Estimation of Mean Squared Error of X-11-ARIMA and Other Estimators Of Time Series Components

Michael Sverchkov
Bureau of Labor Statistics, Washington DC, U.S.A. Email: Sverchkov.Michael@bls.gov

Abstract

This paper considers the old but very important problem of how to estimate the mean squared error (MSE) of seasonally adjusted and trend estimators produced by X-11-ARIMA or other decomposition methods. The MSE estimators are obtained by defining the unknown target components like the trend and seasonal effects to be the hypothetical X-11 estimates of them that would be obtained if there were no sampling errors and the series was sufficiently long to allow the use of the symmetric filters embedded in the programme, which are time invariant. This definition of the component series conforms to the classical definition of the target parameters in design-based survey sampling theory, so that users should find it comfortable to adjust to this definition.

Key words: Bias correction; Seasonal Adjustment; Trend; X-13ARIMA-SEATS.

1. Introduction

We consider estimation of the mean squared error (MSE) of seasonally adjusted and trend estimators produced by X-11-ARIMA or other decomposition methods. We define the target seasonally adjusted and trend components to be the hypothetical X-11 estimates of them that would be obtained in the absence of sampling errors and if the time series under consideration was sufficiently long for application of the symmetric filters embedded in the original X-11 procedure, which are time invariant. This definition of the component series conforms to the classical definition of target finite population parameters in design-based survey sampling theory. In fact, in one variant of the proposed definition, the target components are shown to be linear combinations of finite population means or totals. The MSE of X-11-ARIMA and other estimators are defined with respect to this definition. We estimate the MSE by conditioning on the target components, thus accounting for possible conditional bias in estimating them. More detailed results can be found in Pfeffermann and Sverchkov (2014).

2. Target Components, Bias and MSE of X-11-ARIMA Estimators

2.1. Target components

We begin with the usual notion that an economic time series, $Y_t; t = 1, 2, ...$ can be decomposed into a trend or trend-cycle component T_t, a seasonal component S_t, and an irregular component I_t; $Y_t = T_t + S_t + I_t$. In practice, it is often the case that the series Y_t is unobserved and the available series consists of sample estimates, y_t, obtained from repeated sample surveys. Consequently, the series y_t can be expressed as the sum of the true population value, Y_t, and a sampling error, ε_t. More generally, the observed series can be viewed as the sum of a signal, G_t, and an error, e_t; $y_t = G_t + e_t$, where the signal, and hence the error, may be defined in two alternative ways:

*GE*1. $G_t = T_t + S_t$, $e_t = I_t + \varepsilon_t$. In this case e_t is the combined error of the time series irregular and the sampling error (Pfeffermann, 1994);

*GE*2. $G_t = T_t + S_t + I_t$, $e_t = \varepsilon_t$. In this case the irregular term is part of the signal, and e_t is the sampling error (Bell and Kramer, 1999).

We assume without loss of generality that the series started at time $-\infty < t_{start} < 1$, but y_t is only observed for the time points $t = 1,...,N$, such that

$$y_t = G_t + e_t,\ t = \underbrace{t_{start},...,0}_{unobserved},\ \underbrace{1,...,N}_{y_t\ observed},\ \underbrace{N+1,...,\infty}_{unobserved}. \quad (1)$$

It is assumed also that under both definitions of the signal, e_t is independent of $\mathbf{G} = \{G_t,\ t = t_{start},...,\infty\}$ for all t, with $E(e_t) = 0$, $Var(e_t) < \infty$, although in practice the sampling error, and in particular the variance of the sampling error, sometimes depends on the magnitude of the signal.

The X-11-ARIMA program first forecasts and backcasts the time series under consideration based on an ARIMA model fitted to the observed series, and then applies a sequence of moving averages (linear filters) to the series augmented by the forecasts and backcasts. It follows that the X-11-ARIMA estimators of the trend and the seasonal components can be approximated as,

$$\hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k},\ \hat{S}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^S y_{t+k}, \quad (2)$$

where the coefficients $\{w_{kt}^T\}$, $\{w_{kt}^S\}$ are defined in general by the program options for the observed time interval $t = 1,...,N$, the ARIMA model used to forecast and backcast the series and by the number of backcasts and forecasts. However, at the central part of the series, the filters in (2) are time-invariant and symmetric; $w_{kt}^T = w_k^T$, $w_{-k}^T = w_k^T$ for $a_T < t \le N - a_T$; $w_{kt}^S = w_k^S$, $w_{-k}^S = w_k^S$ for $a_S < t \le N - a_S$, where a_T, a_S are defined by the X-11-ARIMA program options. The length of the symmetric filters is $2a_T + 1$ ($2a_S + 1$), such that $w_{kt}^T = w_k^T = 0$ if $k \notin [-a_T, a_T]$ and $w_{kt}^S = w_k^S = 0$ if $k \notin [-a_S, a_S]$. Note that in the central part of the series the X-11-ARIMA estimators are the same as the X-11 estimators with no ARIMA extrapolations, such that the symmetric filters only depend on the X-11 program options and not on the ARIMA extrapolations.

