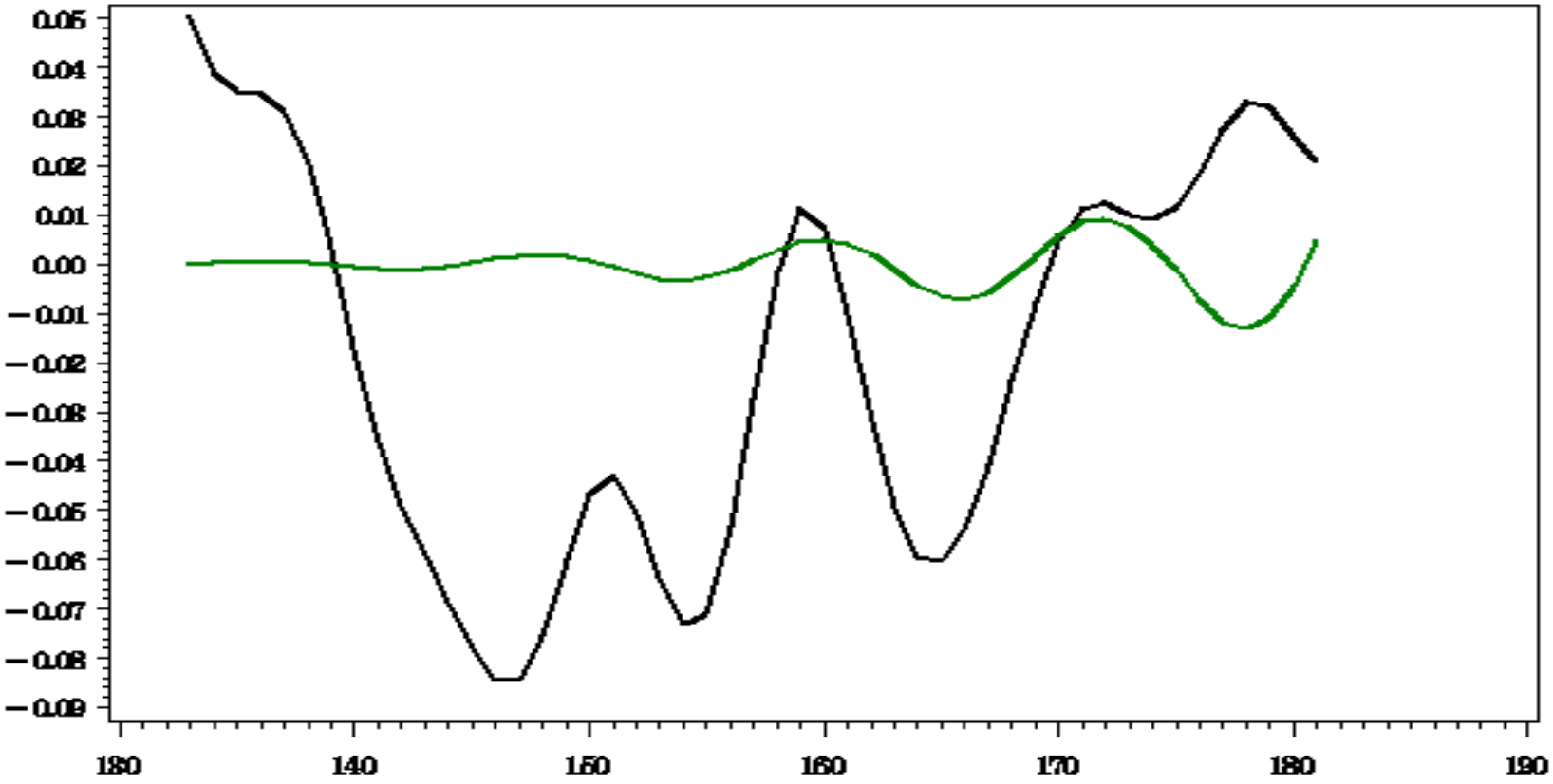
- **First set** of series illustrates **unconditional** properties.
  2nd set illustrates **conditional** properties, given **signal**.

S2- Compute default **X-11** estimator of trend and seasonal component for each simulated signal to obtain the **target X-11 components** for **central 180 months**.

S3- Remove first and last **60** observations from simulated series and apply **X-11 ARIMA** with **12** and **60** forecasts, using default filters and **ARIMA (0,1,1),(0,1,1)** model. (Model chosen by method for first 10 series in each of two data sets.)