Remark 1. The use of X-11-ARIMA involves also 'non-linear' operations such as the identification and estimation of ARIMA models used for forecasting and backcasting the original series, and the identification and gradual replacement of extreme observations. We assume that the time series under consideration is already modified for extreme values, thus robustifying the variance estimates described in Section 2.3. As illustrated in Pfeffermann *et al*. (1995) and Pfeffermann *et al*. (2000), the effects of the identification and non-linear estimation of ARIMA models are generally minor.

Definition 1. Assuming $t_{start} < \min(-a_T, -a_S)$ and following Bell and Kramer (1999), we define the trend component at time t to be $T_t^{X11} = \sum_{k=-a_T}^{a_T} w_k^T G_{t+k}$. Analogously, the seasonal component is defined as $S_t^{X11} = \sum_{k=-a_S}^{a_S} w_k^S G_{t+k}$. The target components T_t^{X11} and S_t^{X11} are thus the hypothetical components that would be obtained by application of the X-11 *symmetric filters* to the signal $\mathbf{G}$ at time point t, $t = 1,...,N$. It follows therefore that the observed series may be decomposed as the sum of the 'X-11-trend', T_t^{X11}, the 'X-11-seasonal component', S_t^{X11}, and the 'X-11 error', $e_t^{X11} = y_t - T_t^{X11} - S_t^{X11}$;

$$y_t = T_t^{X11} + S_t^{X11} + e_t^{X11}. \quad (3)$$

Result 1. For $a_T < t \le N - a_T$, $T_t^{X11} = E(\hat{T}_t \mid \mathbf{G})$ and for $a_S < t \le N - a_S$, $S_t^{X11} = E(\hat{S}_t \mid \mathbf{G})$, where $\hat{T}_t, \hat{S}_t$ are the X-11-ARIMA estimators defined in (2) and the expectation is taken over the distribution of the errors $\{e_t,\ t = 1,...,N\}$, with the signal $\boldsymbol{G}$ held fixed. It follows therefore from our definition

that in the central part of the series, the X-11-ARIMA estimators $\hat{T}_t, \hat{S}_t$ of the trend and the seasonal component are unbiased. (As noted before, we assume that the observed series is already modified for extreme values. The identification and estimation of ARIMA models are irrelevant at the center of the series.)

Remark 2. For X-11 filters $a_T > a_S$ because the final trend filter is applied after the final seasonal and seasonally adjusted components are computed. Thus, $\max(a_T, a_S) = a_T$.

Remark 3. We define the trend and seasonal components to be the (hypothetical) outputs that would be obtained when applying the symmetric filters to the signal, since the filters at the non-central parts of the series are asymmetric and depend on the time points with data. In particular, the filters applied for a time point $t > N - a_T$ change every time that a new observation is added to the series until $t \le N - a_T$, when the symmetric filter is applied. As mentioned before, the decomposition (3) has been used by Bell and Kramer (1999) with the error defined by the sampling error, such that the irregular term is part of the signal; $G_t = T_t + S_t + I_t$ (Definition GE2). Notice that with this definition, the target values are just linear combinations of the unadjusted population values of the series, which in most cases are finite population means or totals, in line with classical survey sampling theory.

2.2. Conditional Bias and MSE of X-11-ARIMA estimators

The conditional bias, variance and MSE of the X-11-ARIMA estimators of the trend with respect to the decomposition (3), given the signal, are as follows:

$$Bias(\hat{T}_t \mid \mathbf{G}) = E[(\hat{T}_t - T_t^{X11}) \mid \mathbf{G}] = \sum_{k=-(t-1)}^{N-t} w_{kt}^T G_{t+k} - \sum_{k=-a_T}^{a_T} w_k^T G_{t+k} \,. \tag{4}$$

$$\begin{aligned} Var[\hat{T}_t \mid \mathbf{G}] &= E\{[\sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k} - E(\sum_{k=-(t-1)}^{N-t} w_{kt}^T y_{t+k} \mid \mathbf{G})]^2 \mid \mathbf{G}\} \\ &= E\{[\sum_{k=-(t-1)}^{N-t} w_{kt}^T (y_{t+k} - G_{t+k})]^2 \mid \mathbf{G}\} = E(\sum_{k=-(t-1)}^{N-t} w_{kt}^T e_{t+k})^2 \end{aligned} \tag{5}$$

$$MSE(\hat{T}_t \mid \mathbf{G}) = E[(\hat{T}_t - T_t^{X11})^2 \mid \mathbf{G}] = Var(\hat{T}_t \mid \mathbf{G}) + Bias^2(\hat{T}_t \mid \mathbf{G}) \,. \tag{6}$$