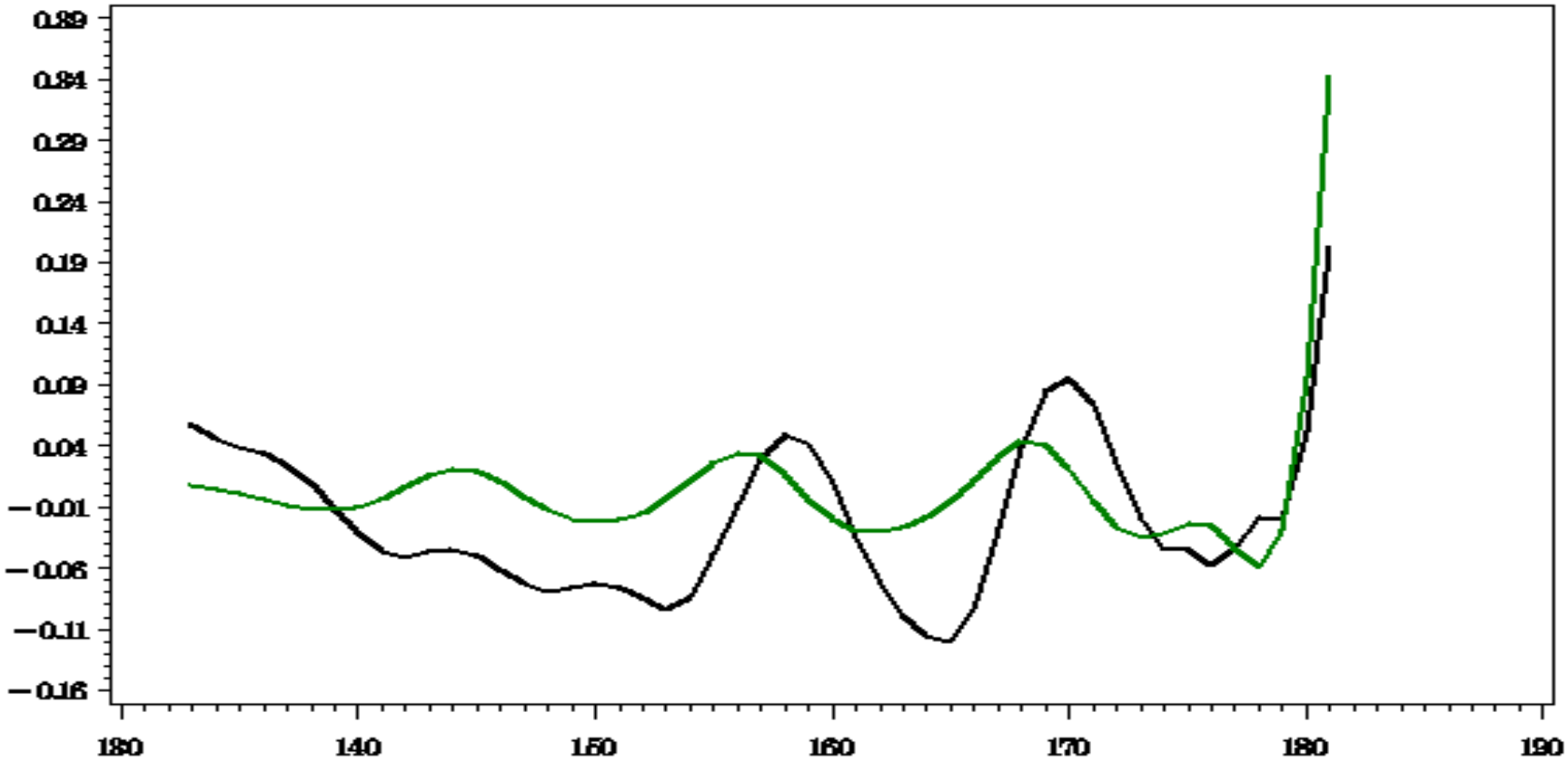
S4- Fit the **BSM** for each of the series of length 180.

Figure 3. Empirical **unconditional bias** of **X-11 ARIMA** trend estimates with **60** forecasts and **mean of bias estimates**, signals estimated by X-11 ARIMA.



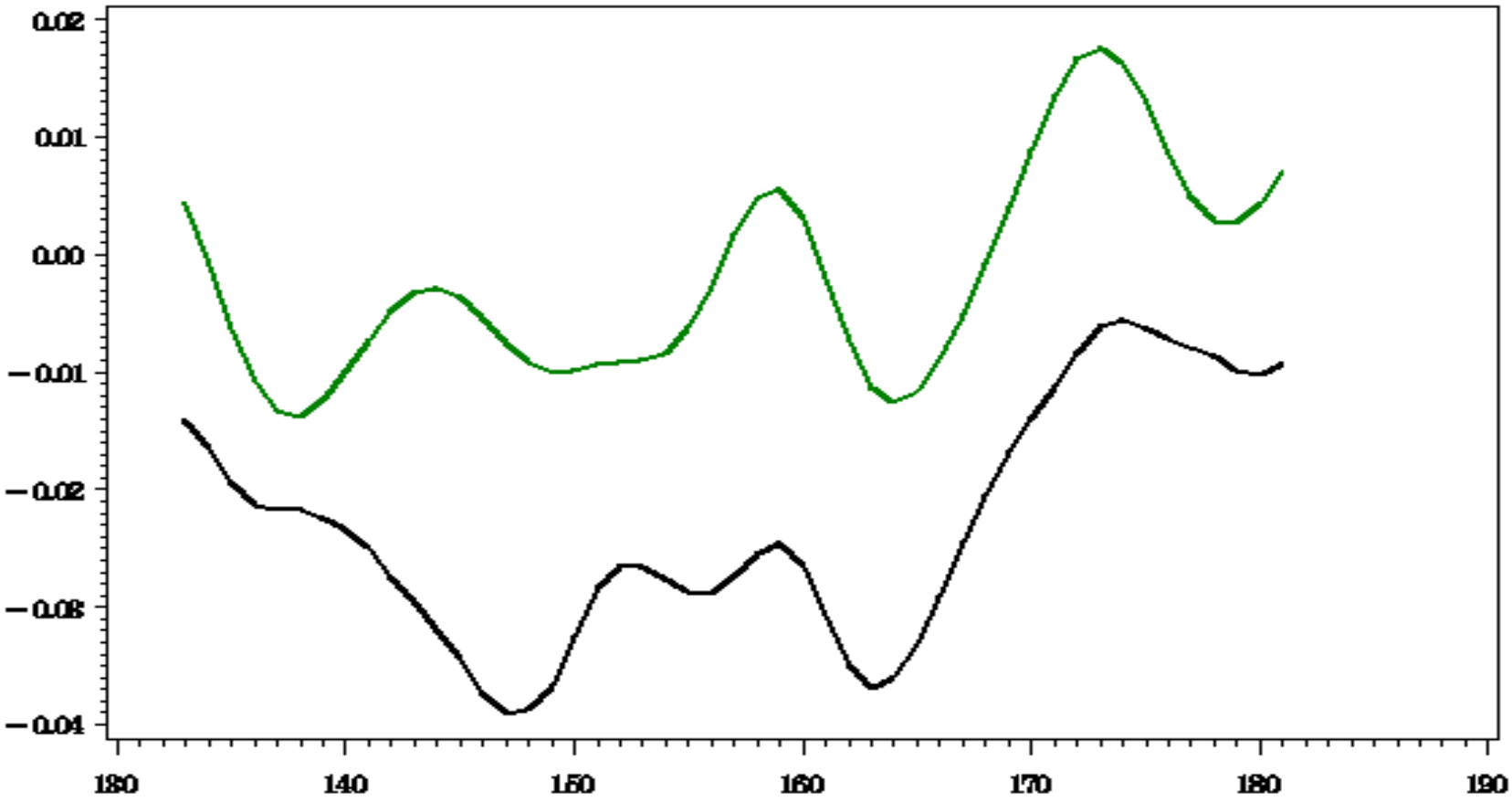
---- Empirical bias ; ---- Mean of estimates

Figure 4. Empirical **conditional Bias** of **X-11 ARIMA** trend estimates with **60** forecasts and **mean of bias estimates**, signal estimated by X-11 ARIMA.



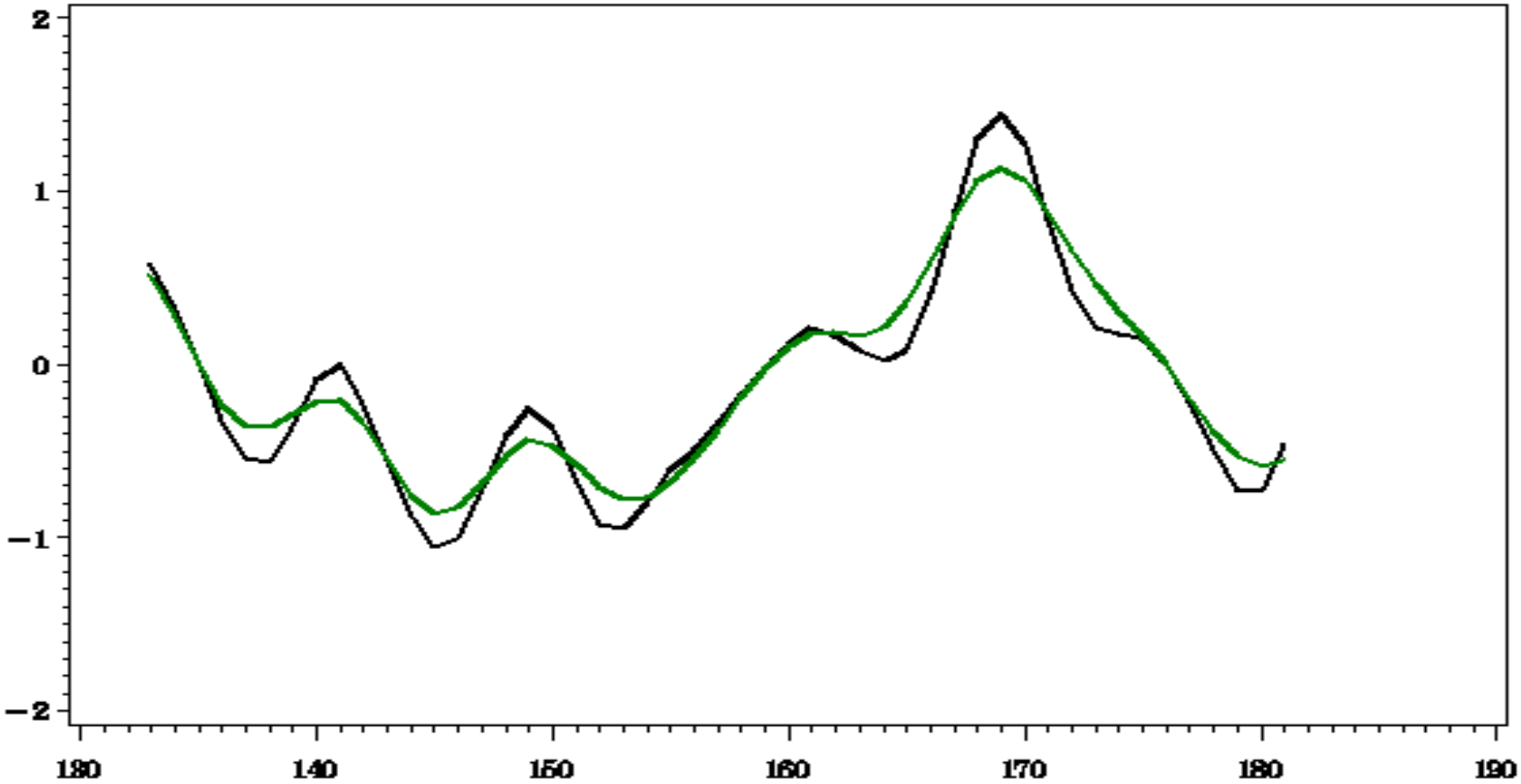
---- Empirical conditional bias ; ---- Mean of estimates