Similar expressions hold for the seasonal and seasonally adjusted estimators.

The expressions (4)-(6) are general and apply to any linear estimator with arbitrary coefficients $\{w_{kt}^T\}$, as defined by the X-11-ARIMA options, the ARIMA model used for extrapolations and the number of forecasts and backcasts. In fact, and as shown in Section 3, the expressions (4)-(6) hold equally for other linear filters, not necessarily embedded in the X-11-ARIMA program. In the next sections we discuss ways of estimating the MSE in (6).

2.3. Variance estimation

Under Definition GE2 of the signal and error in Section 2.1, $e_t = \varepsilon_t$ is the sampling error, and by (5), $Var(\hat{T}_t \mid \mathbf{G}) = E(\sum_{k=-(t-1)}^{N-t} w_{kt}^T \varepsilon_{t+k})^2 = \sum_k \sum_l w_{kt}^T w_{lt}^T Cov(\varepsilon_{t+k}, \varepsilon_{t+l})$. Similar expressions apply when estimating the seasonal or the seasonally adjusted component. We assume the availability of estimates of the variances and covariances of the sampling errors, which enables estimation of the variance $Var(\hat{T}_t \mid \mathbf{G})$ and the variance of any other component estimator.

Next, consider the estimation of the variance under Definition GE1 of the signal and error, by which $e_t = I_t + \varepsilon_t$. By (5), the variance of the X-11-ARIMA trend estimator is in this case a linear combination of the covariances $v_{tm} = Cov(e_t, e_m)$, $t, m = 1, \ldots, N$. Following Pfeffermann (1994) and Pfeffermann

and Scott (1997), let $R_t = y_t - \hat{S}_t - \hat{T}_t = \sum_{k=-(t-1)}^{N-t} w_{kt}^R y_{t+k}$ define the linear approximation of the X-11-ARIMA residual term at time t, where $w_{0t}^R = 1 - w_{0t}^S - w_{0t}^T$ and $w_{kt}^R = -w_{kt}^S - w_{kt}^T$ for $k \neq 0$. Then,

$$Var(R_t \mid \mathbf{G}) = E\{[\sum_{k=-(t-1)}^{N-t} w_{kt}^R (y_{t+k} - E(y_{t+k} \mid \mathbf{G}))]^2 \mid \mathbf{G}\} = Var(\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k})$$
$$Cov(R_t, R_m \mid \mathbf{G}) = Cov[\sum_{k=-(t-1)}^{N-t} w_{kt}^R e_{t+k}, \sum_{l=-(m-1)}^{N-m} w_{lm}^R e_{m+l}] = \sum_k \sum_l w_{kt}^R w_{lm}^R Cov(e_{t+k}, e_{m+l}) \quad (7)$$

The residuals R_t are not stationary because of the use of asymmetric filters towards the two ends of the series. However, Let $U(m) = \frac{1}{N-m} \sum_{t=1}^{N-m} Cov(R_t R_{t-m})$, $m = 0, ..., N-1$ and suppose that the errors $e_t = I_t + \varepsilon_t$ are stationary (see Remark 4 below). Then, by (7), the vector $\mathbf{U}$ of the means $U(m)$ and the vector $\mathbf{V}$ of the covariances $V_k = Cov(e_t, e_{t+k}) = Cov(I_t + \varepsilon_t, I_{t+k} + \varepsilon_{t+k})$, $k = 0, ..., N-1$ are related by the system of linear equations,