Figure 5. Empirical unconditional bias of BSM trend estimates and mean of bias estimates, signals estimated by X-11 ARIMA.



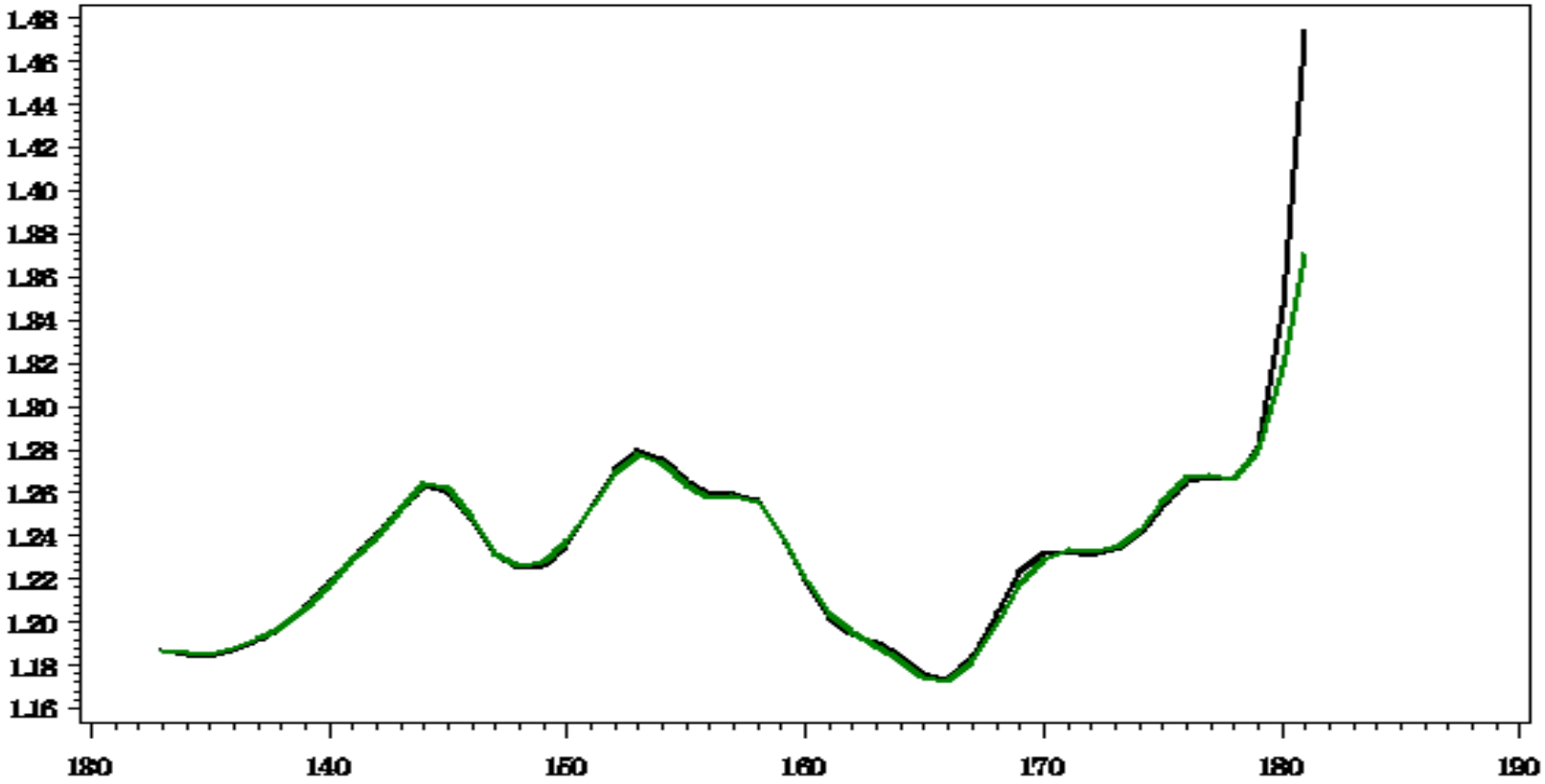
---- Empirical bias ; ---- Mean of estimates

Figure 6. Empirical conditional bias of BSM trend estimates and mean of bias estimates, signal estimated by X-11 ARIMA.



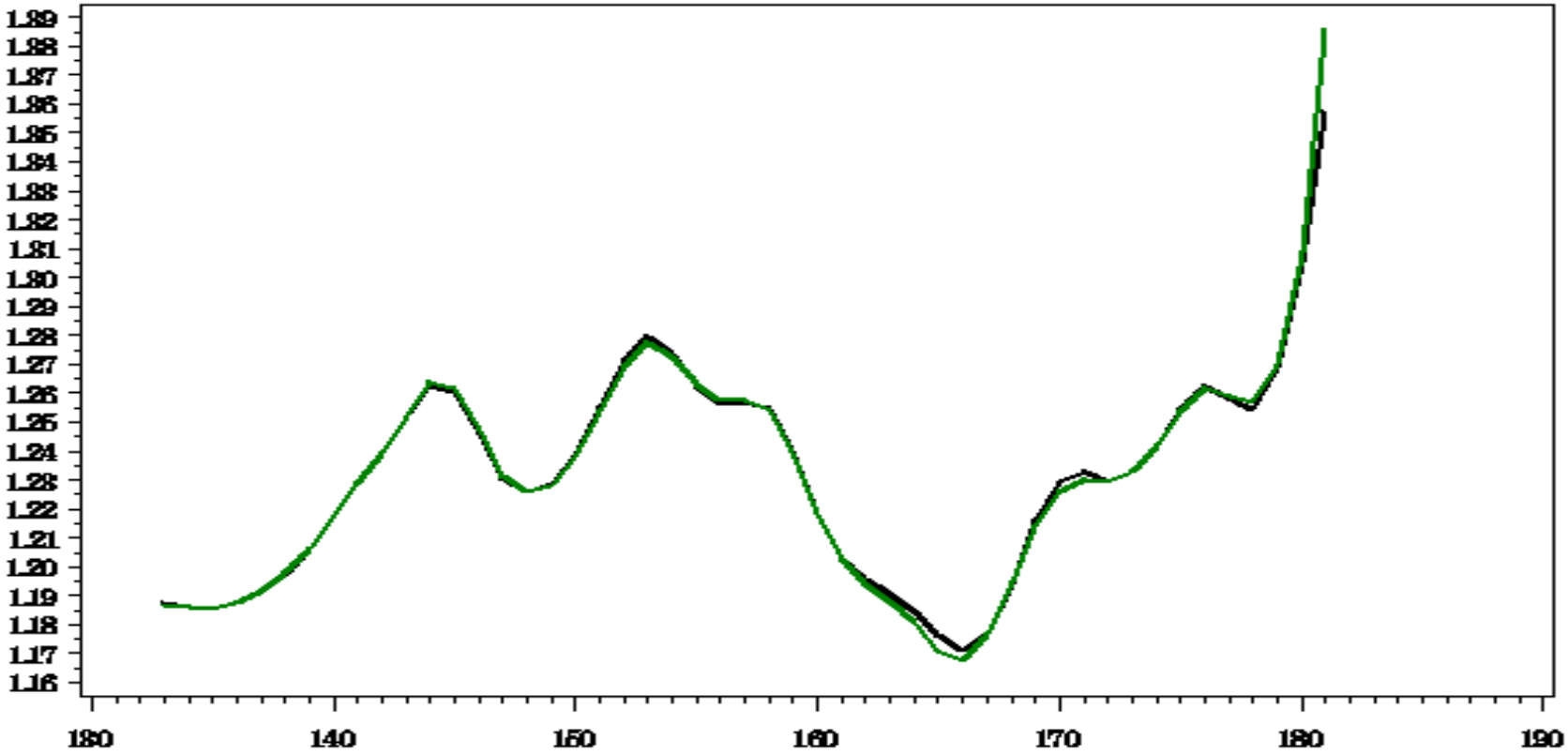
---- Empirical bias ; ---- Mean of bias estimates

Figure 7. Empirical unconditional RMSE of **X-11 ARIMA** trend estimates with **60** forecasts and **mean of RMSE estimates**, signals estimated by X-11 ARIMA.