$$\mathbf{U} = D\mathbf{V}, \quad (8)$$

where the matrix D is defined by the known weights $\{w_{kt}^R\}$. Since the X-11-ARIMA residuals are known for every $t = 1, ..., N$, one may estimate $U(m)$ by $\tilde{U}(m) = \frac{1}{N-m} \sum_{t=1}^{N-m} R_t R_{t-m}$. Substituting $\tilde{U}(m)$ for $U(m)$ in (8) enables estimation of $\mathbf{V}$ by solving the resulting equations; see Pfeffermann (1994) and Pfeffermann and Scott (1997). Notice that the use of (8) does not require the availability of estimates of the variances and covariances of the sampling errors. However, the estimators obtained in this way can be very unstable since the number of unknown variances and covariances generally equals the number of equations. A possible solution to this problem is to assume that the covariances V_k are negligible beyond some lag C and hence can be set to zero, and then solve the reduced set of equations for $V_0, ..., V_C$. This is a mild ergodicity condition assumed for the series e_t. Notice that with this assumption it is no longer needed to consider the estimates $\tilde{U}(m)$ for large m. Additionally, when estimates for the autocovariances of the sampling errors are available, they can be substituted into the vector $\mathbf{V}$ and taken as known, in which case one only needs to estimate the unknown variance and covariances of the time series irregular terms, I_t. This reduces the number of unknown covariances and hence the number of equations very drastically. Note that all these procedures are basically 'model free'. See Chen *et al*. (2003) for a different approach to estimating $\mathbf{U}$ and $\mathbf{V}$. Bell and Kramer (1999) consider model based estimation of the variance and covariances of the sampling errors.

Remark 4. The linear equations in (8) can easily be extended to the case of heteroscedastic sampling errors for which $V_{tk} = Cov(e_t, e_{t+k}) = L_{tk} V_k$ with known coefficients L_{tk}. Another potential modification consists of utilizing all the equations (or most of them) in (8), and estimating $V_0, ..., V_C$ by a discounted least-squares procedure.

2.4. Bias and MSE estimation

Estimation of the conditional bias of the estimator $\hat{T}_t$ (or any other linear estimator) given the signal, and hence the conditional MSE is more involved. We propose to estimate the bias by estimating the signal and then substituting the estimate in the right hand side of the bias expression (4). A possible way of estimating the signal is by application of the programme X-13ARIMA-SEATS, which is now in common use in many statistical bureaus around the world (replacing X-12-ARIMA). The programme

enables extraction of the models holding for the trend and the seasonal effects from the ARIMA model fitted to the observed series, and use of these models in order to estimate the signal within the observation period, and to forecast and backcast the signal for a_T time points with no observations. Denote by $\hat{G}_t$ the estimated signal for time t, including before or after times $1,...,N$. The bias is estimated then as,

$$\hat{Bias}[\hat{T}_t \mid \mathbf{G}] = \hat{E}[(\hat{T}_t - T_t^{X11}) \mid \mathbf{G}] = \sum_{k=-(t-1)}^{N-t} w_{kt}^T \hat{G}_{t+k} - \sum_{k=-a_T}^{a_T} w_k^T \hat{G}_{t+k}, \quad t = 1,...,N. \tag{9}$$

Use a similar expression for estimating the bias of the seasonally adjusted estimator.
The SEATS models are obtained by application of canonical signal extraction and under correct model specification, the estimators have minimum MSE (Hilmer and Tiao, 1982).

Having estimated the conditional variance and bias, a conservative estimator of the conditional MSE defined by (6) is obtained by adding the variance estimator to the square of the bias, i.e.,

$$\hat{MSE}(\hat{T}_t \mid \mathbf{G}) = \hat{Var}(\hat{T}_t \mid \mathbf{G}) + \hat{Bias}^2(\hat{T}_t \mid \mathbf{G}). \tag{11}$$

The estimator in (11) is conservative since $E[\hat{Bias}^2(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}] = \{E[\hat{Bias}(\hat{T}_t \mid \boldsymbol{G})] \mid \boldsymbol{G}\}^2 + Var[\hat{Bias}(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}] > \{E[\hat{Bias}(\hat{T}_t \mid \boldsymbol{G})] \mid \boldsymbol{G}\}^2$. The overestimation of the MSE can be corrected by subtracting an estimate of $Var[\hat{Bias}(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}]$. Notice that $\hat{Bias}(\hat{T}_t \mid \mathbf{G})$ is a linear combination of the signal estimates, $\hat{G}_t$, which in turn are linear combinations of the observed series, y_t. Thus, $\hat{Bias}(\hat{T}_t \mid \mathbf{G})$ is a linear combination of the y_t's and hence $Var[\hat{Bias}(\hat{T}_t \mid \mathbf{G}) \mid \mathbf{G}]$ can be estimated similarly to the estimation of $Var[\hat{T}_t \mid \mathbf{G}]$ discussed in Section 2.3. The weights defining $\hat{Bias}(\hat{T}_t \mid \boldsymbol{G})$ can be obtained similarly to Burck and Sverchkov (2001).

3. Estimation of MSE of Model-Based and Other Estimators of X-11 Components

Consider any other set of component estimators of the form,

$$\tilde{T}_t = \sum_{k=-(t-1)}^{N-t} h_{kt}^T y_{t+k}, \quad \tilde{S}_t = \sum_{k=-(t-1)}^{N-t} h_{kt}^S y_{t+k}. \tag{13}$$

Then, similar to the X-11-ARIMA estimators in Section 2, we can calculate the conditional bias and MSE with respect to the target X-11 components defined in Definition 1, yielding the same expressions as in (4)-(6) but with the weights $w_{kt}^T (w_{kt}^S)$ replaced by the weights $h_{kt}^T (h_{kt}^S)$. Notice that unlike the X-11 estimators, the estimators defined by (13) are potentially biased when conditioning on the signal even at the center of the series.
The weights in (13) can be calculated as in Burck and Sverchkov (2001). As in Section 2.4, the bias is estimated in this case by estimating the augmented signal $\mathbf{G}^{aug} = (G_{-a_T+1},...,G_0,...,G_N,...,G_{N+a_T})$ under an appropriate model. The bias and MSE estimators are obtained similarly to Eqs. (9)-(11).

4. Summary

In this paper we propose a new method for estimation of MSE of X-11 ARIMA estimators or other linear estimators of the underlying components of a time series. Our approach has some important advantages over other approaches proposed in the literature. First, we follow Bell and Kramer (1999) by defining the target component values as the corresponding X-11 estimators that would be obtained if the series was free of sampling errors and long enough to permit the use of the symmetric filters embedded in the program. In other words, the target components are real entities defined as linear combinations of finite population means or totals over time, in close correspondence to the target values in classical finite population sampling. In particular, under definition GE2 of the signal, the target component values are just linear combinations of the unadjusted finite population values. Interestingly,

while the programme X-11 for seasonal adjustment and its previous and subsequent versions have been in wide use for many decades, the target estimated values were never defined in a precise form. This is rather unusual in statistics where an estimator is defined but not what is estimated. This problem does not exist when using model dependent methods where the targets are defined by the model, such as in the BSM, the Tramo and Seats program (Gomez and Maravall, 1996) and in one of the modules of X-13ARIMA-SEATS, but purely model dependent estimators are not in common use, at least not in national statistical offices.

A second important advantage of the procedure is its flexibility in terms of the target values and the estimators used. It is applicable to the case where the signal consists of only the trend and the seasonal effect and the time series irregular is part of the error (definition GE1 of the signal and error), and to the case where the irregular is part of the signal, as under the Bell and Kramer (1999) approach. It is up to the user to decide which definition of the signal is more appropriate. In addition, the procedure is applicable to any linear estimator with known coefficients.

Taking into account the clear interpretation of the target values and the estimated MSE and the other advantages listed above, we hope that our proposed procedure will be experimented with by other users and we shall be happy to receive questions arising from these experiments.

References

Bell, W. R. and Kramer, M. (1999), Toward Variances for X-11 Seasonal Adjustments, Survey Methodology **25**, 13-29.

Burck, L., and Sverchkov, M. (2001) A general method for estimating the variances of X-11-ARIMA estimators, Federal Committee on Statistical Methodology Research Conference, **3**, 1 - 11

Chen, Z.G., Wong, P., Morry, M., and Fung, H. (2003), Variance Estimation for X-11 Seasonal Adjustment Procedure: Spectrum Approach and Comparison, Statistics Canada Report BSMD-2003-001E.

Gómez, V. and Maravall, A. (1996). Programs TRAMO and SEATS, Introduction for User (Beta Version). Banco de España Working Papers 9628, Banco de España.

Hilmer, S.C., and Tiao, G.S. (1982) An ARIMA-Model-Based Approach to Seasonal Adjustment, Journal of the American Statistical Association, **77**, pp. 63-70.

Pfeffermann, D. (1994), A General Method for Estimating the Variances of X-11 Seasonally Adjusted Estimators, Journal of Time Series Analysis **15**, 85-116.

Pfeffermann, D., Morry, M., and Wong, P. (1995), Estimation of the Variances of X-11 ARIMA Seasonally Adjusted Estimators for a Multiplicative Decomposition and Heteroscedastic Variances, International Journal of Forecasting, **11**, 271-283.

Pfeffermann, D., and Scott, S. (1997), Variance Measures for X-11 Seasonally Adjusted Estimators; Some Developments with Application to Labor Force Series, Proceedings of the ASA Section on Business & Economic Statistics, 211-216.

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