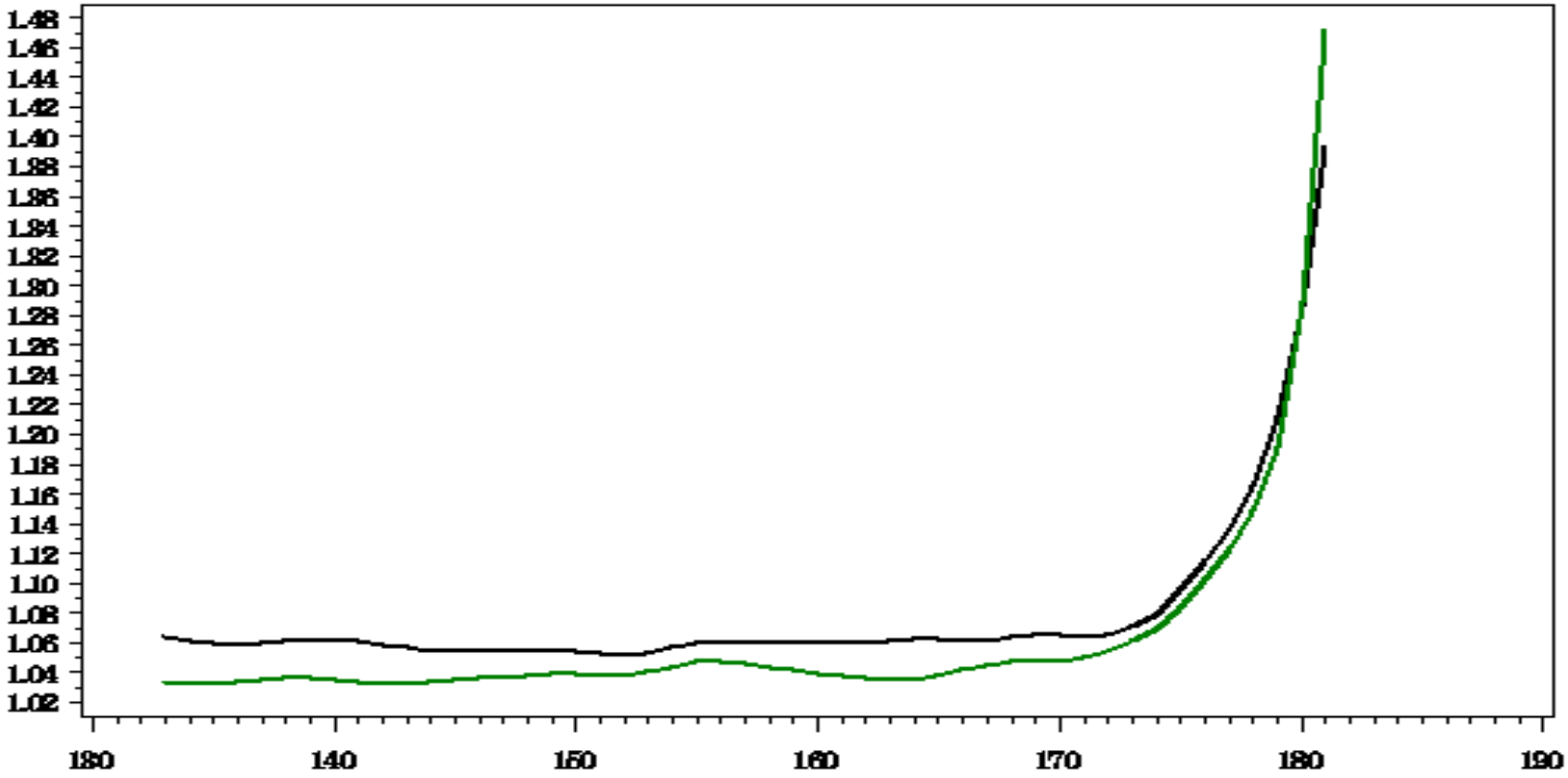
---- Empirical RMSE ; ---- Mean of estimates

Figure 8. Empirical **conditional RMSE** of **X-11 ARIMA** trend estimates with **60** forecasts and **mean of RMSE estimates**, signal estimated by X-11 ARIMA.



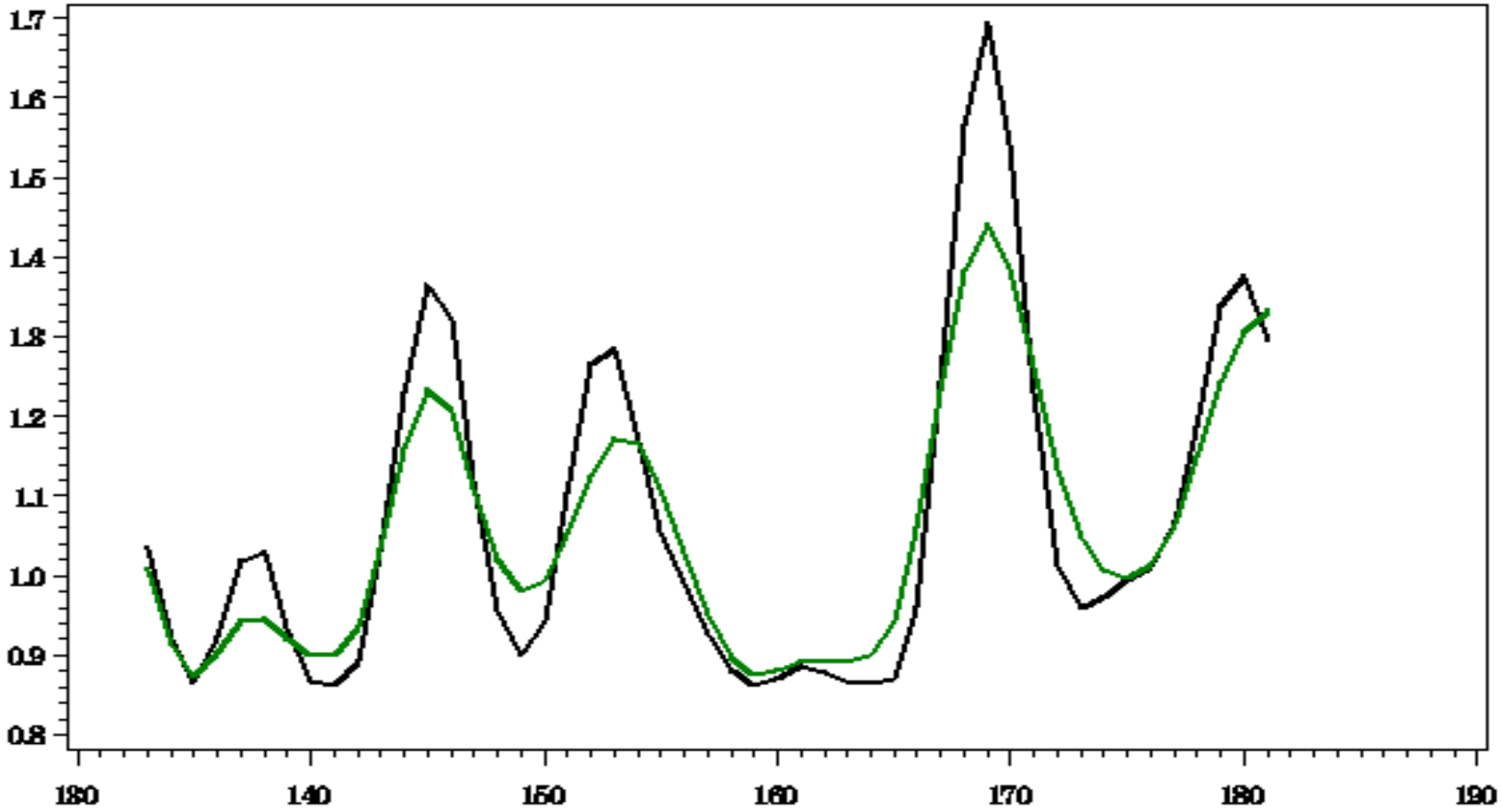
---- Empirical RMSE ; ---- Mean of estimates

Figure 9. Empirical **unconditional RMSE** of **BSM** trend estimates and **mean of RMSE estimates**, signals generated from **BSM** and estimated by X-11 ARIMA.



---- Empirical RMSE ; ---- Mean of estimates

Figure 10. Empirical **conditional RMSE** of **BSM** trend estimates and **mean of RMSE estimates**, signal estimated by X-11 ARIMA.



---- Empirical RMSE ; ---- Mean of estimates

Figure 11. **Empirical unconditional RMSE** of **X-11 ARIMA** trend estimates with **12** forecasts (**blue**), **60** forecasts (**red**) and **BSM (black)**, signals estimated by X-11 ARIMA.

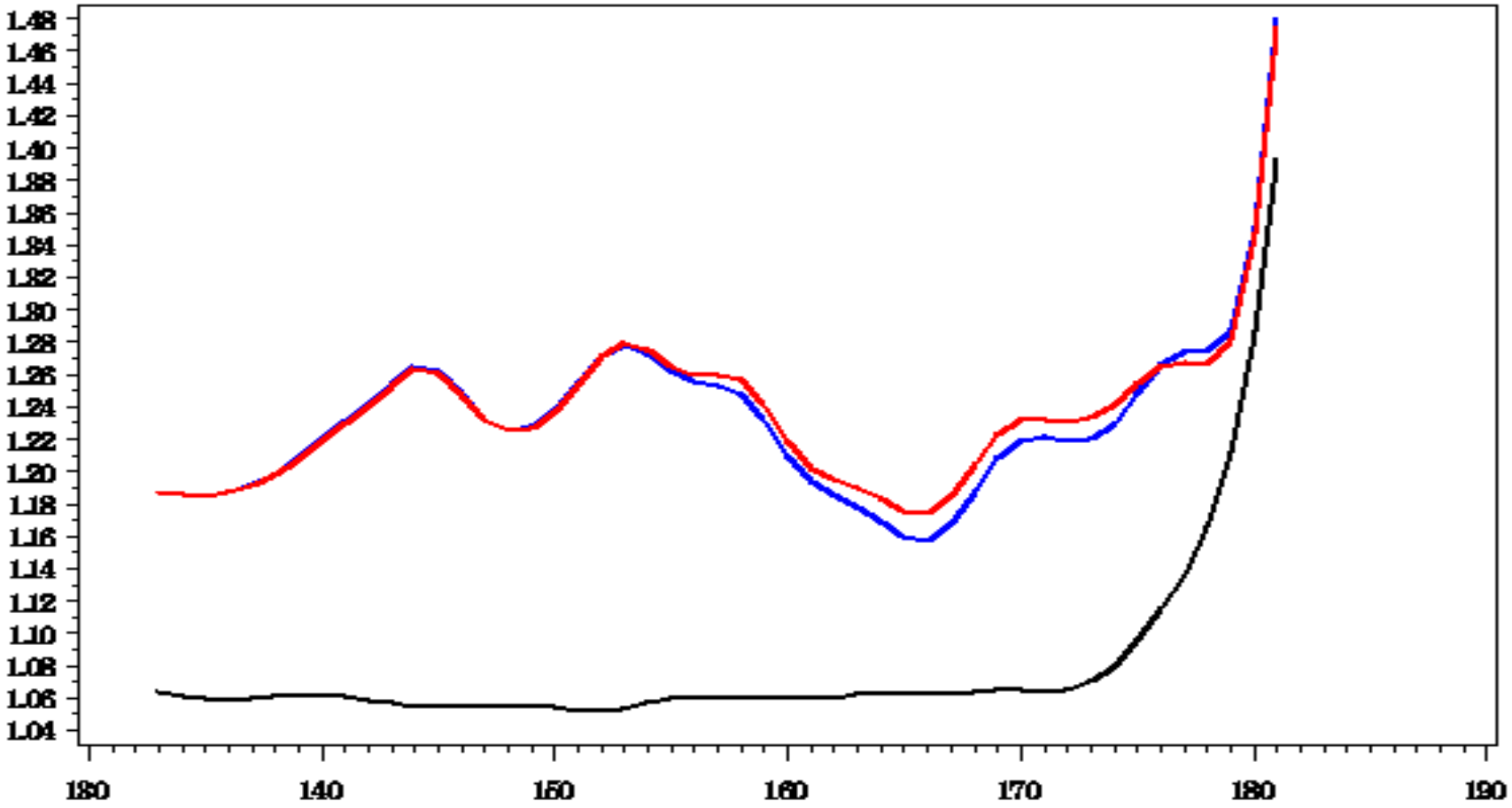
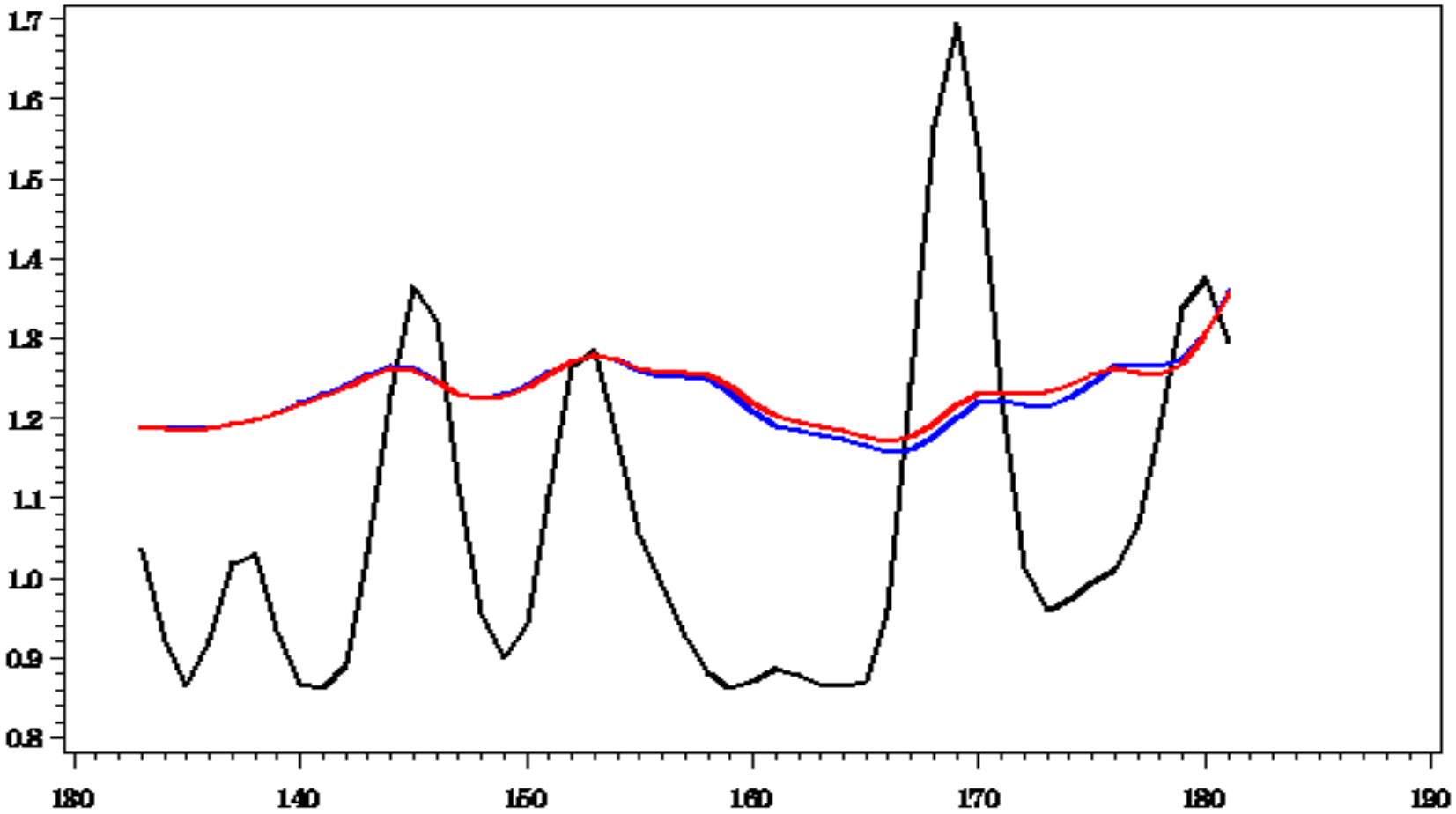


Figure 12. Empirical conditional RMSE of **X-11 ARIMA** trend estimates with **12** forecasts (**blue**), **60** forecasts (**red**) and **BSM (black)**, signal estimated by X-11 ARIMA.



Conclusions

- Method seems to work but need to experiment with many more simulated and real series,
- Investigate the robustness of the method to possbile model misspecification
- Study **efficient** estimation of signal under appropriate models.

Thanks!!! **(Sverchkov.Michael@bls.gov)**